RealisID: Scale-Robust and Fine-Controllable Identity Customization via Local and Global Complementation

Zhaoyang Sun[1,2*] , **Fei Du**[2,3*], **Weihua Chen**[2,3], **Fan Wang**[2,3]
Yaxiong Chen[1], **Yi Rong**[1†], **Shengwu Xiong**[4,5†]
[1]Wuhan University of Technology
[2]DAMO Academy, Alibaba Group
[3]Hupan Laboratory
[4]Shanghai AI Laboratory
[5]Interdisciplinary Artificial Intelligence Research Institute, Wuhan College

Abstract

Recently, the success of text-to-image synthesis has greatly advanced the development of identity customization techniques, whose main goal is to produce realistic identity-specific photographs based on text prompts and reference face images. However, it is difficult for existing identity customization methods to simultaneously meet the various requirements of different real-world applications, including the identity fidelity of small face, the control of face location, pose and expression, as well as the customization of multiple persons. To this end, we propose a scale-robust and fine-controllable method, namely RealisID, which learns different control capabilities through the cooperation between a pair of local and global branches. Specifically, by using cropping and up-sampling operations to filter out face-irrelevant information, the local branch concentrates the fine control of facial details and the scale-robust identity fidelity within the face region. Meanwhile, the global branch manages the overall harmony of the entire image. It also controls the face location by taking the location guidance as input. As a result, RealisID can benefit from the complementarity of these two branches. Finally, by implementing our branches with two different variants of ControlNet, our method can be easily extended to handle multi-person customization, even only trained on single-person datasets. Extensive experiments and ablation studies indicate the effectiveness of RealisID and verify its ability in fulfilling all the requirements mentioned above.

Introduction

The tremendous success of text-to-image synthesis techniques (Saharia et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Wang et al. 2024b) have spawned and driven a variety of customized generation tasks (Yang et al. 2023; Kim et al. 2024b; Wei et al. 2024). In this paper, we study one of the most prominent of these tasks, namely identity (ID) customization. It aims to adapt text-to-image synthesis models to generate new images, which match the identities depicted in the given reference face images and follow the controls prompted by the input text.

Most early researches on ID customization typically fine-tune some specific parameters on a set of images containing the same ID so that the information associated with this ID can be integrated into the model. Although fine-tuning-based methods (Hu et al. 2021; Ruiz et al. 2023; Gal et al. 2022) have achieved commendable results, their computational costs are often significant. Even with advanced GPUs, fine-tuning the model parameters for every single ID can takes several minutes, which may make these methods infeasible in practical applications. To alleviate this issue, a series of recent works (Xiao et al. 2023; Wang et al. 2024a; Zhang et al. 2024; Guo et al. 2024) attempt to leverage ID-related prior knowledge learned from other large-scale face datasets to support fast inference without the need for fine-tuning. Generally, these approaches merge the ID features encoded from CLIP (Radford et al. 2021) or pre-trained face recognition model into the text embeddings or generative models by introducing trainable ID adapters (Ye et al. 2023; Mou et al. 2024; Zhang, Rao, and Agrawala 2023).

Despite the effectiveness and efficiency, all existing methods fail to simultaneously satisfy the various requirements (as illustrated in Fig. 1) of different real-world applications: (1) **Identity Fidelity of Small Face.** The faces in the generated images may need to be of different sizes. While most previous methods are effective in generating large faces, they often struggle with maintaining the identity details for small ones. (2) **Flexible and Fine Control.** In addition to identity information, we may also wish to finely adjust certain factors in the target images, such as face location, pose and expression. Since these factors cannot be precisely described through text, existing approaches that rely solely on text prompts will face challenges in appropriately controlling them. (3) **Multi-person Customization.** Most current methods are limited to customization for a single individual and lack the flexibility required for practical applications involving multiple persons. In addition, the scarcity of multi-person datasets available for model training further exacerbates the difficulty of meeting this requirement.

To this end, in this paper we propose a novel unified identity customization framework, namely RealisID, which is

[*]Equal contribution. Work done during internship of Zhaoyang Sun at DAMO Academy, Alibaba Group
[†]Corresponding author.

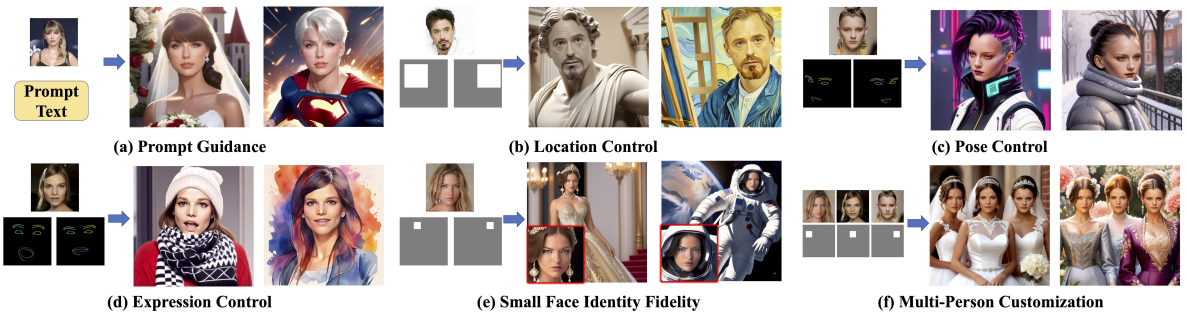

Figure 1: Our RealisID can flexibly and finely control the face location, pose and expression factors of the generated facial images. It is also able to keep high identity fidelity for small faces and easily generalizes to multi-person customization.

scale-robust and fine-controllable, thus can fulfill all the requirements mentioned above. Similar to the recently proposed state-of-the-art methods (Wang et al. 2024a; Guo et al. 2024; Li et al. 2024b), our RealisID framework also attempts to inject additional condition signals into a pre-trained generative model (e.g., Stable Diffusion (Rombach et al. 2022; Podell et al. 2023)). But different from all existing works, RealisID performs the condition information injection both locally and globally through two complementary branches. Specifically, we observe that the face-irrelevant contents in the reference image will prevent its latent space embedding from aligning well with the face-focusing control conditions (e.g., pose and expression). This may lead to difficulties in achieving the fine control of facial details and the identity fidelity for small faces. To address this problem, our local branch crops the face region from the latent embedding of the whole image, and up-samples it to the same spatial size. By taking this cropped embedding as input, the local branch will focus on injecting facial details information. And also, the up-sampling operation narrows the face scale differences across different reference images, making our local branch to be scale-robust. In addition, we also design a global branch to integrate face location information and manage the overall harmony of the entire image. The complementarity of these two branches allows RealisID to generate high-quality facial images that can be flexibly manipulated. Furthermore, by implementing our two branches with the ControlNet (Zhang, Rao, and Agrawala 2023), RealisID can be easily extend to handle multi-person customization, even without training on multi-person data. Our main contributions are summarized as follows:

- We introduce RealisID, which achieves scale-robust and fine-controllable ID customization through the collaborative efforts of two complementary branches.
- To the best of our knowledge, by equipping the newly proposed local branch, our RealisID is the first attempt for scale-robust ID customization, which can effectively maintain the identity fidelity for small faces.
- Extensive comparisons indicate that RealisID outperforms five recent state-of-the-art ID customization methods, especially in the small face scenario. And ablation studies validate the ability of RealisID in facial factors fine control and multi-person customization.

Related Works

Text-to-Image Synthesis Empowered by diffusion models, text-to-image synthesis techniques have achieved significant advances. The typical procedure involves encoding textual prompts into latent vectors using pre-trained text encoders such as CLIP (Radford et al. 2021) to steer the diffusion process. Imagen (Saharia et al. 2022), GLIDE (Nichol et al. 2021), and DALL-E (Ramesh et al. 2022) perform denoising directly in the original pixel space. Stable Diffusion, a standout work in latent diffusion models (Rombach et al. 2022), utilizes autoencoders for denoising in latent space to optimize computational efficiency while maintaining the ability to generate high-quality images. In continuation, SDXL (Podell et al. 2023) introduces a larger UNet, a refiner model, and a second text encoder to bolster image quality and text control. Owing to the remarkable achievements in image synthesis, text-to-image models have been widely used as the backbone for many generative tasks.

Identity Customization ID customization is a challenging task in subject-driven image generation. Current research can be categorized into two groups based on the necessity of fine-tuning during inference. Fine-tuning methods (Hu et al. 2021; Ruiz et al. 2023; Gal et al. 2022; Wang et al. 2024b) involve adjusting parameters on multiple images with the same ID or injecting the ID condition into the model using inversion techniques. However, optimization-based approaches require individual training for each new role, leading to substantial computational expenses and constraining flexibility and practicality. In contrast, fine-tuning free methods show promising results on zero-shot identity customization. These methods (Xiao et al. 2023; Liang et al. 2024; Cui et al. 2024; Li et al. 2024b; Peng et al. 2024; Wang et al. 2024a; Zhang et al. 2024; Guo et al. 2024) initially encode reference facial features into one or more labels, then train additional adapters (such as IP-Adapter, ControlNet) and integrate identity conditioning into the pre-trained text-to-image model. However, these methods may encounter poor identity fidelity in small faces, lack precise and adaptable facial control, or fail to scale to multi-person customization. To overcome these challenges, this paper introduces RealisID, which eliminates the need for fine-tuning during the inference stage and can be seamlessly extended to handle multi-person customization task (Kim et al. 2024a).

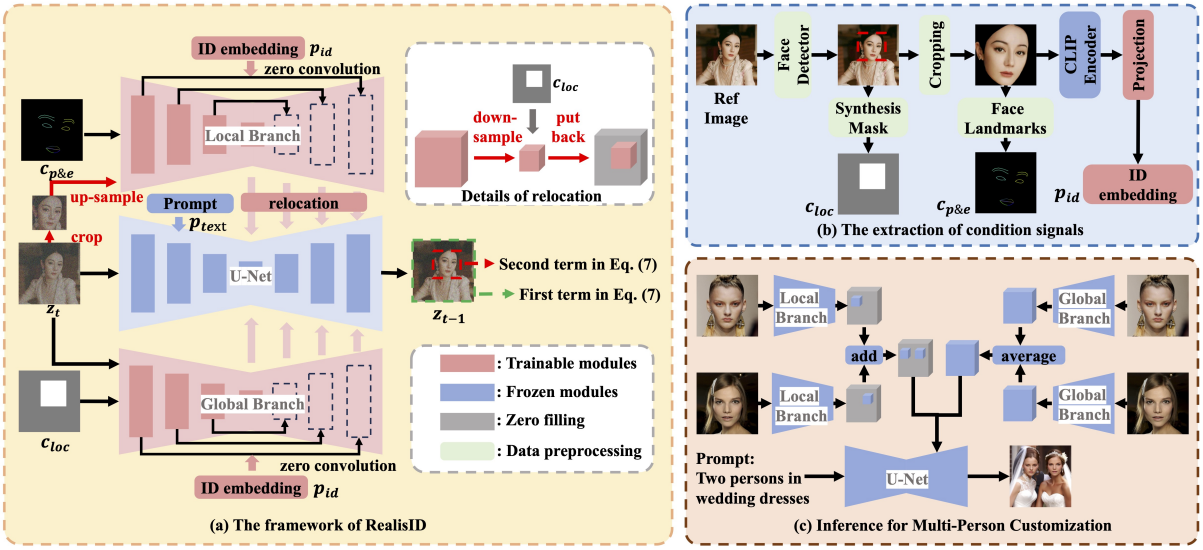

Figure 2: (a) The overall architecture of our RealisID framework, which constructs a pair of local and global branches to inject additional condition information into the U-Net denoiser of a pre-trained stable diffusion model. (b) The procedure of extracting different condition signals from the input reference images (c) The inference strategy for handling multi-person customization.

Methodology

Preliminary

Since our RealisID framework is built based on Stable Diffusion and ControlNet models, we briefly review them first.

Stable Diffusion (SD) is a large-scale text-to-image latent diffusion model trained on LAION (Schuhmann et al. 2022) dataset. It consists of three modules: The encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ transform the input images into latent space and back to pixel space, respectively, playing a crucial role in reducing the computational complexity. And the U-Net denoiser ϵ_θ is trained to predict the applied diffusion noise in the latent space as follow:

$$\mathcal{L}_{sd} = \mathbb{E}_{z_t,p,\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}(1,T)}[\| \epsilon_\theta(z_t, p, t) - \epsilon \|_2], \quad (1)$$

where z_t is the noisy latent variable at the t-th timestep, p is a prompt signal. The noise ϵ is sampled from the standard Gaussian distribution and T denotes the maximum timestep.

ControlNet is designed to integrate additional control conditions into pre-trained generative models (e.g., Stable Diffusion). Specifically, it injects the following condition information into the intermediate representation produced by each decoding layer of the U-Net denoiser ϵ_θ:

$$i_t = \mathcal{Z}(\mathcal{F}(z_t + \mathcal{Z}(c; \Theta_{z1}), p, t; \Theta_c); \Theta_{z2}), \quad (2)$$

where c is the input condition. $\mathcal{Z}(\cdot;\Theta_{z1})$ and $\mathcal{Z}(\cdot;\Theta_{z2})$ indicate two different zero convolution layers with parameters Θ_{z1} and Θ_{z2}, respectively. $\mathcal{F}(\cdot,\cdot,\cdot;\Theta_c)$ consists of trainable copy blocks whose parameter is Θ_c.

The Proposed RealisID Framework

Our RealisID framework consists of two branches, whose structures are two different variants of ControlNet. As illustrated in Fig. 2(a), these two branches introduce distinct condition control information to the pre-trained Stable Diffusion U-Net denoiser, respectively. The local branch focuses on injecting facial details information within the local face region, such as head pose and facial expressions as well as person identity. In contrast, the global branch manages the overall harmony of the entire image, with the face location and the corresponding body and background layouts being manipulated through this branch. The cooperation of these two branches enables the U-Net denoiser to exploit both of their complementary information, thus allowing the generative process to be flexibly and finely controlled.

Condition Signals for Model Training In order to train both branches in our RealisID framework and achieve the corresponding control capabilities mentioned above, we first need to obtain the relevant condition signals from the training samples. Specifically, as shown in Fig. 2(b), given a reference image and its associated description text, we extract the following three types of condition signals from them:

(1) ID and Text Embeddings. For each reference image, we first apply a face detection model to generate a bounding box that locates the face region, which is then manually adjusted into a square based on its longer edge. After that, the input image is cropped according to the obtained bounding box, and a pre-trained face parsing model is used to zero-out its irrelevant background areas. Subsequently, this cropped face image is fed into the CLIP (Radford et al. 2021) image encoder to extract image features from the penultimate hidden layer. Finally, we obtain the ID embedding p_{id} of the reference image by aligning these extracted CLIP features with the U-Net latent space through a projection layer. For the input text, we use the same operations as SDXL (Podell et al. 2023) to get the corresponding text embedding p_{text}. During the model training, p_{id} acts as the prompt signal of both local and global branches in our framework, and p_{text} is fed into the U-Net denoiser as the text prompt input.

(2) Pose-Expression Representation. Apart from iden-

tity, the head pose and facial expression are also essential elements of the human face. To achieve the control of these factors, inspired by (Ma et al. 2024; Wei, Yang, and Wang 2024), we take the facial landmarks of the above cropped face image as the pose-expression representation $c_{p\&e}$. It will be utilized as the condition input of our local branch.

(3) Location Guidance. Based on the squared bounding box, we can generate a mask as the face location guidance c_{loc}. This mask is a single-channel matrix that fills the region inside the facial bounding box with 1 and other areas with 0. In addition to face region, c_{loc} also implicitly indicates the locations of human body and image background. For model training, c_{loc} is input into the global branch as the condition, and is also used for relocation operations in the local branch.

Local Branch According to Eq. (2), with the ID embedding p_{id} as the prompt signal, the pose-expression representation $c_{p\&e}$ as the condition, and the noisy latent variable z_t, the injection information of our local branch can be produced as follow:

$$i_{t\_l} = \mathcal{Z}(\mathcal{F}(z_t + \mathcal{Z}(c_{p\&e}; \Theta_{z1\_l}), p_{id}, t; \Theta_{c\_l}); \Theta_{z2\_l}). \quad (3)$$

By observing Eq. (3), we can find that the face-irrelevant information in z_t will significantly influence the effectiveness of i_{t_l}. This problem can be even worse when the face region becomes relatively smaller in the whole reference image. In such case, the face-irrelevant information will dominate z_t, making it unable to align with the pose-expression condition information $\mathcal{Z}(c_{p\&e}; \Theta_{z1})$, thus decreasing the ability of i_{t_l} for fine control and identity preservation. To deal with this problem, we takes cues from previous studies (Bai et al. 2018; Noh et al. 2019) on improving small object detection through super-resolution. Specifically, instead of using the entire z_t, we crop and bilinearly up-sample the face area from it, and then take the obtained latent embedding $\hat{z}_t$ as the input of our local branch network. In this way, $\hat{z}_t$ will mainly focus on facial details information and can be well aligned with $\mathcal{Z}(c_{p\&e}; \Theta_{z1})$. And moreover, since the face regions in different reference images are up-sampled to the same input size, their scale differences can be effectively reduced, **making our local branch to be robust to face scale.** However, cropping the face from z_t will result in the loss of its location information. To address this, we further insert a relocation operation before the information injection, which can be formulated as:

$$\hat{i}_{t\_l} = \mathcal{R}(i_{t\_l}, c_{loc}), \quad (4)$$

$$i_{t\_l} = \mathcal{Z}(\mathcal{F}(\hat{z}_t + \mathcal{Z}(c_{p\&e}; \Theta_{z1\_l}), p_{id}, t; \Theta_{c\_l}); \Theta_{z2\_l}). \quad (5)$$

As shown in Fig. 2, the relocation operation $\mathcal{R}(i_{t\_l}, c_{loc})$ first down-samples i_{t_l} according to the relative size of the face region to the entire input image. Then based on the location guidance c_{loc}, it puts the down-sampled features back to the face position in a zero tensor of the same size as i_{t_l}. Finally, the obtained $\hat{i}_{t\_l}$ is injected into the U-Net desnoier to achieve the flexible and fine control of facial details as well as the scale-robust identity fidelity.

Global Branch By taking the ID embedding p_{id} as the prompt signal, the location guidance c_{loc} as the condition, and the entire noisy latent variable z_t, our global branch generates the following injection information:

$$i_{t\_g} = \mathcal{Z}(\mathcal{F}(z_t + \mathcal{Z}(c_{loc}; \Theta_{z1\_g}), p_{id}, t; \Theta_{c\_g}); \Theta_{z2\_g}). \quad (6)$$

Based on the face region, c_{loc} can implicitly indicate the locations of human body and image background, therefore it can provide the overall layout information of the whole image for i_{t_g}. As a result, by injecting i_{t_g} in the U-Net desnoier, we can accurately control the face location and obtain harmonious generation results.

Training Loss During the training phase, we only update the parameters of our two branches and the projection layer for generating the ID embedding, while freezing those of the pre-trained Stable Diffusion model. To this end, we optimize the following objective function:

$$\mathcal{L} = \mathbb{E}_{z_t, p_{text}, \epsilon \sim \mathcal{N}(0,1), t \sim \mathcal{U}(1,T)} [\| \epsilon_\theta(z_t, p_{text}, t) - \epsilon \|_2 + \lambda \| (\epsilon_\theta(z_t, p_{text}, t) - \epsilon) \odot c_{loc} \|_2]. \quad (7)$$

Here, $\odot$ denotes the element-wise multiplication operation. The first term in Eq. (7) is a standard stable diffusion loss, and the second term only calculates the noise discrepancy within the face area indicated by c_{loc}, thus facilitating the generation of facial details. $\lambda > 0$ is a positive hyperparameter that balances the importance of these two terms.

Inference Strategy Following (Zhang et al. 2024), we employ the classifier-free guidance scheme (Ho and Salimans 2022) for model inference. It combines different types of noise predicted under three conditions: without ID embedding nor text prompt ϵ_{none}, with only text prompt ϵ_t, and with both ID embedding and text prompt $\epsilon_{t\&i}$, as follow:

$$\epsilon_{prd} = \epsilon_{none} + \lambda_t(\epsilon_t - \epsilon_{none}) + \lambda_i(\epsilon_{t\&i} - \epsilon_t), \quad (8)$$

where ϵ_{prd} denotes the final noise prediction for inference procedure. $\lambda_t > 0$ and $\lambda_i > 0$ serve as two weighting factors for text and image guidance, respectively.

Inference for Multi-Person Customization For the customization of multiple persons, given the reference image and the target location guidance of each individual, we need to integrate the injection information for all these conditions that are produced by our two branches, respectively. As illustrated in Fig. 2(c), for the local branch, the injection information of different reference images are simply summed, since each of them only affects the local areas of the whole image, while those generated by the global branch are averaged. The integrated information is then injected into the U-Net denoiser for noise prediction.

Experiments

Implementation Details

To improve the realism of generated results, we establish our framework based on the pre-trained SDXL-1.0 (Podell et al. 2023) text-to-image synthesis model. The trainable parameters in our RealisID model are learned from the publicly available CosmicMan dataset (Li et al. 2024a), which comprises 2 million image-text pairs of single individuals. All the reference images are manually cropped and resized to

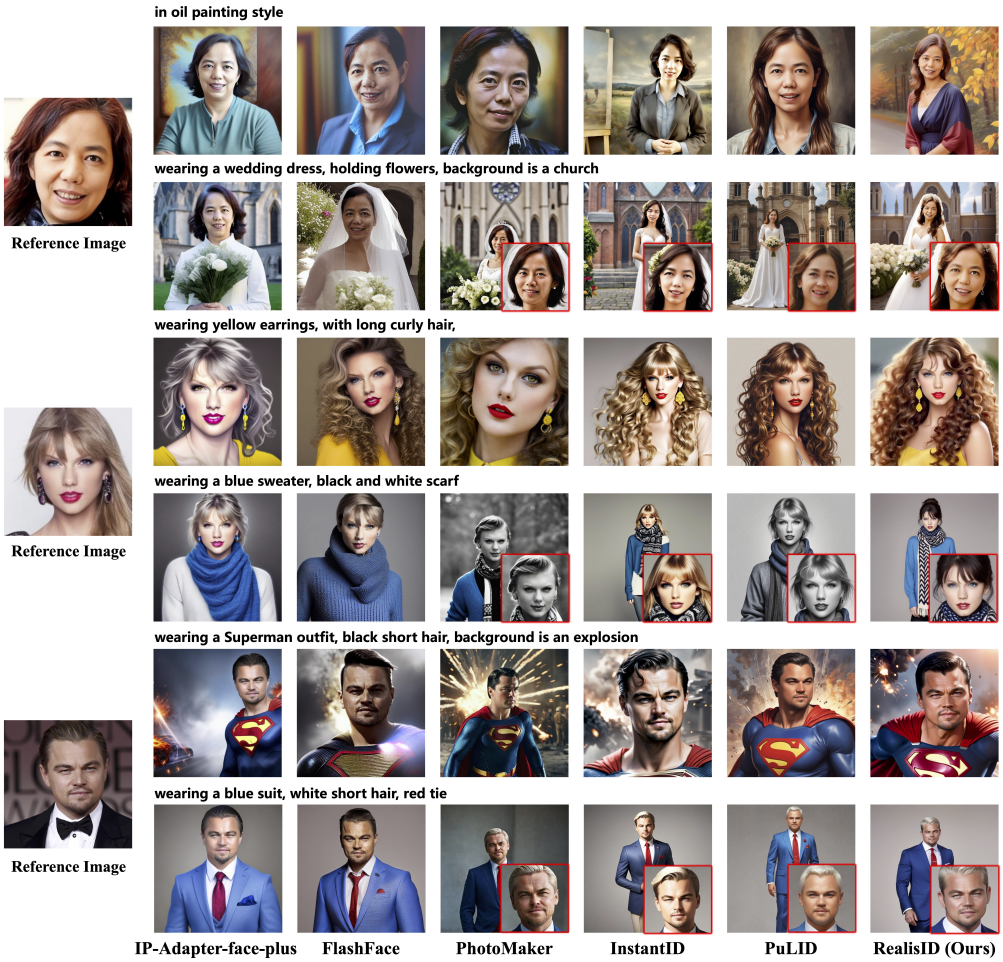

Figure 3: Qualitative comparison between different methods. The odd and even rows correspond to regular and small face scenarios, respectively. Regardless of face scales, our RealisID framework achieves the high fidelity of identity and facial details, thus generating visually appealing portrait images.

1024×1024 pixels. To extract the condition signals from image samples, we use MTCNN (Zhang et al. 2016) for face detection, BiSeNet (Yu et al. 2018) for face parsing and MediaPipe (Lugaresi et al. 2019) for facial landmark detection. During the training phase, we follow the learning strategy of IP-Adapter (Ye et al. 2023) that randomly drops either the image prompt (i.e., ID embedding) or the text prompt or both of them with a probability of 0.05. The hyperparameter λ in Eq. (7) is set to 1.0. The framework is optimized on 8 NVIDIA H20 GPUs, using the Adam optimizer with batch size of 16, learning rate of 1e-5 and weight decay of 1e-2. For inference, we adopt the same delayed subject conditioning technique as in (Xiao et al. 2023). We set $\lambda_t = 7.5$ and $\lambda_i = 5.0$ in Eq. (8), and use a 30-step DDIM (Song, Meng, and Ermon 2020) sampler to generate the target images.

Baseline Methods

We compare our RealisID framework with five recently proposed state-of-the-art ID customization methods, including IP-Adapter-face-plus (Ye et al. 2023), FlashFace (Zhang et al. 2024), PhotoMaker (Li et al. 2024b), InstantID (Wang et al. 2024a), and PuLID (Guo et al. 2024). All these methods are implemented with their officially released code and pre-trained model parameters. Notably, our method keeps the same setting with PhotoMaker, InstantID and PuLID that utilizes SDXL-1.0 as the basic model, while IP-Adapter-face-plus and FlashFace are constructed based on SD-1.5. Following InstantID, when only a reference image is provided, the facial landmarks of the reference image are used as the pose-expression control condition input for our model.

Qualitative Comparisons

Regular Case In Fig. 3, the odd rows display the quantitative results generated by different competing methods in a regular scenario, without involving particular conditions other than the input image and text. We can see that all these methods except PhotoMaker exhibit high identity fidelity. It is mainly because that PhotoMaker projects the identity features into the text embedding space that is contextually ambiguous, thus resulting in the loss of identity detail informa-

Methods	Regular Case				Small Face			
	ASP ↑	CLIP-T ↑	FaceNet ↑	CLIP-I ↑	ASP ↑	CLIP-T ↑	FaceNet ↑	CLIP-I ↑
IP-Adapter-face-plus	5.64	0.201	0.760	0.714	-	-	-	-
FlashFace	5.46	0.212	**0.809**	**0.754**	-	-	-	-
PhotoMaker	5.74	0.223	0.508	0.651	5.78	0.229	0.516	0.658
InstantID	6.01	0.211	0.768	0.706	5.72	0.210	<u>0.693</u>	<u>0.686</u>
PuLID	**6.37**	**0.254**	0.772	0.697	<u>5.95</u>	**0.249**	0.497	0.585
RealisID (Ours)	<u>6.22</u>	<u>0.234</u>	<u>0.796</u>	<u>0.739</u>	**6.11**	<u>0.236</u>	**0.767**	**0.701**

Table 1: Quantitative comparison between different methods. Four metrics ASP, CLIP-T, FaceNet, and CLIP-I are calculated for both regular and small face cases. The best and second best results are highlighted in **bold** and <u>underlined</u>, respectively.

Methods	Small Face			
	ASP	CLIP-T	FaceNet	CLIP-I
w/o B_{local}	6.07	**0.243**	0.681	0.673
w/o B_{global}	5.97	0.241	0.734	0.689
Full Model	**6.11**	0.236	**0.767**	**0.701**

Table 2: Effects of each individual branch in RealisID.

Methods	Face Relative Size			
	1/4	1/5	1/6	1/7
InstantID	0.765	0.745	0.708	0.664
RealisID (Ours)	**0.791**	**0.787**	**0.772**	**0.748**

Table 3: Identity fidelity (FaceNet) of different face sizes.

tion. Moreover, the SDXL-1.0-based methods typically produce visually appealing and high-quality facial images due to the more powerful generation capabilities of SDXL-1.0.

Images with Small Faces Ensuring the identity fidelity for faces of different sizes is a critical requirement for ID customization. The even rows in Fig. 3 present the quantitative comparisons of all competing approaches in generating target images with small faces. *In this study, we define the "small face" as a human face that enclosed in a bounding box whose long side is less than 1/6 of the image edge.* For the methods that cannot control the face size, we introduce an additional text prompt "a full-body people image" to guide them in generating outcomes with small faces. From Fig. 3, we observe that IP-Adapter-face-plus and FlashFace do not have the ability to synthesize small face images. PhotoMaker, InstantID and PuLID also experience a significant degradation in identity fidelity when compared to their results in the regular case. In contrast, our RealisID framework consistently maintains high identity fidelity for small faces. This is mainly attributed to the detail fine control ability and the scale robustness of our local branch, demonstrating the effectiveness of its scale alignment (i.e., face cropping and up-sampling) and feature relocation operations.

Quantitative Comparisons

Evaluation Dataset Our evaluation data consists of 40 unseen identities obtained from another CelebA-HQ (Karras et al. 2017) dataset. For a comprehensive evaluation, we prepare 35 prompts that covering cover a variety of clothing, attributes, actions, backgrounds, and styles. Every method generates 2 images for each prompt of each identity, resulting in a total of 2800 synthesized images for evaluation.

Evaluation Metrics Similar to the previous work (Li et al. 2024b), we use CLIP-T (Radford et al. 2021) to measure the prompt fidelity of the generated images and use ASP[1] metric to evaluate their aesthetic. We also employ FaceNet (Schroff, Kalenichenko, and Philbin 2015) and CLIP-I (Gal et al. 2022) to assess the identity fidelity. Specifically, we calculate the identity similarity between the generated image and the reference image based on the face embeddings extracted by FaceNet from the face regions detected by MTCNN (Zhang et al. 2016). For all these metrics, the higher value indicates the better generative performance.

Quantitative Results For the methods capable of controlling the face size, we randomly set the relative size of face regions (i.e., the ratio of face bounding box long side length to the whole image edge length) in the range of $[1/4, 1/2]$ for the regular scenario, and set those in the range of $[1/7, 1/6]$ for the small face scenario. Similar to our quantitative analysis, for the methods that cannot manipulate the face size, an additional text prompt of "a full-body people image" is introduced for small face image generation. The quantitative results are presented in Table 1. Since IP-Adapter-face-plus and FlashFace fail to generate small face images (see Fig. 3), their corresponding evaluation metrics are not calculated. It can be seen that our RealisID achieves comparable results with other state-of-the-art methods in the regular case, which is consistently the second best method on all four metrics. Furthermore, in the small face scenario, RealisID showcases its superiority by achieving the best results on three out of four metrics. In particular, the significant improvements on FaceNet (0.767 vs. 0.693) and CLIP-I (0.701 vs. 0.686) indicate the effectiveness of our framework in maintaining high identity fidelity for small faces.

Ablation Studies

Complementarity of Local and Global Branches We construct two ablated models by separately removing the lo-

[1]https://github.com/christophschuhmann/improved-aesthetic-predictor

Figure 4: Effects of our local and global branches.

Figure 5: Ablation study on scale robustness.

cal and global branches from our full RealisID framework to investigate their effects on final generated results. Here, we use B_{local} and B_{global} to denote the two branches, respectively. We conduct this ablation study only in the small face scenario. As shown in Fig. 4, the removal of B_{local} leads to distortions in identity and facial details, which is also reflected in the reduction of FaceNet and CLIP-I metrics in Table 2. This suggests that B_{local} plays an important role in preserving and controlling face-relevant details information. When B_{global} is removed, the ablated model indeed synthesizes a face with higher identity fidelity (FaceNet: 0.734 and CLIP-I: 0.689) at the specified location. But this face is isolated from other image contents and does not blend well with the overall layout, thus achieving the lowest ASP value of 5.97. This demonstrates that B_{global} manages the global harmony of the entire image. In contrast, benefiting from the complementarity and cooperation between B_{local} and B_{global}, the full RealisID model produces more satisfactory and realistic results, both qualitatively and quantitatively.

Scale Robustness To study the scale robustness of our RealisID, we vary the relative size of target faces from 1/7 to 1/4, and compare the generation performance of InstantID and our method. As depicted in Fig. 5, compared to InstandID, our method is more effective in preserving facial structure and details across different face sizes. Table 3 further confirms that RealisID consistently surpasses InstantID in identity fidelity (measured by FaceNet metric) across all sizes, demonstrating its superior scale robustness.

Fine Control Our RealisID framework also provides fine control over other essential factors of the target face. Specif-

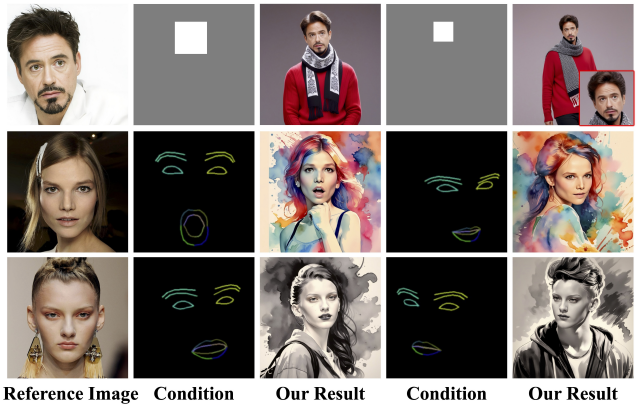

Figure 6: Fine control of location, size, pose and expression.

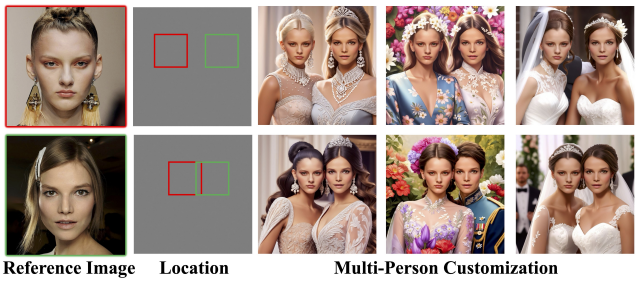

Figure 7: Results of multi-person customization.

ically, the location and size can be manipulated by manually setting the face bounding box. The pose and expression can also be controlled using the facial landmarks of other images with the desired pose and expression. The corresponding generation results are shown in Fig. 6.

Multi-Person Customization Given multiple reference images and their corresponding location guidance, RealisID can realize multi-person customization by separately integrating the injection information of different inputs in our local and global branches, as shown in Fig. 2(c). In this subsection, we study the multi-person customization in both non-overlapping and overlapping situations. From the results displayed in Fig. 7, it can be seen that our method maintains the identity independence of two persons, even when there is overlap in their face locations.

Conclusion

In this paper, we propose RealisID, a scale-robust and fine-controllable method for identity customization. It introduces a newly designed local branch to achieve the fine control of facial details and the identity fidelity across different face sizes. In addition, through another global branch, the overall harmony of the entire output image can be ensured, meanwhile the face location and the corresponding body and background layouts can be appropriately manipulated. Extensive experiments have verified the effectiveness of each individual branch, and the superiority of our whole RealisID framework over existing state-of-the-art methods.

Acknowledgments

This work was in part supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160604), the National Natural Science Foundation of China (Grant No. 62176194), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62306219), the Key Research and Development Program of Hubei Province (Grant No. 2023BAB083), the Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ- JYRC-2022-76, SKJC-2022-PTDX-031), the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031), Alibaba Group through Alibaba Research Intern Program.

References

Bai, Y.; Zhang, Y.; Ding, M.; and Ghanem, B. 2018. Sodmtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European conference on computer vision (ECCV)*, 206–221.

Cui, S.; Guo, J.; An, X.; Deng, J.; Zhao, Y.; Wei, X.; and Feng, Z. 2024. IDAdapter: Learning Mixed Features for Tuning-Free Personalization of Text-to-Image Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 950–959.

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*.

Guo, Z.; Wu, Y.; Chen, Z.; Chen, L.; and He, Q. 2024. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. *arXiv preprint arXiv:2404.16022*.

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

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kim, C.; Lee, J.; Joung, S.; Kim, B.; and Baek, Y.-M. 2024a. InstantFamily: Masked Attention for Zero-shot Multi-ID Image Generation. *arXiv preprint arXiv:2404.19427*.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024b. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8176–8185.

Li, S.; Fu, J.; Liu, K.; Wang, W.; Lin, K.-Y.; and Wu, W. 2024a. CosmicMan: A Text-to-Image Foundation Model for Humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6955–6965.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024b. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650.

Liang, C.; Ma, F.; Zhu, L.; Deng, Y.; and Yang, Y. 2024. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6400–6409.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

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*.

Noh, J.; Bae, W.; Lee, W.; Seo, J.; and Kim, G. 2019. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9725–9734.

Peng, X.; Zhu, J.; Jiang, B.; Tai, Y.; Luo, D.; Zhang, J.; Lin, W.; Jin, T.; Wang, C.; and Ji, R. 2024. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27080–27090.

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.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Wei, Z.; Su, Q.; Qin, L.; and Wang, W. 2024. MM-Diff: High-Fidelity Image Personalization via Multi-Modal Condition Integration. *arXiv preprint arXiv:2403.15059*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

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*.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

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.

Zhang, S.; Huang, L.; Chen, X.; Zhang, Y.; Wu, Z.-F.; Feng, Y.; Wang, W.; Shen, Y.; Liu, Y.; and Luo, P. 2024. FlashFace: Human Image Personalization with High-fidelity Identity Preservation. *arXiv preprint arXiv:2403.17008*.