A Dual-Branch Disentanglement Diffusion for ID-Attribute Conditional Face Generation

Anonymous authors
Paper under double-blind review

Abstract

Face identity customization, i.e., face generation with specified identity, has received increasing attention owing to its extensive applications in personalized content creation. Although existing methods achieve high consistency in identity with reference faces, they still struggle to precisely manipulate fine-grained facial attributes. We attribute this issue to the inherent entanglement of identity and attribute information, as well as the lack of attribute-specific supervision. Accordingly, to address this issue, we propose AttPortrait, a high-quality identity-attribute conditional face generation framework. Based on a foundational face diffusion model, we introduce an extra disentanglement branch alongside the conventional denoising branch during the training stage. This extra branch employs explicit attribute supervision to encourage the model to capture the attribute information from the text prompts, effectively disentangling the identity and attributes and achieving precise attribute manipulation with high identity consistency. Comprehensive experiments demonstrate that our method substantially improves attribute accuracy by 34%, while maintaining identity similarity on par with state-of-the-art methods and achieving competitive FID scores across both real and synthetic datasets.

1 Introduction

Recent advances in diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023), along with the availability of large-scale text-image pair datasets (Schuhmann et al., 2022; Changpinyo et al., 2021), have led to significant progress in text-to-image (T2I) generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Balaji et al., 2022; Li et al., 2024a; Chen et al., 2024a; Esser et al., 2024; BlackForestLab; Xie et al., 2025). Correspondingly, face identity customization, as one of the key applications of T2I generation, has attracted growing research interest and reached new heights. Typically, existing methods incorporate facial features (Xiao et al., 2024; Li et al., 2024c; Wang et al., 2024b; Huang et al., 2024; Zhang et al., 2023; Varanka et al., 2024) into diffusion models, enabling the specification of face identities when generating novel images.

Despite their success in specifying face identities, existing methods exhibit inherent limitations in specifying fine-grained facial attributes, such as hair style and age. As demonstrated in Figure 1, although the generated images have correct identities, their facial attributes do not match the given text prompts. In other words, the capacity to manipulate facial attributes via textual descriptions is substantially constrained when identity customization is required, which impedes the practical application of these approaches. We attribute this issue to the following two factors:

1) The attribute information is entangled with the identity (ID) information. Previous methods (Xiao et al., 2024; Valevski et al., 2023; Li et al., 2024c; Wang et al., 2024b; Huang et al., 2024) employ either general vision models such as CLIP image encoder (Radford et al., 2021) or specialized face recognition models such as ArcFace (Deng et al., 2019) to extract ID embeddings from the reference images for ID-conditional generation. However, these extractors usually fail to disentangle facial attribute information from the ID embeddings. In consequence, the attributes of the generated faces are often similar to those of the reference faces, even when the text prompts specify different attributes. For example, as can be seen in Figure 1,

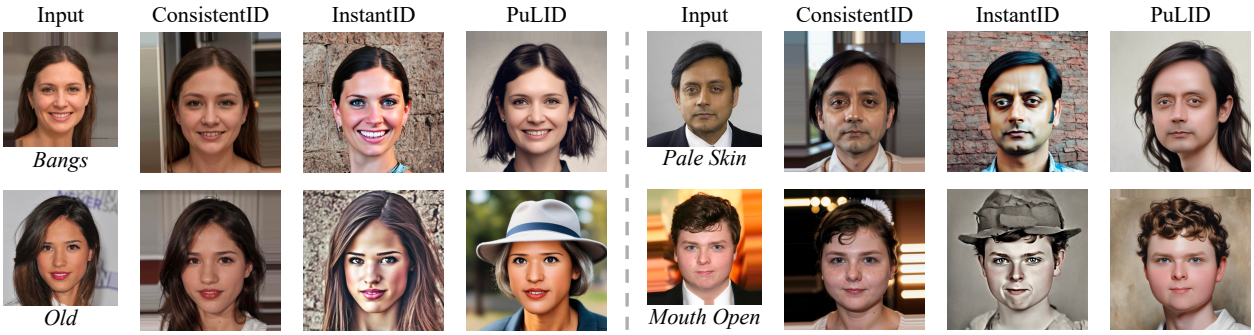

Figure 1: Given the reference images and the text description of the target attributes, the generated images of existing methods have the correct identity but fail to reflect the target attributes.

both the reference and generated images display nearly identical facial attributes, demonstrating strong entanglement between the attribute information and the ID information. *2) Previous methods lack explicit supervision to capture the attribute information from text prompts*, which is a more direct reason. Since there is no explicit supervision, most previous models tend to follow the entangled attributes in the ID embeddings rather than the attributes described in the text prompts when generating images.

In response to the aforementioned analysis, we propose **AttPortrait**, a face generation framework conditioned on both attributes and identity, which is capable of precise attribute control through textual descriptions while maintaining high identity consistency. Our framework consists of two branches: the denoising branch and the disentanglement branch, as shown in Figure 2. *The denoising branch* is a conditional diffusion model with classifier-free guidance (Ho & Salimans, 2022) that only employs an MSE loss, which is similar to existing methods (Cui et al., 2024; Xiao et al., 2024; Valevski et al., 2023; Chen et al., 2023). This branch is mainly responsible for generating high-quality and identity-consistent face images given the ID embeddings of reference faces. However, as discussed above, the attribute information is highly entangled with the ID information, which makes it difficult for the model to manipulate the attributes via text prompts. To solve this problem, we introduce an extra *disentanglement branch* with explicit supervision for attribute manipulation. Specifically, to ensure that the model faithfully captures the target attributes, we employ a pre-trained facial attribute predictor to assess the attributes present in the generated images. An attribute matching loss is then applied to minimize the discrepancy between the desired input attributes and those manifested in the generated images. Furthermore, we introduce a dual cross-attention module, which utilizes two parallel cross-attention blocks to separately incorporate the attribute information and the ID information. In this manner, the interference between attribute information and ID information is reduced, which improves both the image quality and the identity consistency.

Our contributions are summarized below:

- We reveal the entanglement of attribute and identity information in ID embeddings and quantify how this entanglement degrades attribute manipulation in diffusion models with comprehensive studies, providing critical guidance for future ID customization in diffusion models.
- We propose AttPortrait, an identity-attribute conditional generation framework, which adopts a denoising branch and a disentanglement branch to guide the generation process to ensure the correct attributes and identities. To our knowledge, AttPortrait is the first customization method that achieves satisfactory attribute manipulation with high identity consistency.
- Extensive experiments demonstrate that our AttPortrait significantly outperforms existing approaches in attribute accuracy, while maintaining identity consistency comparable to state-of-the-art approaches. Moreover, our model enables robust multiple attribute manipulation and partial zero-shot attribute manipulation with high controllability.

2 Related Work

2.1 Subject-driven Text-to-Image Diffusion Models

Owing to the powerful generative capability of text-to-image diffusion models (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Balaji et al., 2022; Li et al., 2024a; Chen et al., 2024a; Esser et al., 2024; BlackForestLab; Xie et al., 2025), subject-driven methods have attracted increasing research attention. These methods aim to adapt large-scale diffusion models to synthesize images conditioned on specified subjects. Based on whether fine-tuning is required when testing a new subject, these methods can be divided into two types: test-time fine-tuning methods (Ruiz et al., 2023; Gal et al., 2022; Dong et al., 2022; Kumari et al., 2023; Smith et al., 2023; Wang et al., 2024a; Yuan et al., 2023) and test-time free-tuning methods (Wei et al., 2023; Ye et al., 2023; Gal et al., 2023; Shi et al., 2024; Ma et al., 2024; He et al., 2024; Zhang et al., 2024). The test-time fine-tuning methods optimize the diffusion model at test time for each individual subject using one or more reference images. However, per-subject optimization is time-consuming and limits the applications. Test-time free-tuning methods, also known as encoder-based methods, typically incorporate diverse subject features into diffusion models with various fusion mechanisms. Specifically, in order to better capture subject-specific details, ELITE (Wei et al., 2023) inserts global subject features into the textual embedding and incorporates the local subject feature with an additional cross-attention, which inspires many subsequent works (Ye et al., 2023; Ma et al., 2024; Xiao et al., 2024; Shi et al., 2024). IP-adapter (Ye et al., 2023) is another representative subject-driven method, which introduces a lightweight adapter along with a separate cross-attention to fuse the subject information into the model and only finetune the lightweight adapter during training. To achieve better subject consistency, Subject-Diffusion (Ma et al., 2024) incorporates more subject information, including segmentations and bounding boxes, into the diffusion model and sets an adapter between self-attention and cross-attention.

2.2 ID Customization in Diffusion Models

A prominent direction within subject-driven methods is ID customization (Cui et al., 2024; Peng et al., 2024; Xiao et al., 2024; Huang et al., 2024; Li et al., 2024c; Wang et al., 2024b; Chen et al., 2024b; Wu et al., 2024b; Wang et al., 2024a; Yuan et al., 2023; Lin et al., 2025; Xu et al., 2025; Borse et al., 2025), which aims to generate images with specified identities. For example, CelebBasis (Yuan et al., 2023) leverages the inherent knowledge of celebrities in pretrained T2I model by projecting the embeddings of celebrity names to construct a "celebrity basis". Building upon this basis, the model can represent new identities with only a small number of learnable parameters, enabling flexible and controllable facial identity manipulation. Following this idea, StableIdentity (Wang et al., 2024a) employs adaptive instance normalization to project face identity into the constructed celebrity embedding space. Other works focus on incorporating additional facial conditioning signals to generate faces more precisely aligned with reference images. For example, InstantID (Wang et al., 2024b) utilizes a modified ControlNet to incorporate facial landmarks as an extra condition, providing finer-grained control over generated facial structures. ConsistentID (Huang et al., 2024) employs a fine-grained multimodal feature extractor to capture detailed features from different facial regions, along with corresponding textual descriptions, to generate more precise facial details. Separately, several methods utilize multiple reference images of the same identity to improve identity representation. PhotoMaker (Li et al., 2024c) combines multiple reference images to improve the representation of a single identity, while IDAdapter (Cui et al., 2024) also adopts a multi-image design and further inserts adapters between attention blocks to better fuse identity features.

FastComposer (Xiao et al., 2024) integrates multi-identity features with textual embeddings. It employs localized attention control guided by face segmentation masks, ensuring that each identity is accurately represented in its designated region. Building on FastComposer, PortraitBooth (Peng et al., 2024) further improves the localized attention control with a truncated cross-attention mechanism, which additionally facilitates emotion control. MultiID (Lin et al., 2025) leverages the localized attention control and incorporates DDIM-inversion (Song et al., 2020) features extracted from reference images along with depth-guided spatial control, enabling fully training-free multi-identity generation.

To further improve identity fidelity, several existing methods (Chen et al., 2023; Peng et al., 2024; Cui et al., 2024; Gal et al., 2024; Guo et al., 2024) introduce optimization-based techniques. Early approaches like PortraitBooth, IDAdapter, and PhotoVerse (Chen et al., 2023) directly generate images within one step at

an early diffusion timestep and then calculate ID loss between the generated images and the reference images. However, such generated images are often noisy and low-quality, which reduces the effectiveness of the ID loss. To solve this issue, LCM-lookahead (Gal et al., 2024) and PuLID (Guo et al., 2024) employ fast sampling methods (Luo et al., 2023; Ren et al., 2024; Lin et al., 2024) to generate images of higher quality compared to previous approaches, which enables more reliable computation of ID loss. WithAnyone (Xu et al., 2025) adds a contrastive ID loss combined with keypoint alignment from real images, further enhancing both the identity consistency and the discriminability. Unlike these supervision-based approaches, Ar2Can (Borse et al., 2025) leverages reinforcement learning with GRPO (Shao et al., 2024) on large-scale synthetic data to optimize multi-ID generation.

Although these approaches demonstrate improved identity fidelity, most of them still struggle to accurately manipulate fine-grained facial attributes. To address this issue, some methods (Li et al., 2024b; Parihar et al., 2024) attempt to integrate an editable W+ space like StyleGAN (Karras et al., 2019) into diffusion models. However, these methods exhibit limited capability in controlling complex attributes and often fail to preserve fine-grained details, leading to degraded identity fidelity in many cases.

3 Methods

3.1 Preliminaries

Stable Diffusion The Stable Diffusion model (Rombach et al., 2022) consists of three components: a CLIP (Radford et al., 2021) text encoder, a Variational Autoencoder (VAE) (Kingma et al., 2013), and a U-Net (Ronneberger et al., 2015). In the training phase, the VAE compresses the image $\mathbf{x}$ into the latent code $\mathbf{z}$. The latent code is then perturbed by Gaussian noise ϵ. Subsequently, the U-Net $\epsilon_\theta(\cdot)$ is optimized to denoise the noisy latent code $\mathbf{z}_t$, conditioned on the CLIP text embedding $\mathbf{e}$, with the following objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}\sim\text{VAE}(\mathbf{x}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,\mathbf{e}}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{e})\|^2\right]. \tag{1}$$

In the inference phase, a Gaussian noise $\mathbf{z}_T$ is iteratively denoised by the U-Net to obtain a clean latent $\mathbf{z}_0$, which is then decoded into the final image by the VAE decoder.

Embedding-Conditioning Cross-Attention Mechanism In the Stable Diffusion model, the U-Net uses multiple cross-attention (Vaswani et al., 2017) layers to incorporate the textual information, guiding the generation process according to the text prompt. Specifically, the text embedding $\mathbf{e}$ is projected to the key $K_\mathbf{e} = W_k\mathbf{e}$ and value $V_\mathbf{e} = W_v\mathbf{e}$, and the noisy latent code $\mathbf{z}_t$ is projected to the query $Q = W_q\mathbf{z}_t$. The cross-attention mechanism is formulated as follows:

$$Attn(Q, K_\mathbf{e}, V_\mathbf{e}) = \text{Softmax}\left(\frac{QK_\mathbf{e}^T}{\sqrt{d}}\right)V_\mathbf{e}, \tag{2}$$

where d is the dimension of $K_\mathbf{e}$, $V_\mathbf{e}$, and Q. In this work, we employ cross-attention to incorporate attribute information and identity information.

3.2 ID-Attribute Conditional Face Generation

Task Definition Let $\mathbf{I}^\mathbf{a}_{\text{ref}}$ denote a reference image with the target identity, and $\mathbf{a} = [\,a_1, \cdots, a_n\,]$ denote the corresponding attributes where each component a_i is one kind of attribute, such as "Bangs", "Eyeglasses", and "Gender". Let $\mathbf{b} = [\,b_1, \cdots, b_n\,]$ denote the target attributes. Our objective is to develop a face diffusion model G capable of generating faces with the target attributes and identities, which is formulated as follows:

$$\mathbf{I}^\mathbf{b}_{\text{gen}} = G(\mathbf{b}, \mathbf{I}^\mathbf{a}_{\text{ref}}), \tag{3}$$

where the generated image $\mathbf{I}^\mathbf{b}_{\text{gen}}$ is expected to faithfully exhibit the target attributes $\mathbf{b}$ while maintaining the same identity as the reference image $\mathbf{I}^\mathbf{a}_{\text{ref}}$.

To clarify, we use ArcFace (Deng et al., 2019) to extract the feature $\phi(\mathbf{I}^\mathbf{a}_{\text{ref}})$ from the reference image. This feature, following Arc2Face (Paraperas Papantoniou et al., 2024), is further mapped into an embedding $\mathbf{e}_{\text{id}}$, which is used as a condition input to the diffusion model. In the rest of the paper, we refer to $\phi(\mathbf{I}^\mathbf{a}_{\text{ref}})$ as the "face recognition feature" and refer to $\mathbf{e}_{\text{id}}$ as the "ID embedding".

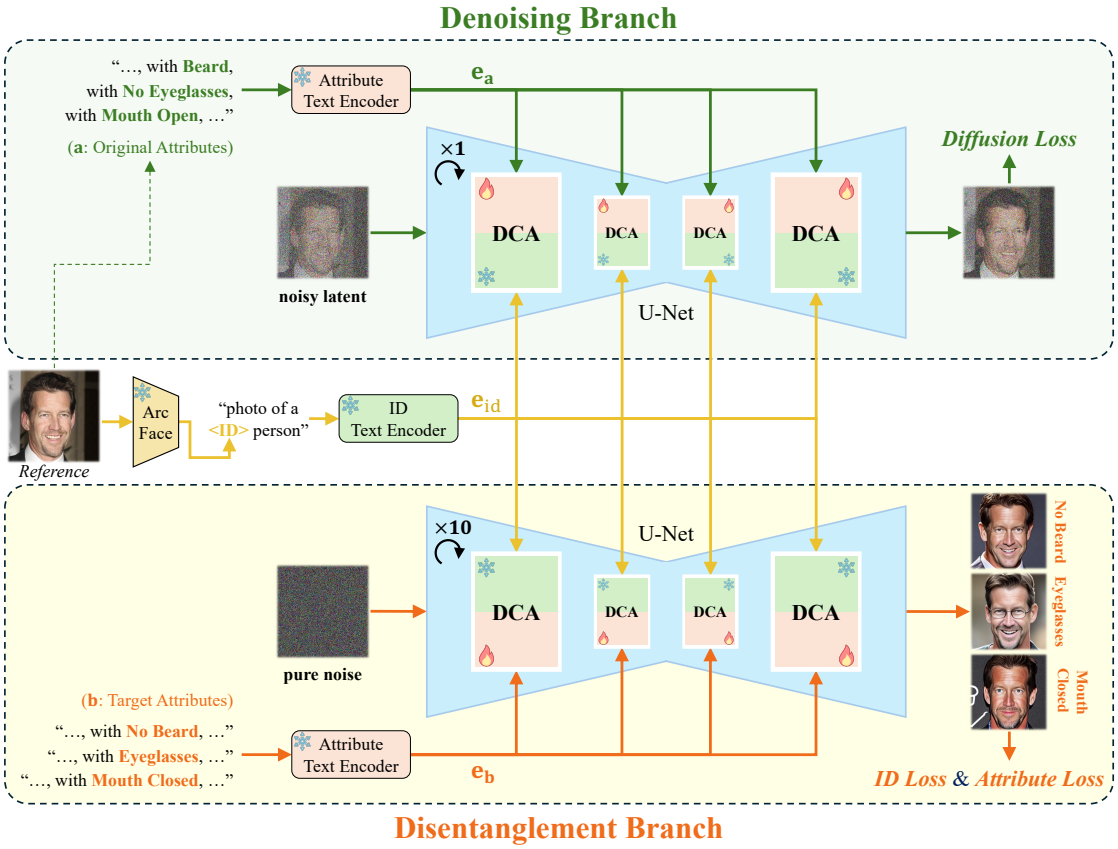

Figure 2: The overall framework of our AttPortrait. The upper half of this framework illustrates the denoising branch, which uses the original attribute embedding $\mathbf{e_a}$, containing attribute information present in the reference face. The lower part shows the disentanglement branch, which uses the target attribute embedding $\mathbf{e_b}$, containing target attribute information not present in the reference face. Both branches employ the identical ID embedding $\mathbf{e}_{\text{id}}$ and share the same attribute text encoder and U-Net. For clarity, textual descriptions are used to represent tokenized text embeddings.

Denoising Branch As shown in the upper branch of Figure 2, to generate identity-consistent and high-quality face images, we employ the classifier-free guidance (Ho & Salimans, 2022) diffusion model conditioned on the ID embeddings. Following Arc2Face (Paraperas Papantoniou et al., 2024), we first map the reference image $\mathbf{I}^{\mathbf{a}}_{\text{ref}}$ into the ID embedding $\mathbf{e}_{\text{id}}$. Besides, we employ CLIP to map the text description of the original attributes $\mathbf{a}$ into the attribute embedding $\mathbf{e_a}$. Using the ID embedding and the original attribute embedding as a joint condition, our diffusion model learns to predict the noise using an MSE loss according to Eq. (1), which is formulated as follows:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{e_a}, \mathbf{e}_{\text{id}})\|^2\right]. \tag{4}$$

Identity-Attribute Disentanglement Branch Although the denoising branch can generate identity-consistent faces, it fails to precisely manipulate the attributes due to the inherent entanglement between attribute and identity information within the ID embeddings. Specifically, the ID embedding $\mathbf{e}_{\text{id}}$ generally covers most of the attribute information, which makes the model directly ignore the attribute information from $\mathbf{e_a}$. As a consequence, we cannot effectively control the attributes by modifying the attribute input.

To address this issue, as shown in the lower part of Figure 2, we introduce a disentanglement branch, which applies explicit supervision to encourage the model to capture the information from the attribute embeddings. Let $\mathbf{b}$ denote the target attributes, which is distinct from the original attributes $\mathbf{a}$ of the reference image $\mathbf{I}^{\mathbf{a}}_{\text{ref}}$. Given the target attributes and the reference identity image, in this branch, the diffusion model generates

the final image $\mathbf{I}^{\mathbf{b}}_{\text{gen}} = G(\mathbf{b}, \mathbf{I}^{\mathbf{a}}_{\text{ref}})$ through 10 sampling steps, rather than just predicting the noise of one specific step. Further, to make the generated image accurately exhibit the target attributes, we apply an attribute matching loss as follows:

$$\mathcal{L}_{\text{att}} = \sum_{i=1}^{n} -b_i \log \mathcal{C}_i(\mathbf{I}^{\mathbf{b}}_{\text{gen}}) - (1 - b_i) \log(1 - \mathcal{C}_i(\mathbf{I}^{\mathbf{b}}_{\text{gen}})), \tag{5}$$

where $\mathcal{C}$ is a pretrained multi-attribute classifier and $\mathcal{C}_i$ is the prediction of the i^{th} attribute. The objective in Eq. (5) encourages the generated image $\mathbf{I}^{\mathbf{b}}_{\text{gen}}$ to be classified as possessing the target attributes $\mathbf{b}$.

Besides, to avoid identity drift, we incorporate an auxiliary identity loss, formulated as follows:

$$\mathcal{L}_{\text{id}} = 1 - \frac{\phi(\mathbf{I}^{\mathbf{b}}_{\text{gen}}) \cdot \phi(\mathbf{I}^{\mathbf{a}}_{\text{ref}})}{\|\phi(\mathbf{I}^{\mathbf{b}}_{\text{gen}})\| \|\phi(\mathbf{I}^{\mathbf{a}}_{\text{ref}})\|}, \tag{6}$$

where ϕ is the face recognition feature extracted from ArcFace (Deng et al., 2019). The objective in Eq. (6) encourages a high identity similarity between the generated image and the corresponding reference image.

As mentioned above, we perform 10 sampling steps in this branch to generate the final image $\mathbf{I}^{\mathbf{b}}_{\text{gen}}$. However, backpropagation through 10 diffusion steps requires a large amount of GPU memory and a large number of FLOPs. To alleviate the computational requirements, we adopt DRTune (Wu et al., 2024a), which only maintains the gradient for a small subset of sampling steps while stopping the gradient of the U-Net input for the rest steps. In this manner, the attribute loss and identity loss can be effectively backpropagated through our model.

Identity-Attribute Dual Cross-Attention To effectively capture both the attribute and identity information, we design a dual cross-attention (DCA) module. Specifically, as shown in Figure 2, the attribute embedding and the ID embedding are independently processed by distinct cross-attention blocks and then summed, formulated as follows:

$$Q^*_{\mathbf{a}} = Attn(Q, K_{\mathbf{e}_{\text{id}}}, V_{\mathbf{e}_{\text{id}}}) + Attn(Q, K_{\mathbf{e_a}}, V_{\mathbf{e_a}}), \tag{7}$$

$$Q^*_{\mathbf{b}} = Attn(Q, K_{\mathbf{e}_{\text{id}}}, V_{\mathbf{e}_{\text{id}}}) + Attn(Q, K_{\mathbf{e_b}}, V_{\mathbf{e_b}}), \tag{8}$$

where $Attn$ is defined in Eq. (2), $Q^*_{\mathbf{a}}$ and $Q^*_{\mathbf{b}}$ correspond to the DCA outputs in the denoising branch and disentanglement branch respectively. In this manner, the interference between the attribute and identity information is effectively mitigated, thereby further enhancing both the image quality and the identity consistency of the generated images.

Objective Function The full objective function is given by:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{att}} \mathcal{L}_{\text{att}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}}, \tag{9}$$

where λ_{att} and λ_{id} are the hyperparameters to modulate the strength of the attribute loss and identity loss. In this paper, both λ_{att} and λ_{id} are set to 0.01.

4 Experiments

4.1 Setup

Datasets Our training dataset consists of about 500K highest-quality faces selected from LAION-Face (Schuhmann et al., 2022) and their corresponding attribute labels predicted by our pretrained attribute predictor, including "Bald", "Young", "Male", "Bangs", "Black Hair", "Blond Hair", "Bushy Eyebrows", "Eyeglasses", "Mouth Slightly Open", "No Beard" and "Pale Skin". Each attribute is encoded as 1 if present and 0 if absent. For evaluation, we use two datasets with attribute labels, one is the 500 synthetic faces from Karras et al. (2020), referred to as Synth-test, and the other is 5K faces randomly selected from CelebA test dataset (Liu et al., 2015), referred to as CelebA-test.

Table 1: **Quantitative comparison against existing methods on CelebA-test (Liu et al., 2015) and Synth-test (Karras et al., 2020).** We report mean ID similarity (ID Sim), mean attribute accuracy (Att Acc), and FID. Here we do not report the attribute accuracy of Arc2Face (Paraperas Papantoniou et al., 2024) because it fails to accept the attribute text prompts.

Method	CelebA-test			Synth-test		
	ID Sim ↑	Att Acc ↑	FID ↓	ID Sim ↑	Att Acc ↑	FID ↓
Arc2Face	**0.795**	-	12.43	<u>0.746</u>	-	9.75
ConsistentID	0.499	0.094	**7.99**	0.475	0.074	**3.98**
PhotoMaker	0.234	<u>0.525</u>	<u>9.01</u>	0.249	<u>0.357</u>	5.21
InstantID	0.764	0.171	19.46	0.699	0.163	27.96
PuLID	0.631	0.371	9.56	0.609	0.322	<u>5.16</u>
AttPortrait (ours)	<u>0.793</u>	**0.869**	9.32	**0.768**	**0.911**	7.33

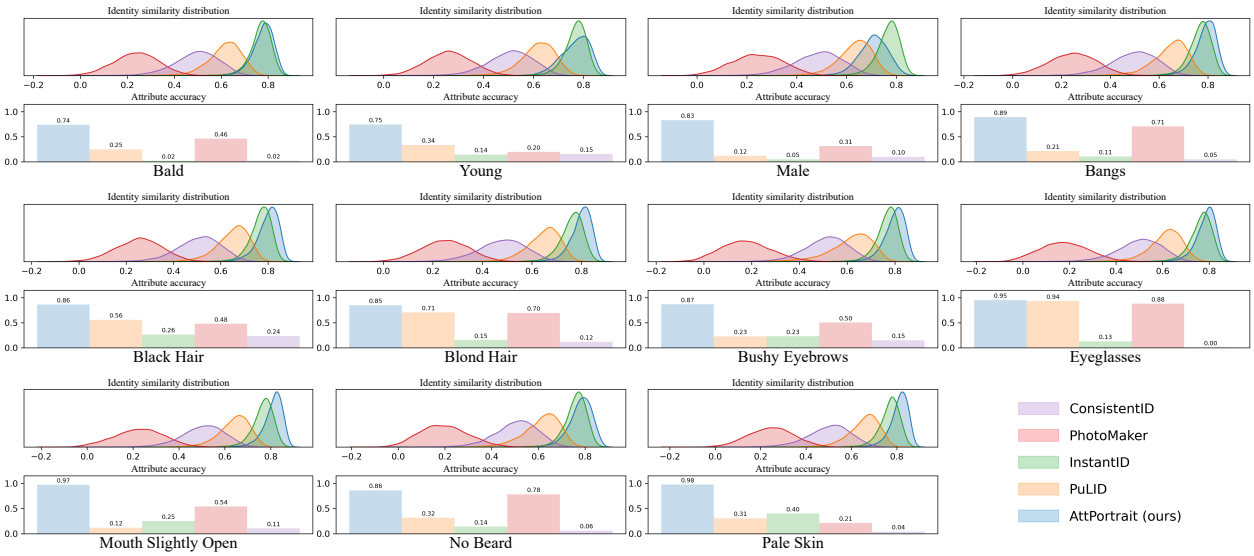

Figure 3: Comparison of identity consistency and attribute accuracy of different methods across diverse facial attributes on CelebA-test. For each attribute, the top subplot shows the distribution of identity similarity between the generated images and the reference images, while the bottom subplot reports the corresponding attribute accuracy. Different colors denote different methods. Here, we exclude Arc2Face since it does not support attribute text prompts.

Implement Details Our AttPortrait is built upon the SD-v1.5 architecture and uses two text encoders, namely ID text encoder and attribute text encoder, as illustrated in Figure 2. The ID text encoder and the U-Net are initialized from Arc2Face (Paraperas Papantoniou et al., 2024), and the attribute text encoder and the VAE are initialized from SD-v1.5. During the training phase, the denoising branch and the disentanglement branch utilize the shared U-Net and text encoders. Only the key and the value matrices in the cross-attention layers of attribute embeddings are trained with AdamW (Loshchilov & Hutter, 2017). In each iteration, the disentanglement branch generates images using a classifier-free guidance scale (Ho & Salimans, 2022) of 3.0 with 10 DPM-Solver (Lu et al., 2022) sampling steps, and for 8 randomly chosen steps we stop gradients of the U-Net input. During the inference phase, images are generated using the same classifier-free guidance scale with 25 DPM-Solver sampling steps. Our model is trained for 1 epoch on 8 NVIDIA A100 GPUs, with a constant learning rate of 1e-6 and a total batch size of 8.

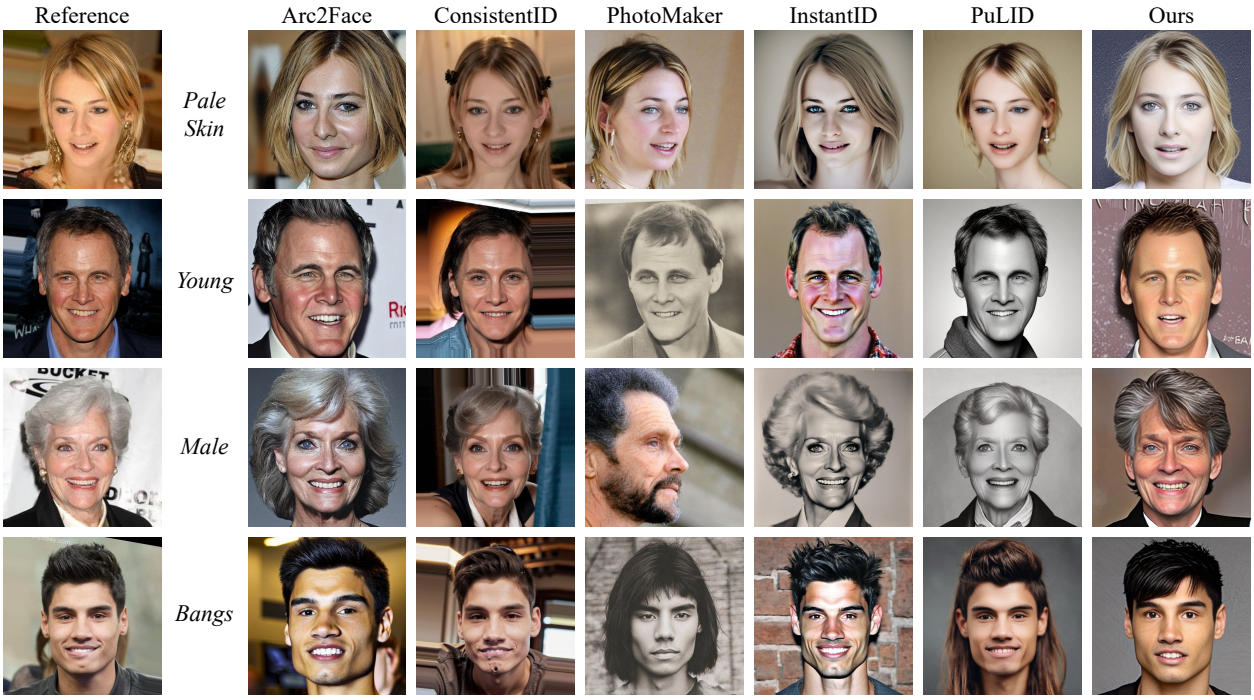

Figure 4: Qualitative comparison against existing works. We compare our method against Arc2Face, ConsistentID, PhotoMaker, InstantID, and PuLID across four distinct identities and four diverse attributes. For each row, we input the reference image and the text of the attribute displayed on its right into these models. Since Arc2Face does not support attribute text prompts, we only compare generated image quality.

4.2 Evaluation Protocols

Identity and FID Evaluation Protocol In this protocol, *Identity (ID) Similarity* and *FID* (Heusel et al., 2017) are employed to assess visual fidelity and identity consistency. Specifically, for each image, we randomly modify one attribute and generate a sample. This procedure is repeated twice. Then we report (a) the FID scores between the generated and reference sets, and (b) the average ID similarity for each generated–reference pair.

Attribute Evaluation Protocol In this protocol, we use *Attribute Accuracy* to evaluate whether the generated images correctly reflect the target attributes. Specifically, for each attribute, we modify it and generate two samples per reference image. Then we use the pretrained attribute predictor to predict the attribute of generated samples and compute the attribute accuracy, which is defined as the proportion of generated images where the target attributes are correctly present. Finally, we report the mean attribute accuracy over all attributes to show the overall performance.

4.3 Comparison with Existing Methods

Quantitative comparison We compare our AttPortrait with Arc2Face (Paraperas Papantoniou et al., 2024), ConsistentID (Huang et al., 2024), PhotoMaker (Li et al., 2024c), InstantID (Wang et al., 2024b), and PuLID (Guo et al., 2024) on our evaluation protocols. As shown in Table 1, AttPortrait achieves the highest attribute accuracy, with gains of **55.4**% on Synth-test and **34.4**% on CelebA-test. Our method also achieves the best identity similarity on Synth-test and the second-best on CelebA-test while maintaining comparable FID scores across both datasets.

To gain deeper insights beyond the overall metrics, we perform an attribute-level analysis that jointly evaluates identity consistency and attribute accuracy for each attribute, as shown in Figure 3. The results show that AttPortrait achieves consistently higher attribute accuracy across all evaluated attributes. Some prior

Figure 5: Visual results on multiple attribute manipulation.

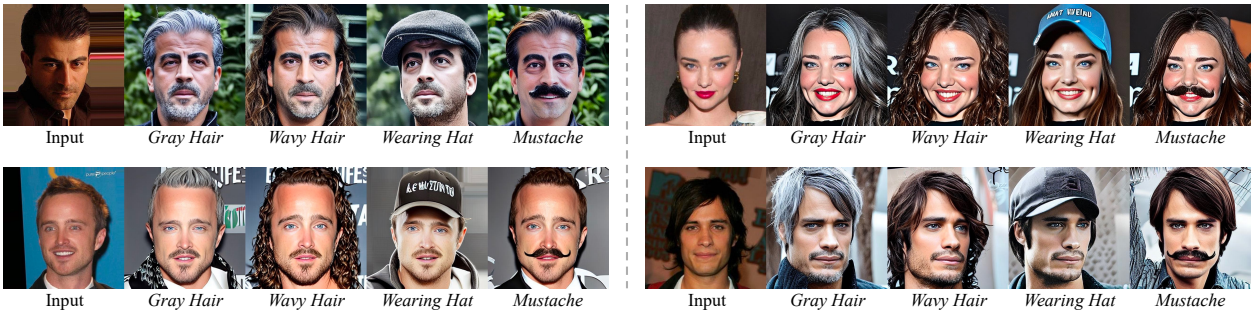

Figure 6: Visual results on zero-shot attribute manipulation.

methods achieve comparable attribute accuracy on specific attributes, but exhibit a noticeable reduction in identity consistency, as illustrated by PuLID under the "Eyeglasses" attribute. Moreover, AttPortrait maintains strong identity consistency when manipulating each attribute. Although certain methods slightly outperform ours in identity consistency on a few attributes, this often comes at the cost of severely compromised attribute control, as observed for InstantID under the "Male" attributes. Overall, these results demonstrate that AttPortrait effectively manipulates attributes while maintaining identity consistency, surpassing prior approaches.

Qualitative comparison Beyond quantitative metrics, AttPortrait demonstrates superior visual results compared to existing approaches. As shown in Figure 4, other methods exhibit shortcomings to varying degrees. Arc2Face and ConsistentID generate high-quality faces but fail to control attributes. PhotoMaker is capable of controlling certain attributes, such as gender and hairstyle (3rd and 4th rows), but the outputs exhibit low identity consistency. InstantID often generates less realistic faces (2nd and 4th rows) and shows limited ability to control attributes. PuLID can handle simple attributes, but it still has difficulty manipulating more abstract attributes such as age, gender (2nd and 3rd rows). In contrast, all images generated by AttPortrait exhibit high identity fidelity and accurately reflect the attributes specified in the text prompts, even under challenging cases like gender or age changes. More examples can be found in Appendix B.

4.4 Extensive Visual Results

Multiple Attribute Manipulation We further evaluate the performance of AttPortrait in scenarios involving simultaneous manipulation of multiple facial attributes, which imposes higher demands on attribute manipulation and identity maintaining. Figure 5 illustrates several cases where multiple facial attributes are successfully manipulated while maintaining identity consistency, demonstrating the robustness of our method. More results can be found in Appendix B.

Zero-Shot Attribute Manipulation Our method supports zero-shot control of novel attributes. Even though certain attributes such as "Gray Hair", "Wavy Hair", "Wearing Hat", and "Mustache" are never seen

Table 2: **Effect of each component, evaluated on CelebA-test† and Synth-test.** The CelebA-test† contains 1K randomly selected images from CelebA-test (Liu et al., 2015). For clarity, we denote "Disent." as "Disentanglement Branch", "Denoise" as "Denoising Branch", "Target" as "Target Attributes", and "Att. Loss" as "Attribute Loss".

Ablations	Disent.			Denoise	DCA	CelebA-test†			Synth-test		
	Target	ID Loss	Att. Loss			ID Sim ↑	Att. Acc ↑	FID ↓	ID Sim ↑	Att. Acc ↑	FID ↓
w/o Target Attributes	✗	✓	✓	✓	✓	0.810	0.481	13.31	0.788	0.520	9.75
w/o ID Loss	✓	✗	✓	✓	✓	0.539	**0.959**	13.56	0.430	**0.965**	9.88
w/o Attribute Loss	✓	✓	✗	✓	✓	**0.829**	0.159	15.61	**0.814**	0.169	11.79
w/o Disentanglement Branch	✗	✗	✗	✓	✓	0.774	0.207	20.65	0.735	0.212	18.20
w/o Denoising Branch	✓	✓	✓	✗	✓	0.805	0.932	24.99	0.772	0.949	26.17
w/o DCA	✓	✓	✓	✓	✗	0.694	0.886	28.35	0.660	0.902	28.74
Full	✓	✓	✓	✓	✓	0.792	0.866	**9.18**	0.768	0.911	**7.33**

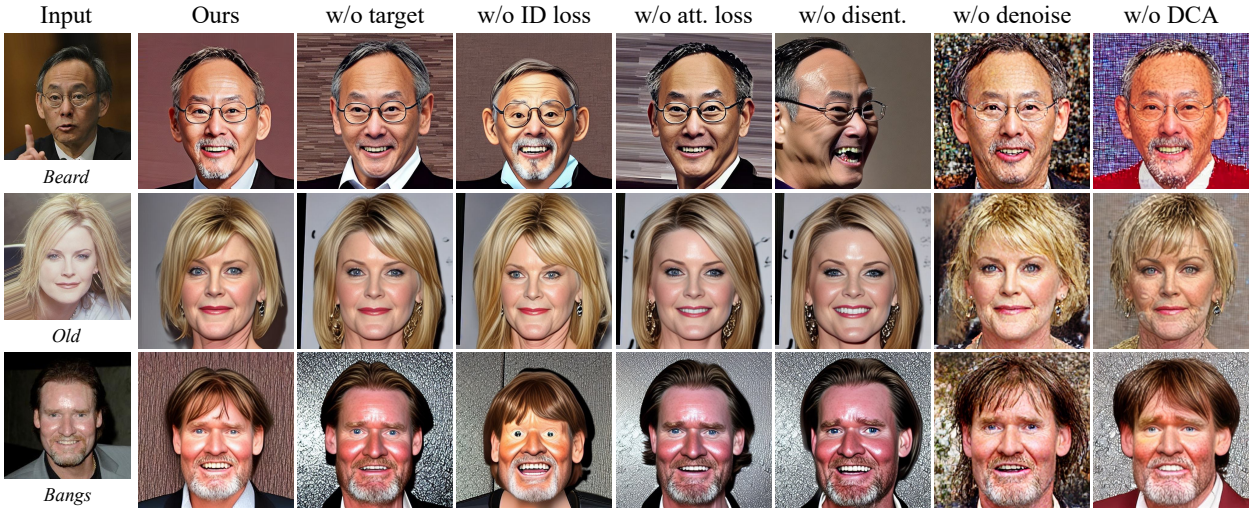

Figure 7: Qualitative comparison for ablation studies. For clarity, we denote "w/o target" as "w/o target attributes", "w/o att. loss" as "w/o attribute loss", "w/o disent." as "w/o disentanglement branch", and "w/o denoise" as "w/o denoising branch". Zoom in for better observation.

during training, our model is able to generate identity-consistent faces reflecting these attributes. Figure 6 presents several examples where these unseen attributes are correctly rendered, demonstrating the strong generalization capability of our model.

4.5 Ablation Study

Effect of Target Attributes In the disentanglement branch, target attributes that differ from the original attributes are used as text prompts. As shown in Table 2, replacing target attributes with original ones reduces attribute accuracy on CelebA-test† from 0.866 to 0.481 and on Synth-test from 0.911 to 0.520. Figure 7 also illustrates that the model fails to capture attributes like age and bangs without target attributes. These results demonstrate the effectiveness of target attributes for precise attribute manipulation.

Effect of ID Loss Removing the ID loss causes ID similarity to drop from 0.792 to 0.539 on CelebA-test† and from 0.768 to 0.430 on Synth-test, as shown in Table 2. Figure 7 shows that some generated images present noticeable distortions with significant drops in identity fidelity, indicating that the model fails to preserve subject-specific facial characteristics. These results clearly highlight the critical role of the ID loss in preventing identity drift during attribute manipulation.

Effect of Attribute Loss As shown in Table 2, removing the attribute loss leads to a drastic drop in attribute accuracy, from 0.866 to 0.159 on CelebA-test† and from 0.911 to 0.169 on Synth-test, while identity similarity slightly increases. Figure 7 shows that the generated images fail to exhibit the specified attributes. This indicates that, without attribute supervision, the model tends to preserve the original facial appearance, underscoring the necessity of the attribute loss for effective attribute manipulation.

Effect of Disentanglement Branch As shown in Table 2, removing this branch causes attribute accuracy to drop from 0.866 to 0.207 on CelebA-test† and from 0.911 to 0.212 on Synth-test, with a substantial increase in FID from 9.18 to 20.65 and from 7.33 to 18.20, respectively. Figure 7 further shows that the generated images fail to reflect the specified attributes and exhibit noticeable distortions. These results highlight the essential role of the disentanglement branch in attribute manipulation while preserving facial quality.

Effect of Denoising Branch As reported in Table 2, removing the denoising branch increases FID scores from 9.18 to 24.99 on CelebA-test† and from 7.33 to 26.17 on Synth-test, indicating significant quantitative degradation. Figure 7 shows that the outputs exhibit severe noise and visibly lower image quality. These results highlight the crucial role of the denoising branch in generating high-quality, realistic face images.

Effect of DCA We evaluate the impact of our dual cross-attention by comparing it with a baseline where the ID and attribute embeddings are concatenated and injected through a single cross-attention block. As shown in Table 2, replacing it with a single block lowers ID similarity from 0.792 to 0.694 on CelebA-test† and from 0.768 to 0.660 on Synth-test, accompanied by a substantial increase in FID from 9.18 to 28.35 and from 7.33 to 28.74 respectively. Generated images also show more pronounced noise, as illustrated in Figure 7. These results show that our dual cross-attention effectively maintains image quality and identity fidelity when manipulating attributes.

Understanding Trade-offs in Ablation Results A careful examination of Table 2 and Figure 7 reveals that different ablation settings bias the model towards specific aspects of the task, rather than uniformly improving performance. Notably, several ablated variants achieve higher scores than the full model on individual metrics. For example, removing the attribute loss or target attributes leads to higher identity similarity, as attribute manipulation is largely suppressed. Similarly, removing the ID loss, DCA, or denoising branch can lead to higher attribute accuracy than the full model, but this comes at the cost of noticeable identity degradation or a significant decline in visual quality. These observations show that when any of these components is removed, the balance among identity fidelity, attribute controllability, and generation quality is disrupted. Overall, the full model consistently achieves a balanced performance across both datasets, combining strong identity consistency, accurate attribute manipulation, and the best overall generation quality, as supported by both quantitative metrics and qualitative results.

5 Conclusion and Limitations

In this paper, we present AttPortrait, a face generation framework that precisely manipulates facial attributes through text prompts while maintaining high identity consistency. Existing methods fail to follow given attributes due to the entanglement of identity and attribute information in ID embeddings. However, we effectively overcome this limitation by employing a novel dual-branch framework with explicit attribute supervision. Extensive experiments clearly demonstrate that AttPortrait significantly outperforms prior methods in attribute accuracy. We hope our work can provide new inspiration for future research on personalized face generation.

While AttPortrait excels at identity consistency and precise attribute control, it has two main limitations. Firstly, our method exhibits a slight decline in image realism compared to Arc2Face (Paraperas Papantoniou et al., 2024). This may be caused by the attribute matching loss, which encourages the model to align with the given attribute, but also introduces a mild adversarial effect. Secondly, AttPortrait cannot generate full-body images with diverse backgrounds and styles, as it is trained exclusively on face-centric data. In the future, we plan to incorporate datasets with diverse backgrounds and contexts to enable simultaneous control over attributes, scenes, and styles.

References

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022.

BlackForestLab. Flux.1. https://blackforestlabs.io/flux-1/.

Shubhankar Borse, Phuc Pham, Farzad Farhadzadeh, Seokeon Choi, Phong Ha Nguyen, Anh Tuan Tran, Sungrack Yun, Munawar Hayat, and Fatih Porikli. Ar2can: An architect and an artist leveraging a canvas for multi-human generation. arXiv preprint arXiv:2511.22690, 2025.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3558–3568, 2021.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In European Conference on Computer Vision, pp. 74–91. Springer, 2024a.

Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793, 2023.

Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreamidentity: enhanced editability for efficient face-identity preserved image generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 1281–1289, 2024b.

Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. 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, pp. 950–959, 2024.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019.

Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. arXiv preprint arXiv:2211.11337, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG), 42(4):1–13, 2023.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. In European Conference on Computer Vision, pp. 322–340. Springer, 2024.

Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. Advances in neural information processing systems, 37:36777–36804, 2024.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. arXiv preprint arXiv:2408.05939, 2024.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. IEEE transactions on image processing, 28(11):5464–5478, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. arXiv preprint arXiv:2404.16771, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110–8119, 2020.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1931–1941, 2023.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024a.

Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2187–2196, 2024b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8640–8650, 2024c.

Jiawei Lin, Guanlong Jiao, and Jianjin Xu. A training-free approach for multi-id customization via attention adjustment and spatial control. arXiv preprint arXiv:2511.20401, 2025.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems, 35:5775–5787, 2022.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556, 2023.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In ACM SIGGRAPH 2024 Conference Papers, pp. 1–12, 2024.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR), 54(1):1–41, 2021.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In Proceedings of the European Conference on Computer Vision (ECCV), 2024.

Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In European Conference on Computer Vision, pp. 469–487. Springer, 2024.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27080–27090, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. arXiv preprint arXiv:2404.13686, 2024.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pp. 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems, 35:25278–25294, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8543–8552, 2024.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. arXiv preprint arXiv:2304.06027, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In SIGGRAPH Asia 2023 Conference Papers, pp. 1–10, 2023.

Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. arXiv preprint arXiv:2407.20175, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 2555–2563, 2023.

Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. arXiv preprint arXiv:2401.15975, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519, 2024b.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9168–9178, 2021.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15943–15953, 2023.

Mika Westerlund. The emergence of deepfake technology: A review. Technology innovation management review, 9(11), 2019.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In European Conference on Computer Vision, pp. 108–124. Springer, 2024a.

Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In European Conference on Computer Vision, pp. 279–296. Springer, 2024b.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision, pp. 1–20, 2024.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Towards controllable and id consistent image generation. arXiv preprint arXiv:2510.14975, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. arXiv preprint arXiv:2306.00926, 2023.

Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. arXiv preprint arXiv:2304.03117, 2023.

Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. Flashface: Human image personalization with high-fidelity identity preservation. arXiv preprint arXiv:2403.17008, 2024.

Appendix

A Additional Experiment Details

A.1 Details of Dataset Construction

Our training dataset is selected from LAION-Face (Schuhmann et al., 2022) and contains approximately 500,000 face images. We start by using InsightFace (Deng et al., 2019) to detect and align faces in the original LAION-Face images. Next, we filter these faces and retain those with size larger than 133×133 pixels, face score above 0.85, and CLIP-IQA+ score (Wang et al., 2023) above 0.645, resulting in approximately 500,000 high-quality face images. We then employ GFPGAN v1.4 (Wang et al., 2021) to further enhance the image quality. Finally, we use a facial attribute predictor trained on CelebA (Liu et al., 2015) to assign attribute labels to each face.

A.2 Details of the Input Prompts

For existing methods, we follow the prompt templates described in the corresponding papers to guarantee their optimal performance. Specifically, given an attribute <att>, we use "photo of a person, <att>" for ConsistentID (Huang et al., 2024) and PhotoMaker (Li et al., 2024c), "<att>" for InstantID (Wang et al., 2024b), and "portrait, <att>" for PuLID (Guo et al., 2024). For our method, we construct the prompt with format: "who is " + $\sum_{i=1}^{11}(\mathbf{M}(\text{att}_i) + \text{", "})$, where att_i denotes the i^{th} attribute label and $\mathbf{M}$ is a label to text mapping shown in Table 3.

Table 3: Our mapping $\mathbf{M}$ from attribute labels to texts. N/A indicates an empty string.

Attribute (att_i)	Label = 1	Label = 0
1: Bald	"Bald"	N/A
2: Young	"Young"	"Old"
3: Male	"Male"	"Female"
4: Bangs	"with Bangs"	N/A
5: Black Hair	"with Black Hair"	N/A
6: Blond Hair	"with Blond Hair"	N/A
7: Bushy Eyebrows	"with Bushy Eyebrows"	N/A
8: Eyeglasses	"with Eyeglasses"	N/A
9: Mouth Slightly Open	"with Mouth Slightly Open"	N/A
10: No Beard	"with No Beard"	N/A
11: Pale Skin	"with Pale Skin"	N/A

B Additional Visual Results

We present additional qualitative comparisons of single attribute manipulation in Figure 8 and Figure 9, which demonstrate that AttPortrait achieves more effective attribute control than existing methods while maintaining high identity consistency. We also show more results of multi-attribute manipulation in Figure 10, which accurately reflect the given attributes, demonstrating the ability to handle multiple attributes.

C Failure Cases

Three types of failure cases of our model are shown in Figure 11. The first is visual deformation, where generated faces exhibit irregularities or distorted features that reduce visual realism. The second is unexpected artifacts, such as colored marks or undesired accessories. These two types of failure might be related to adversarial effects (Szegedy et al., 2013; He et al., 2019) caused by the attribute matching loss. The third is ineffective manipulation, where the given attributes are not correctly reflected in the generated faces.

D Broader Impact Statement

The ability of our model to generate realistic, attribute-controllable face images also carries risks such as disinformation, privacy violations, and misuse in identity fraud (Westerlund, 2019; Mirsky & Lee, 2021). To mitigate potential harm, we will release the model with mandatory watermarking, a clear usage license prohibiting malicious applications, and guidelines for explicitly labeling generated content. We encourage the community to follow ethical best practices and develop traceability mechanisms to discourage misuse.

Figure 8: Additional qualitative comparisons (1/2). Arc2Face (Paraperas Papantoniou et al., 2024) is included solely to show image quality and identity consistency, as it does not support attribute manipulation.

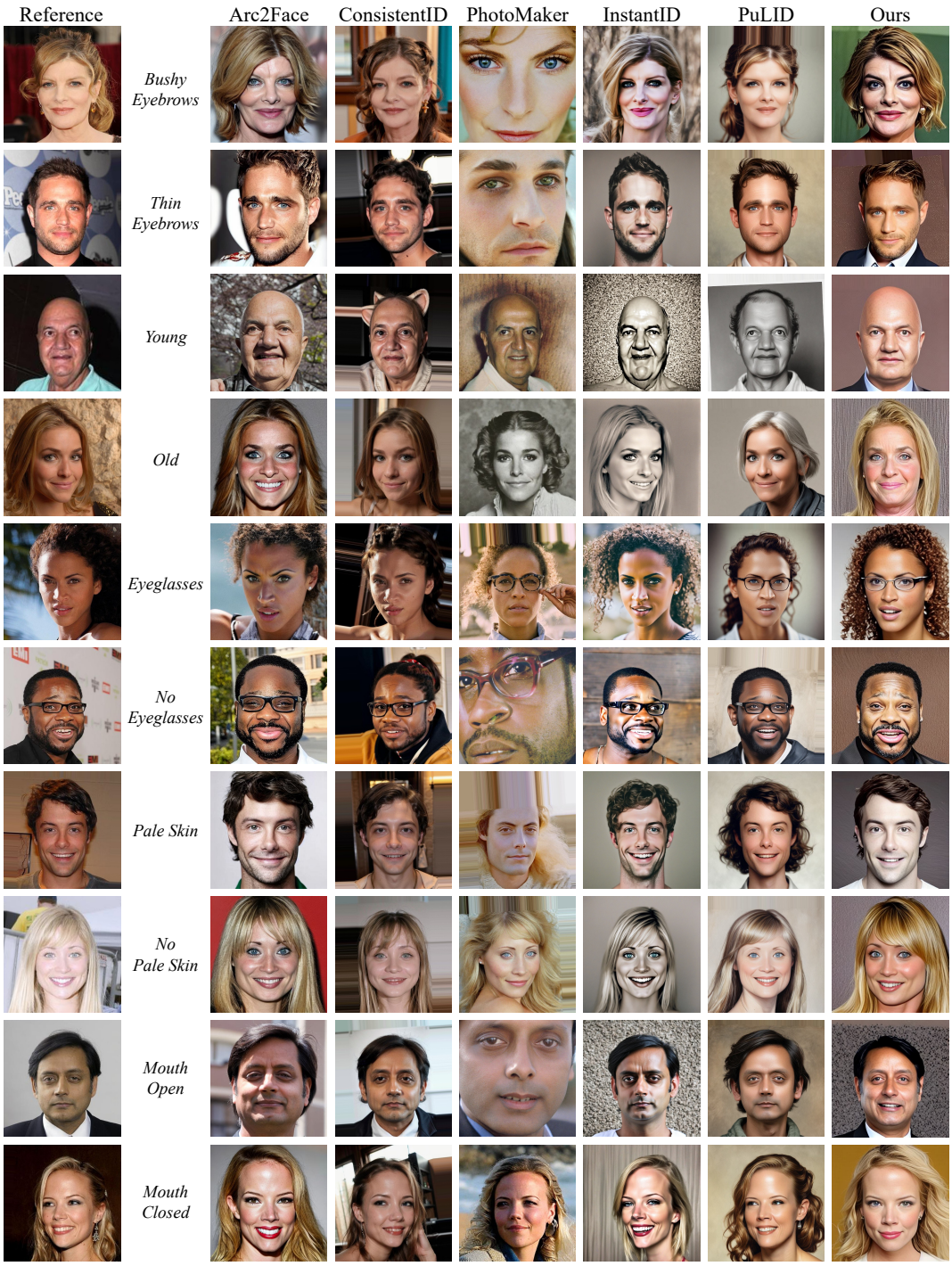

Figure 9: Additional qualitative comparisons (2/2). Arc2Face (Paraperas Papantoniou et al., 2024) is included solely to show image quality and identity consistency, as it does not support attribute manipulation.

Figure 10: Additional results of multi-attribute manipulation.

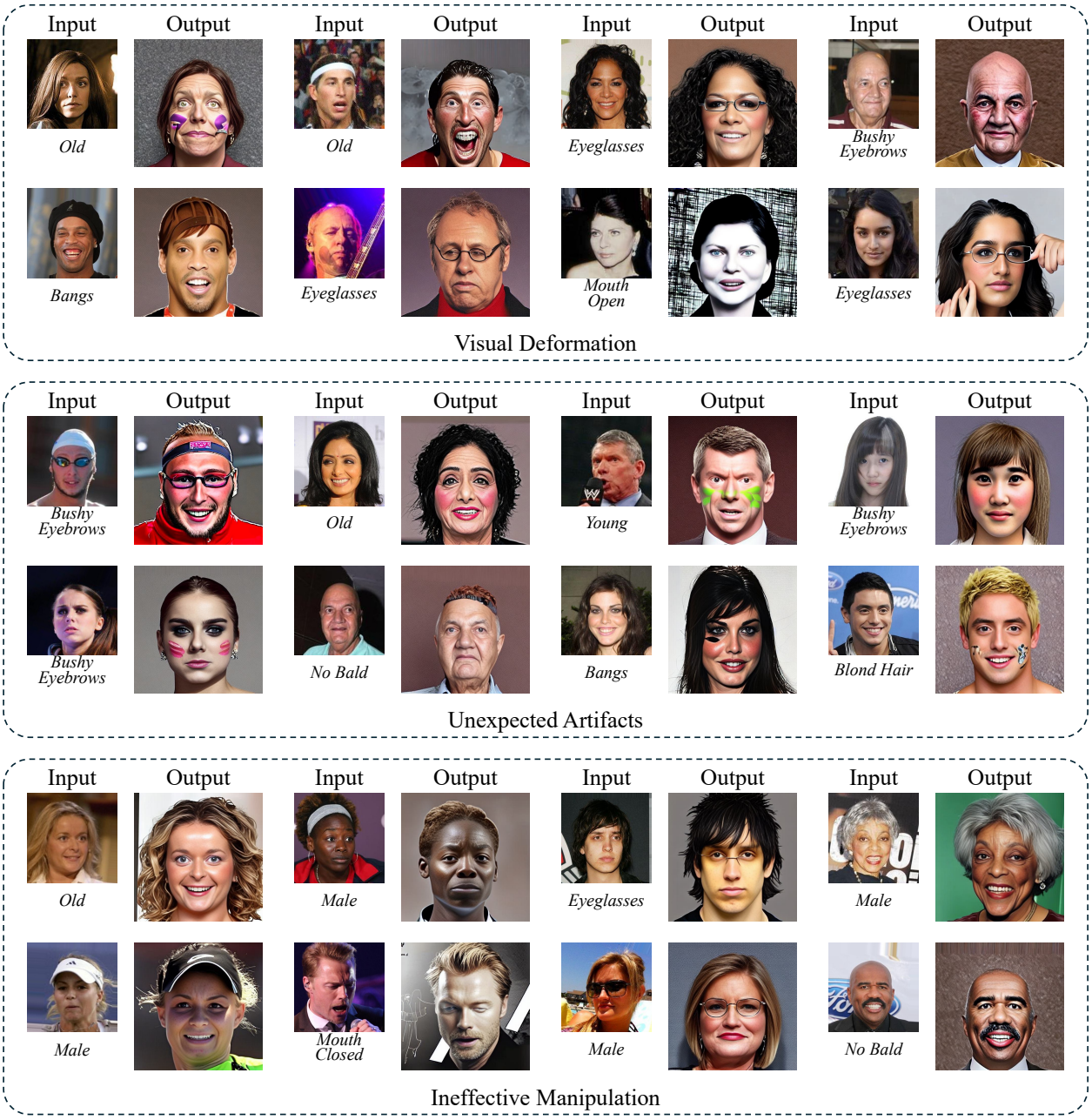

Figure 11: Failure cases.