PIDiff: Image Customization for Personalized Identities with Diffusion Models

Jinyu Gu, Haipeng Liu*, Meng Wang, Yang Wang
School of Computer Science and Information Engineering
Hefei University of Technology, China
2023170666@mail.hfut.edu.cn, hpliu_hfut@hotmail.com,
eric.mengwang@gmail.com, yangwang@hfut.edu.cn

Abstract

Text-to-image generation for personalized identities aims at incorporating the specific identity into images using a text prompt and an identity image. Based on the powerful generative capabilities of denoising diffusion probabilistic models (DDPMs), many previous works adopt additional prompts, such as text embeddings and CLIP image embeddings, to represent the identity information, while they fail to disentangle the identity information and background information. This is because they either mix identity information with text information for backgrounds or extract prompts from content with mixed semantics. As a result, the generated images not only lose key identity characteristics but also suffer from significantly reduced diversity. To address this issue, previous works have combined the $\mathcal{W}_+$ space from StyleGAN with diffusion models, leveraging this space to provide a more accurate and comprehensive representation of identity features through multi-level feature extraction. However, the entanglement of identity and background information in in-the-wild images during training prevents accurate identity localization, resulting in severe semantic interference between identity and background. In this paper, we aim to answer two major questions: 1) how to extract personalized identity features accurately and integrate them into the image generation process effectively. 2) how to leverage training strategies to improve the accuracy of visual prompt localization. To this end, we propose a novel fine-tuning-based diffusion model for personalized identities text-to-image generation, named PIDiff, which leverages the $\mathcal{W}_+$ space and an identity-tailored fine-tuning strategy to avoid semantic entanglement and achieves accurate feature extraction and localization. Style editing can also be achieved by PIDiff through preserving the characteristics of identity features in the $\mathcal{W}_+$ space, which vary from coarse to fine. Through the combination of the proposed cross-attention block and parameter optimization strategy, PIDiff preserves the identity information and maintains the generation capability for in-the-wild images of the pre-trained model during inference. Our experimental results validate the effectiveness of our method in this task. Our code and dataset are available here.

Keywords: Diffusion model, $\mathcal{W}+$ space, Personalized identity customization, Image Synthesis

1. Introduction

In recent years, text-to-image generative models [1, 21, 28, 29, 32, 40, 17] have attracted significant attention due to their ability to synthesize vivid and diverse images from text prompts. The growing demand for customized content has made text-to-image generation for personalized identities a popular research direction. Specifically, the specific identity is expected to be the main subject of generated images. This introduces two key challenges for generative models: first, how to effectively incorporate the personalized identity into the generated images; second, how to ensure text-image semantic consistency while preserving the unique characteristics of the given identity.

Recent methods of text-to-image generation have adopted various approaches to represent specific concepts. Some methods [5, 31, 10] attempt to inverse specific concepts into the text embedding space. They optimize text embeddings and fine-tune the generative model, allowing it to quickly capture the characteristics of the concept. This strategy allows for the optimization of fewer parameters without compromising the model's performance [7, 38, 24, 25]. While these methods perform well in generating simple concepts (such as dogs or doors), they face challenges when generating images of personalized identities. As personalized identities often involve many intricate details that require precise representation. Inverting identity into the text embedding space leads to semantic entanglement with textual information, making it difficult to accurately learn and preserve key identity attributes.

To achieve more accurate identity representation, some

*Haipeng Liu is the corresponding author.

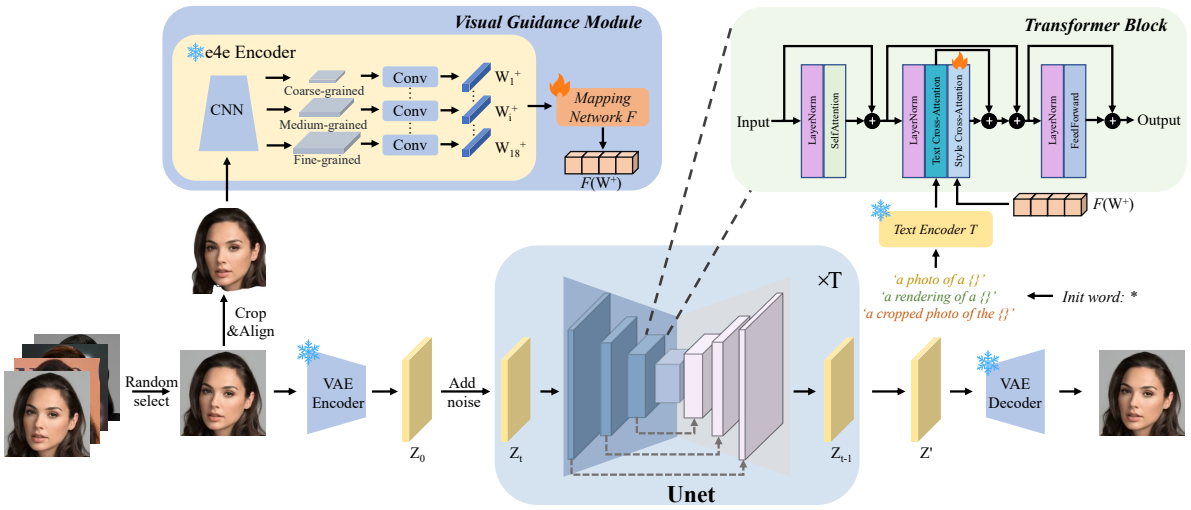

Figure 1. Overview of the proposed PIDiff. PIDiff consists of two modules: Visual Guidance Module(VGM) and the diffusion model. VGM processes the preprocessed image and outputs it to the diffusion model. The diffusion model is based on SDV1.5 and uses new transformer blocks to incorporate both text and visual prompts.

methods [41, 6, 11] modify the generative model by adding additional modules to process visual prompts. Although these methods improve feature representation accuracy through additional visual prompts and processing modules, images generated by these methods exhibit poor diversity, and some key features are often overlooked. This is because their approaches to obtain visual prompts is problematic. For example, IP-Adapter [41] utilizes the image encoder of CLIP. However, extracting visual prompts from image patches is affected by the entanglement of identity and irrelevant region information. As a result, the generated images not only lose crucial personalized identity attributes but also closely resemble the background and some identity attributes of the reference image.

It is worth noting that many works [22, 13, 37, 15, 2, 12, 23, 3] have utilized the $\mathcal{W}_+$ space from StyleGAN for more accurate personalized identity image generation. Therefore, $\mathcal{W}_+$ Adapter [13] combines the $\mathcal{W}_+$ space with diffusion models to generate more accurate personalized identity images. However, in in-the-wild images, a large amount of information from identity-irrelevant regions entangles with the information of identity, making it difficult to accurately localize the visual prompt. As a result, the visual prompts in $\mathcal{W}_+$ Adapter not only fail to accurately localize the face region but also severely interfere with the background.

After analyzing the issues with previous methods, we identify two key problems that must be addressed: 1.*how to accurately extract personalized identity features as visual prompts and integrate them into the image generation process* ; 2. *how to train the model to improve the accuracy of visual prompt localization to ensure the preservation of personalized identity features and avoid semantic entanglement*.

To address the above problems, we propose a novel fine-tuning-based diffusion model called PIDiff for personalized identity text-to-image generation. Due to the excellent performance of diffusion models [1, 21, 28, 29, 32, 40, 19, 16], we adopt the Stable Diffusion as the generative model. We design a Visual Guidance Module(VGM) to process the reference image and provide the visual prompt to the diffusion model. VGM utilize the $\mathcal{W}_+$ space of StyleGAN to represent the specific identity. Notably, PIDiff innovatively applies the $\mathcal{W}_+$ space to customized fine-tuning, enabling personalized identity style editing by preserving the characteristics of the w_+ vector, making the visual prompt more interpretable. During the denoising process, PIDiff utilizes the Style Cross-Attention(SCA) to integrate visual prompts into the image generation process. To improve the accuracy of visual prompt localization and avoid excessive parameter adjustments, we adopt a customized fine-tuning strategy. This training approach effectively addresses the issue of semantic entanglement, which previous methods struggled to avoid. Our pipeline is illustrated in Fig. 1.

Our contributions can be summarized as follows:

1. The utilization of the $\mathcal{W}_+$ space enables more accurate and comprehensive extraction of personalized identity features. SCA cleverly integrates visual prompts provided by VGM into the image generation process. With a customized fine-tuning strategy, PIDiff avoids semantic entanglement and effectively preserves personalized identity features.

2. VGM enables style editing in customized fine-tuning-based methods by preserving the characteristics of the w_+ vector. PIDiff introduces greater diversity to personalized identity text-to-image generation by allow-

ing style combinations.

3. To address the limitation of existing datasets, we propose a small-scale dataset specifically designed for identity image generation. Both qualitative and quantitative analyses demonstrate that PIDiff outperforms state-of-the-art methods in personalized identity text-to-image generation .

2. Methodology

Our work can be summarized into two parts: 1: Customized text-to-image generation for personalized identities:(1) A training strategy for personalized identities (Sec. 2.2.1 and Sec. 2.2.4). (2) Utilizing the w_+ vector to preserve identity features and enable style editing (Sec. 2.2.2). (3) Improving prompt processing capability and training efficiency with a novel Cross-Attention structure (Sec. 2.2.3). (4) The inference phase (Sec. 2.2.5). 2: A comprehensive dataset tailored for personalized identity customization (Sec. 2.3). Before introducing PIDiff, we first elaborate on the preliminaries of diffusion models, which are fundamental to our method.

2.1. Preliminaries

2.1.1 Stable Diffusion

Stable Diffusion (SD) is a variant of diffusion models, referred to as a latent diffusion model (LDM) [29]. It consists of three main components: a Variational Autoencoder (VAE) with an encoder E and a decoder D, a U-Net [30] ϵ_θ, and a text encoder [26] τ. It operates by transforming the input image $I \in \mathbb{R}^{3\times H\times W}$ to the latent code $z_0 \in \mathbb{R}^{4\times H/8\times W/8}$, which is in the higher-dimensional latent space, through the VAE encoder E. DDPM [8] is employed in the training phase for both the forward diffusion process and the reverse denoising process. In the inference phase, DDIM [33] is utilized for the denoising process. Finally, the VAE decoder D will decode the denoised output back into the pixel space. In this way, diffusion models can not only represent images in an efficient way, but also greatly improve computational efficiency.

A crucial component of SD is the attention mechanism, which consists of both self-attention and cross-attention. The self-attention mechanism allows the model to focus on different parts of the image internally, capturing global dependencies [18]. The cross-attention mechanism integrates text conditions into the image generation process, aligning the generated image with the text prompt c. Therefore, inspired by the success of prior methods [6, 10, 13], we primarily focus on optimizing cross-attention. We first obtain the latent code z by adding noise to z_0 through DDPM, then the cross-attention blocks get query features $f(z_t)$ from the hidden state of input image and text embeddings $\tau(c)$ from the text encoder τ. The output of the cross-attention block in the i-th layer can be defined as:

$$\begin{aligned} f_{text}^{i'}(z_t) &= \text{Cross-Attention}(Q^i, K^i, V^i) \\ &= \text{softmax}\left(\frac{Q^i(K^i)^T}{\sqrt{d}}\right)V^i, \end{aligned} \tag{1}$$

where $Q^i = f^i(z_t)W_q^i$, $K^i = \tau(c)W_k^i$, and $V^i = \tau(c)W_v^i$ are the query, key, and value matrices of the i-th cross-attention block, respectively. Specifically, $W_q^i \in \mathbb{R}^{H^{hs}\times W^{hs}}$, $W_k^i \in \mathbb{R}^{H^{hs}\times W^{cd}}$, and $W_v^i \in \mathbb{R}^{H^{hs}\times W^{cd}}$ refer to the projection matrices. Here, cd denotes the cross-attention dimension, which is the dimensionality of the input features used for cross-attention, corresponding to the text embedding size. hs represents the hidden state size. The dimension d of the keys serves to scale the result before applying the softmax function. This mechanism enables the model to align the generated image with the text prompt c by focusing on relevant semantic features.

After introducing the basic principle, components, and the cross-attention mechanism of SD, the training objective of the diffusion model can be written as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c))\|_2^2\right], \tag{2}$$

where z_0 is $E(I)$, z_t is the latent code at timestep t. ϵ is the ground truth noise randomly sampled from a Gaussian distribution. t is uniformly sampled from $\{1, 2, \ldots, T\}$. The ϵ_θ represents the U-Net [30] denoising network.

2.1.2 $\mathcal{W}_+$ latent space

Recently, works based on StyleGAN [9] have achieved great success in the task of human face image generation. The $\mathcal{W}_+$ latent space possesses several unique characteristics that make it particularly powerful for image generation and manipulation. First, $\mathcal{W}_+$ space is multi-dimensional, allowing for fine-grained control over various aspects of the generated image. Additionally, it facilitates style mixing and manipulation by allowing different dimensions of w_+ vectors to be combined. This flexibility makes $\mathcal{W}_+$ latent space an ideal tool for applications such as face image editing, style transfer, and customized image generation.

2.2. Method

To investigate the reasons behind the lower image quality produced by methods (e.g., IP-Adapter [41], $\mathcal{W}_+$ Adapter [13], and Textual Inversion [5]), we visualize the attention maps in Fig. 2. The following analysis provides insights into the causes of the issues with each method:

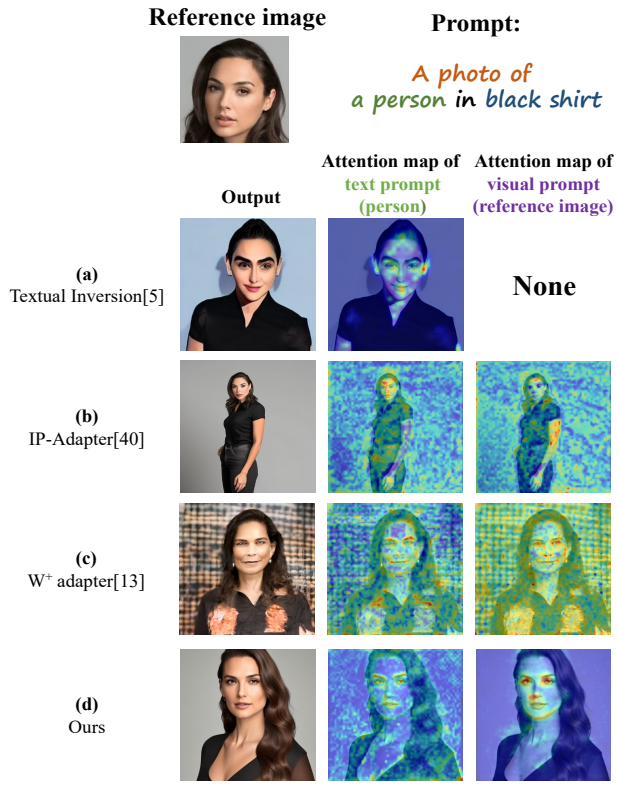

Figure 2. Comparison of Image Generation Results and Attention Maps between Various Methods. The customized training strategy avoids semantic entanglement and effectively preserves key identity features.

1. Although Textual Inversion achieves precise attention localization through its training strategy in Fig. 2(a), the inherent semantic entanglement and limited expressive power of the text embedding space result in the loss of key identity features.

2. The visual and text prompts in IP-Adapter can only roughly localize the person's region in Fig. 2(b). This suggests that CLIP-I causes semantic entanglement by providing embeddings of image patches. Therefore, the visual prompt fails to effectively guide image generation.

3. In Fig. 2(c), the text prompt of $\mathcal{W}_+$ Adapter can localize precisely, whereas the visual prompt's attention is dispersed. This is because the semantic entanglement between identity and background from in-the-wild images prevents the accurate localization of visual prompts.

Therefore, we employ a customized text-to-image generation strategy to avoid semantic entanglement. Then, we utilize w_+ vectors as our visual prompts. To preserve the characteristics of the $\mathcal{W}_+$ space, we process w_+ vectors through our Visual Guidance Module (VGM). Finally, through our Style Cross-Attention (SCA), SD can effectively integrate these prompts into the generated images. Our framework is shown in Fig. 1.

2.2.1 Customized Text-to-Image Generation

Due to the use of pre-trained models, images are influenced by prior knowledge. For example, the word "person" may correspond to the human face images that appear more frequently in the training set. Thus, we use pseudo-words to represent specific identities, such as S^*.

As shown in Fig. 1, during the training process, we only need to provide a few face images of a specific identity, allowing the model to quickly learn accurate localization of visual prompts by avoiding semantic entanglement between identity-relevant and other regions. These images will be encoded by the VAE encoder and perturbed with random noise according to a randomly selected timestep. In the denoising process, we employ random templates as text prompts, for example, "an image of S^* ", "a cropped photo of S^* " and so on. Finally, the UNet outputs the predicted noise based on the text prompt, visual prompt, and the timestep.

During the inference stage, we only need to provide an image of a specific identity and the text prompt that describes the final image. Our model will generate in-the-wild images that not only contain the details of specific identities but also maintain semantic consistency with the text prompt.

2.2.2 Visual Guidance Module

Due to the semantic entanglement in in-the-wild images, simply providing image patch embeddings [39, 41] results in the loss of key features. Therefore, we choose the more accurate and flexible w_+ vector as our visual prompt. To ensure the w_+ vector can be directly utilized by SD, we design the visual guidance module. First, the preprocessing module can align face images and remove backgrounds. The processed images I_{crop} will be input into the e4e encoder [34]. As shown in Fig. 1, the e4e encoder first extracts features of the input image from coarse to fine through a CNN, and then obtains the $w_+ \in \mathbb{R}^{18\times512}$ vector by mapping modules. However, the w_+ vector is designed for StyleGAN's generator, so we use a mapping network F to project the w_+ vector. Through the mapping network, the w_+ vector can be mapped to a visual embedding $F(w_+) \in \mathbb{R}^{4\times768}$ to guide the denoising process of the U-Net.

The visual guidance module also preserves the properties of the $\mathcal{W}_+$ space. As shown in Fig. 1, a CNN in the visual prompt module encodes the image into hidden codes of different levels, enabling the final output w_+ vector to have multi-level semantic expression capabilities. During

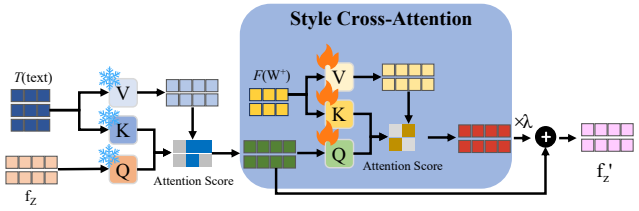

Figure 3. Illustration of Style Cross-Attention(SCA). SCA takes the output of the text cross-attention block as the query and uses the visual prompts as the keys and values. Only the projection matrices is trainable in SCAs.

the inference phase, we can concatenate w_+ vectors of images with different styles to achieve style editing to some extent. For example, in Fig. 11, we can generate a person with desired styles by concatenating hidden codes of different images, which previous methods [6, 31, 39] for customizing specific concepts could not achieve.

2.2.3 Style Cross-Attention

Previous efforts attempted to achieve personalized identity customization through optimizing text embeddings or fine-tuning diffusion models. By analyzing weight changes after training, Custom Diffusion [10] discovered that, although the cross-attention blocks have relatively few parameters, they have a significant impact on the model's performance. Motivated by these findings, we introduce Style Cross-Attention (SCA) to integrate visual prompts into the denoising process.

As shown in Fig. 3, SCA is a cross-attention block behind the text cross-attention block. The text cross-attention block is the original cross-attention block for text in the diffusion model. We add SCA for processing visual prompts after the text cross-attention block to integrate visual prompts using semantically richer queries. This structure helps the model localize visual prompts more accurately and effectively prevents the disruption of the text-image semantic consistency of the pre-trained model. Specifically, SCA takes the output of the text cross-attention block as the query and visual embeddings from Mapping Network F as key and value. The output of SCA can be defined as:

$$\begin{aligned} f_{SCA}^{i''}(z_t) &= \text{Cross-Attention}\left(Q^{i'}, K^{i'}, V^{i'}\right) \\ &= \text{softmax}\left(\frac{Q^{i'}(K^{i'})^T}{\sqrt{d}}\right)V^{i'}, \end{aligned} \tag{3}$$

where $Q^{i'} = f_{text}^{i'}(z_t)W_q^{i'}$, $K^{i'} = F(w_+)W_k^{i'}$, and $V^{i'} = F(w_+)W_v^{i'}$, where $W_q^{i'}$, $W_k^{i'}$, and $W_v^{i'}$ are the projection matrices for query, key, and value. $f_{text}^{i'}(z_t)$ is defined in Eq. (1).

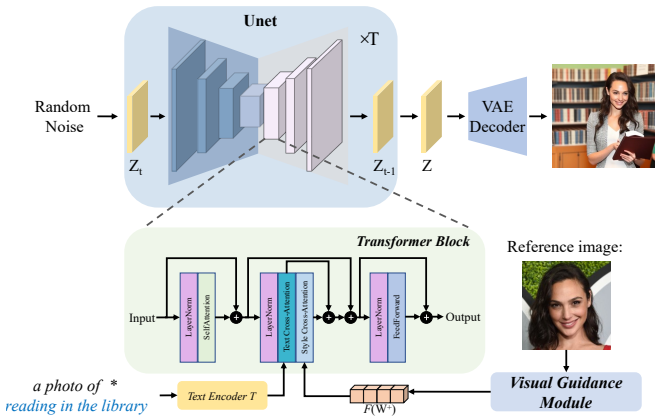

Figure 4. Illustration of inference process in diffusion models: Given a text prompt (e.g.," a photo of * reading in the library") and an identity image, PIDiff can generate images for the identity.

Finally, the output of SCA combines the output of text cross-attention block. The final output can be defined as:

$$f(z_t) = f_{text}^{i'}(z_t) + \lambda \cdot f_{SCA}^{i''}(z_t), \tag{4}$$

where λ is a parameter that controls the contribution of SCA, which processes visual prompts. During training, λ is set to 1. In the inference stage, λ can be adjusted to balance the semantic information from the text with the visual style derived from the visual prompt.

2.2.4 Training Loss

To accelerate model convergence, $W_q^{i'}$, $W_k^{i'}$, $W_v^{i'}$ are initialized from W_q^i, W_k^i, W_v^i respectively. During the training process, only $W_q^{i'}$, $W_k^{i'}$, $W_v^{i'}$ in SCAs and the mapping network F will be trainable. This strategy helps PIDiff better preserve the generative capability of the pre-trained model.

The final optimization objective for the model is given as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c), F(w_+))\|_2^2\right], \tag{5}$$

where Eq. (5) is similar to Eq. (2), except that it incorporates an additional visual condition $F(w_+)$ and adds SCAs after the original text cross-attention blocks in the U-Net.

2.2.5 Inference Stage

During the inference stage, we utilize Stable Diffusion (SD) as the generative model. As shown in Fig. 4, this process can be broken down into several key steps. Specifically, we first sample Gaussian noise. This noise serves as the initial latent code. Next, we employ the DDIM denoising process to iteratively refine the latent code. Given a target text prompt, such as "a photo of * reading in the library ",

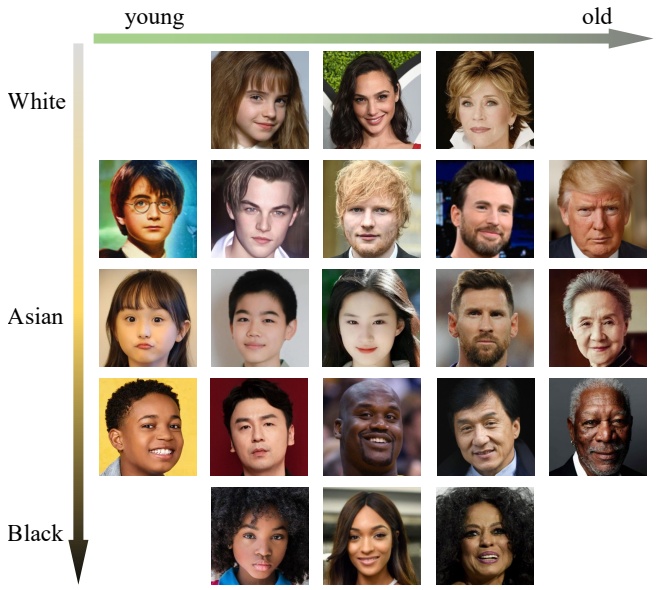

Figure 5. Visualization of our proposed personalized identity customization dataset, we select samples from our dataset showcasing diverse identities, including variations in race, age, and gender.

we obtain the corresponding text embeddings using the text encoder. We also need to provide a specific identity image to guide the generation process through VGM. Finally, the generated image is obtained by decoding the final latent code using the VAE decoder.

2.3. Dataset Construction Method

Although existing datasets such as FFHQ contain high-quality face images, they do not offer multiple images for each identity. However, datasets that contain multiple images for each identity are primarily designed for tasks such as image recognition. As a result, these images are often affected by variations in angles and lighting, leading to suboptimal quality. We also found that while the latest CrossFaceID [36] dataset demonstrates good quality and diversity, it contains a significant number of duplicate samples.

To overcome this limitation, we constructed a novel dataset tailored for personalized identity customization. We provide the dataset here. This dataset was meticulously curated by first selecting high-quality images from the CrossFaceID dataset comprising 40,000 identity samples, followed by supplementing identity categories with images retrieved from Google to ensure comprehensive coverage and diversity.

This carefully curated dataset aims to mitigate potential biases arising from imbalanced representation of demographic groups. The dataset comprises approximately 500 identity images across three racial groups: White, Asian and Black. Each racial group is further divided into three age brackets: 0–20, 21–50, and 51+, where the age of the samples refers to the age at the time the photo was taken, rather than their current actual age (dataset samples are shown in Fig. 5). Moreover, each identity is represented by at least six images with diverse poses, viewing angles, expressions, and environmental variations to ensure the model's robustness in challenging scenarios. By incorporating individuals across various racial, age, and gender groups, our dataset promotes more equitable and accurate research outcomes, fostering a deeper understanding of the complexities in facial recognition technology. The combined dataset effectively addresses the scarcity of identity variations and enhances model generalization. All images in our provided dataset are publicly available and free from privacy issues, ensuring ethical compliance and legal usability for research purposes.

3. Experiment

3.1. Implementation Details

We implement the proposed PIDiff in pytorch framework under the running environment as: python 3.12.4, pytorch 2.2.2 and cuda 12.0. The codes are available here. We use the pre-trained SD V1.5 as the generative model. We train our model on an A40 GPU. We employ the AdamW optimizer [20] with a learning rate of 1×10^{-4} and weight decay of 0.01. We train PIDiff with a batch size of 4 for 600 steps. We employ dropout probabilities of 0.5 for visual embeddings and 0.3 for text embeddings. We also add random noise to w_+ vectors. We adopt DDIM [33] with 50 steps during inference.

3.2. Evaluation Metrics

We use ID, LPIPS, and CLIP-T to evaluate the performance. **Identity Loss (ID↑)**: We first use MTCNN [42] for face alignment. Then we use ArcFace [4] to measure detected faces. Finally, we calculateg the cosine similarity between the feature of the generated image and the original image. **Learned Perceptual Image Patch Similarity (LPIPS↓)** [43]: We use VGG-V0.1 for image feature extraction and evaluate image similarity by comparing the differences between features. **Text-image similarity (CLIP-T↑)** [27]: We use the pre-trained clip-vit-base-patch16 to calculate the similarity between the generated image and the text prompt.

3.3. Comparison with State-of-the-arts

To validate the superiority of PIDiff, we compare it with typical models, including: Textual Inversion [5] proposes finding text embeddings for different concepts. Custom Diffusion [10] introduces fine-tuning the cross-attention mapping matrix. VICO [6] and IP-Adapter [41] attempt to use additional modules to integrate visual prompts. $\mathcal{W}_+$ adapter [13] utilizes the $\mathcal{W}_+$ space to achieve high-quality identity representation. PhotoMaker [14] fuses multiple identity im-

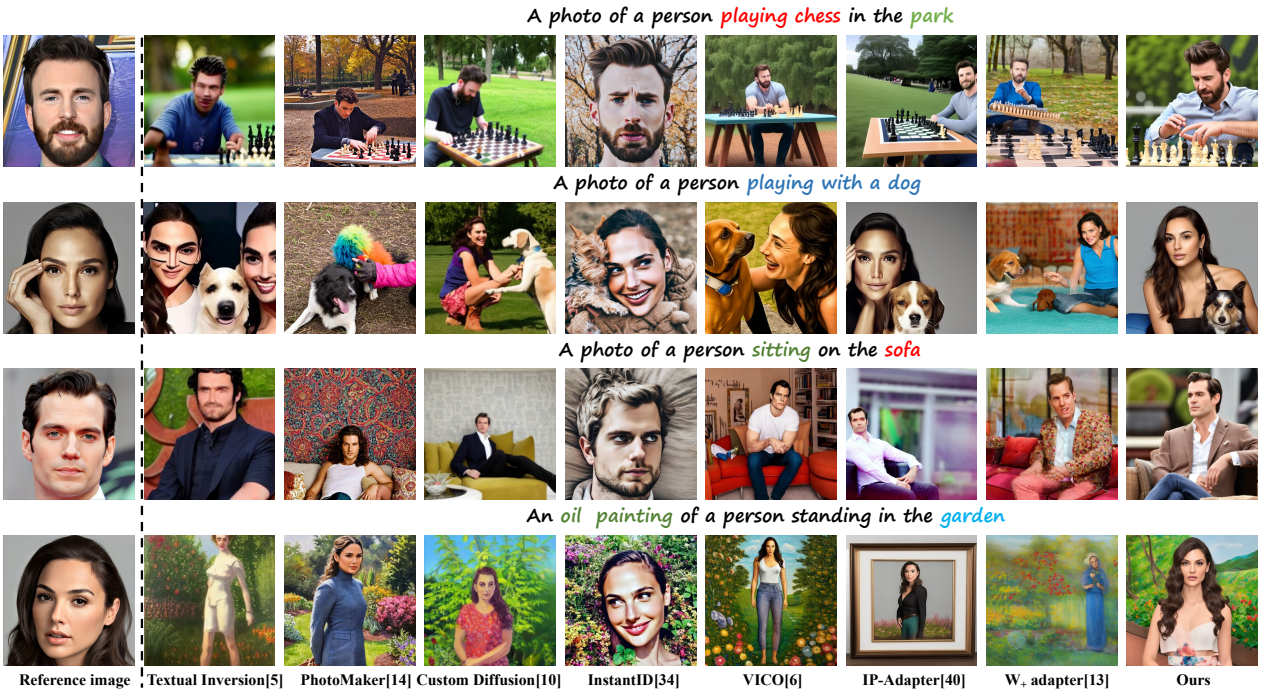

Figure 6. Qualitative Comparison between previous methods and PIDiff. PIDiff not only maintains identity features and text-image semantic consistency but also generates images with significantly high quality.

Table 1. Quantitative comparisons with previous methods. The best result is shown in **bold**, and the second best is underline.

Methods	ID↑	LPIPS↓	CLIP-T↑
Textual Inversion[5]	0.1177	0.6763	0.1111
Custom Diffusion[10]	0.1256	0.7261	0.1484
PhotoMaker[14]	0.2479	0.6712	**0.2124**
InstantID[35]	0.3052	0.6948	0.1105
VICO[6]	0.2490	0.6999	0.1412
IP-Adapter[41]	0.2911	0.6094	0.1593
$\mathcal{W}_+$ adapter[13]	0.2407	0.6740	0.1927
Ours	**0.3109**	**0.5945**	0.1951

Table 2. Quantitative comparisons with previous methods. The best result is shown in **bold**, and the second best is underline.

Methods	IF↑	PF↑
Textual Inversion[5]	0.4516	0.3524
Custom Diffusion[10]	0.4988	0.5647
PhotoMaker[14]	0.5662	**0.8226**
InstantID[35]	0.7144	0.4296
VICO[6]	0.5245	0.4769
IP-Adapter[41]	0.6881	0.6468
$\mathcal{W}_+$ adapter[13]	0.5716	0.7125
Ours	**0.7261**	0.7883

ages into a stacked CLIP embedding and blends them with text prompts via LoRA-adapted cross-attention. InstantID [35] achieves zero-shot ID-preserving generation from a single reference using a face encoder, a cross-attention adapter, and IdentityNet. To ensure a fair comparison, all the experiments are conducted using the proposed dataset.

3.3.1 Quantitative Comparison

In the experiment, we use 12 text prompts as text conditions for each identity. These text prompts include scenarios with single people, multiple people, and multiple objects, as well as requirements for clothing and poses. Since some models, such as Textual Inversion, do not require reference images during inference, we select the most similar facial image to compute the evaluation metrics. This comprehensive evaluation ensures an accurate performance assessment for each model.

As shown in Table 1, our model achieves outstanding results. Notably, our method attains higher ID scores. This stems from the $\mathcal{W}_+$ space, which provides superior expressive power. Additionally, our approach outperforms $\mathcal{W}_+$ adapter. This is because it is affected by the training strategy, which causes visual prompts to fail in accurately localizing relevant regions (e.g., Fig. 2(c)).

Our method also achieves high CLIP-T metric, benefiting from the fusion of the training strategy and SCA. As stated in Sec. 2.2.1 and Sec. 2.2.3, *the customized training*

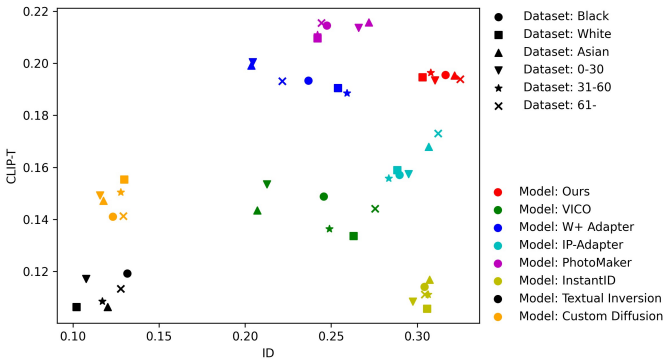

Figure 7. Comparative Analysis of Models Across Various Datasets. Compared to other methods, our approach is free from bias. Outstanding experimental results demonstrate that our method is both superior and more stable.

Table 3. Analysis of training steps for train. The best result is shown in **bold**.

number	400	600	800	1000
ID↑	0.2973	0.3109	0.3219	**0.3390**
CLIP-T↑	0.1535	**0.1951**	0.1265	0.1144

strategy allows the model to quickly learn accurate localization of visual prompts by avoiding semantic entanglement. SCA effectively prevents the disruption of the text-image semantic consistency of the pre-trained model.

Furthermore, we conducted a user study comparing our method with prior works. For identity fidelity (IF), we recruited 100 users to complete a questionnaire comprising ten questions evaluating gender, age, race, and key identity attributes. For prompt fidelity (PF), each generated image was assessed through seven questions, such as "Does the key scene match the prompt?" and "Do the action align with the prompt?". After collecting user responses, we statistically analyzed the results, as shown in Table 2. Although PhotoMaker's mapping network uses MLPs to output a stacked identity embedding for improved prompt fidelity, it sacrifices identity fidelity in the process. Our method, PIDiff, demonstrates superior performance in both metrics. Compared to previous customized fine-tuning approaches like VICO and Textual Inversion, PIDiff significantly enhances identity fidelity through SCA. In contrast to methods such as $\mathcal{W}_+$ adapter, which support customization for arbitrary identities, PIDiff's fine-tuning strategy combined with the $\mathcal{W}_+$ space more effectively mitigates semantic entanglement, thereby improving both prompt fidelity and image quality.

3.3.2 Qualitative Comparison

Fig. 6 demonstrates the visual comparison between PIDiff and other methods. It can be observed that methods without

Figure 8. Qualitative results of using different training steps. Too few steps lead to identity loss. Too many steps cause overfitting, reducing text-image semantic consistency.

Table 4. Analysis of Data Augmentation. The best result is shown in **bold**.

Methods	ID↑	LPIPS↓	CLIP-T↑
w/o aug	0.2744	0.6412	0.1649
w/ aug	**0.3109**	**0.5945**	**0.1951**

visual prompts have poor identity preservation capabilities in the generated images. We can also notice that faces in the images generated by IP-Adapter are relatively fixed, and the background is affected by visual prompts and IP-Adapter rows 2 and 4 in Fig. 6). This aligns with the issue regarding CLIP image encoder that we highlighted in Sec. 2.2.2: *Simply providing image patch embeddings results in the loss of key features due to semantic entanglement, and this also causes a decrease in image diversity.*

These results further validate the effectiveness of the $\mathcal{W}_+$ latent space and the Visual Guidance Module in our method. By aligning faces and cropping backgrounds, the visual prompts provided by the preprocessing module effectively mitigate background interference. Through VGM's multi-level processing approach, the visual prompts can be more effectively utilized by SCA.

Fig. 7 presents our results across various test categories, including different ethnicities and age groups. The stable results suggest that our model is free from bias and is suitable for text-to-image generation tasks for a wide range of identities.

3.4. Ablation Study

3.4.1 Analysis of Training Steps

Different training strategies often require varying numbers of training steps. Previous methods require significantly different numbers of training steps. Therefore, we analyze how many steps are suitable for our model.

As shown in Table 3, we observe that when the number of training steps is fewer than 600, the model fails to integrate the specific identity into the pseudo-words, resulting in generated images that do not effectively capture the specific identity. Conversely, excessive training steps cause the model to overfit to the specific identity, leading to a decline in semantic alignment with text prompts (visual comparison is shown in Fig. 8).

Figure 9. Qualitative results of using data augmentation. PIDiff can more accurately preserve identity features through data augmentation.

Table 5. Quantitative Comparison of Training Configurations. The best result is shown in **bold**.

images	4	6	8	10
ID↑	0.2457	**0.3109**	0.2888	0.2883
CLIP-T↑	0.1410	**0.1951**	0.1663	0.1477

To balance the ID and CLIP-T metrics, we select 600 training steps. This choice is based on the observation that our model achieves the best performance on CLIP-T while also generating high-quality images at 600 steps.

3.4.2 Analysis of Data Augmentation

During training, we randomly add noise to w_+ vector. We drop text prompts with a probability of 0.3 and visual prompts with a probability of 0.5. First, the random noise added to w_+ vector effectively prevents overfitting of vector-image pairs and improve the diversity of generated images. We also discard the prompts with a high probability to avoid overfitting visual and text prompts and help the model customize to a specific identity. Fig. 9 shows the effect of our data augmentation. We also show the experiment results in table 4.

3.4.3 Analysis of Number of Images for Train

Previous methods typically require a few images to customize specific concepts. However, identity face has many unique characteristics that need to be preserved. Therefore, the number of images used for training must be reasonable. We choose to use six images for training. In Fig. 10, we show the visual comparison between our choice and other numbers, and we also show experiment results in Table 5. It can be seen that when there are fewer images for training, the model tends to overfit images and text prompts in dataset. In the inference stage, the generated image is easily affected by the text prompt, which leads to the degradation of image quality. When too many images are provided, the model struggles to learn the key identity characteristics due to the diversity of poses, expressions, and appearances across the images, leading to degraded identity fidelity.

Figure 10. Qualitative results of using different numbers of training images

Table 6. Analysis of Number of Images for Training (SI:Specific identity, ADD:An additional image of a new identity with a specific style). The best result is shown in **bold**.

number	ID↑	LPIPS↓	CLIP-T↑
6(SI)+1(ADD)	0.3097	0.6146	0.1818
6(SI)	**0.3109**	**0.5945**	**0.1951**

3.4.4 Analysis of the Number of Images for Training in Style Editing

One of the contributions of PIDiff is the implementation of style editing under a customized fine-tuning training strategy. PIDiff enables style editing by combining w_+ vectors from different images of a specific identity. However, if the desired style is absent in the images of the specific identity, an additional image is needed to provide the required style during training. Therefore, to ensure the preservation of the original identity features while learning the new style, we train the model using six images of the specific identity along with an additional face image containing the desired style. This method may lead to some degree of performance degradation,but as shown in Table 6, our model still maintains strong performance.

3.4.5 Analysis of Style Cross-Attention Structure

In IP-Adapter, a parallel cross-attention block(PCA) design is adopted, where the query for the visual cross-attention block comes directly from the hidden state. However, in SCA, the query originates from the output of the text cross-attention block. Therefore, we show experimental comparisons in table 7. We found that, due to the processing by the text cross-attention block, different semantic regions in the image are better distinguished, allowing the visual prompt

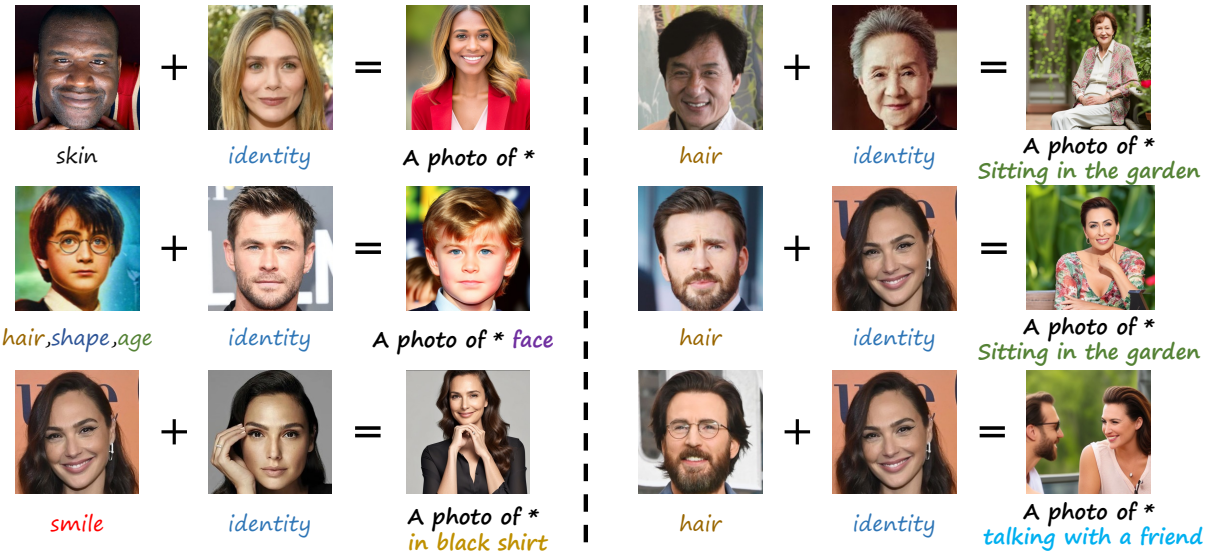

Figure 11. Given face images with coarse-grained features (e.g., hair, skin color, face shape) and fine-grained features (facial details), PIDiff can concatenate the w_+ vectors and generate personalized identity images with new styles

Table 7. Quantitative Comparison of Style Cross-Attention Structure. The best result is shown in **bold**.

Structure	PCA	SCA
ID ↑	0.2235	**0.3109**
CLIP-T↑	0.1468	**0.1951**

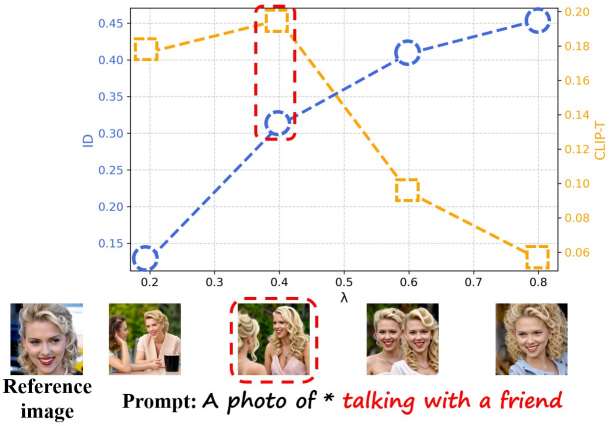

Figure 12. Visual comparisons of images generated by using different λ in SCA. When λ=0.4, PIDiff achieves the highest text-image semantic consistency while effectively preserving identity features.

to be more precisely localized. As a result, our SCA can help the visual prompt focus on the facial region more accurately, ensuring the retention of identity features.

3.4.6 Analysis of Computational Complexity

We report the number of trainable parameters (i.e., those fine-tuned during training) and inference times for each method to analyze computational complexity. We conducted comprehensive testing of all models' average inference time under consistent hardware conditions (using a single NVIDIA A40 GPU). The results in Table 8 demonstrate significant variations in both parameter counts and inference speeds across methods. Notably, PIDiff not only generates high-quality images but also achieves this with fewer parameters while maintaining faster inference speeds compared to alternative methods.

Table 8. Computational comparison of different models. The best result is shown in **bold**, and the second best is underline.

Model	Params (M)	Inf Time (s)
PIDiff	38.5	8
VICO	51.3	65
W+ Adapter	38.5	**6**
Ip-Adapter	**22.5**	**6**
DreamBooth	982.6	11
Custom Diffusion	57.1	8

3.4.7 Analysis of Style Editing

Through the use of w_+ vector, our model is capable of performing a certain degree of style editing. We can synthesize images of identities with specific styles by combining the w_+ vectors of the specific identity. As demonstrated in Fig. 11, the first image provides the desired style, and the second provides the fine-grained characteristics that represent the specific identity. By fusing w_+ vectors, we can add a new style to the specific identity.

3.4.8 Analysis of λ in Style Cross-Attention

We use λ to control the influence of w_+ vectors on the hidden states in SCA. As shown in the Fig. 12, when λ approaches 0, the generated images retain the text alignment capabilities of the pre-trained SD, but the specific identity is not wCell preserved. When λ approaches 1, the generated image fails to match the text prompt. It can be seen that when λ is smaller than 0.4, the features of the specific identity are lost. When the λ is larger than 0.4, CLIP-T rapidly decreases, while ID does not change significantly. Therefore, through experimental analysis, we choose 0.4 as an appropriate choice.

4. Conclusion

In this paper, we propose PIDiff for personalized identities text-to-image generation and demonstrate that: 1) The $\mathcal{W}_+$ space enhances the diffusion model's accuracy in representing identity features and enables flexible style editing. 2) The attention mechanism and customized fine-tuning strategy innovatively integrate the $\mathcal{W}_+$ space with diffusion model fine-tuning, effectively addressing the semantic entanglement issues prevalent in prior diffusion-based approaches. This integration significantly enhances both identity fidelity and textual fidelity. Although PIDiff is designed for single-identity image generation, our future work will explore enhanced injection methods for w_+ vectors to achieve multi-identity image synthesis. Extensive experimental results validate that PIDiff is free from bias and outperforms previous methods.

Acknowledgments

This research is supported by National Natural Science Foundation of China (U21A20470, 62172136, 72188101); Institute of Advanced Medicine and Frontier Technology (2023IHM01080), and sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202513); The computation is completed on the HPC Platform of Hefei University of Technology.

References

[1] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2

[2] A. C. Baykal, A. B. Anees, D. Ceylan, E. Erdem, A. Erdem, and D. Yuret. Clip-guided stylegan inversion for text-driven real image editing. *ACM Transactions on Graphics*, 42(5):1–18, 2023. 2

[3] D. Bobkov, V. Titov, A. Alanov, and D. Vetrov. The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9337–9346, 2024. 2

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[5] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 3, 6, 7

[6] S. Hao, K. Han, S. Zhao, and K.-Y. K. Wong. Vico: Plug-and-play visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2, 3, 5, 6, 7

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[9] T. Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 3

[10] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 3, 5, 6, 7

[11] D. Li, J. Li, and S. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[12] H. Li, M. Huang, L. Zhang, B. Hu, Y. Liu, and Z. Mao. Gradual residuals alignment: a dual-stream framework for gan inversion and image attribute editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3064–3072, 2024. 2

[13] X. Li, X. Hou, and C. C. 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*, pages 2187–2196, 2024. 2, 3, 6, 7

[14] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 6, 7

[15] H. Liu, Y. Song, and Q. Chen. Delving stylegan inversion for image editing: A foundation latent space viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10082, 2023. 2

[16] H. Liu, Y. Wang, B. Qian, M. Wang, and Y. Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024. 2

[17] H. Liu, Y. Wang, and M. Wang. One stone with two birds: A null-text-null frequency-aware diffusion models for text-guided image inpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[18] H. Liu, Y. Wang, M. Wang, and Y. Rui. Delving globally into texture and structure for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1270–1278, 2022. 3

[19] J. Long, G. Ye, T. Chen, Y. Wang, M. Wang, and H. Yin. Diffusion-based cloud-edge-device collaborative learning for next poi recommendations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2026–2036, 2024. 2

[20] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[22] H. Pehlivan, Y. Dalva, and A. Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1828–1837, 2023. 2

[23] H. Pehlivan, Y. Dalva, and A. Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1828–1837, 2023. 2

[24] B. Qian, Y. Wang, R. Hong, and M. Wang. Adaptive data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7960–7968, 2023. 1

[25] B. Qian, Y. Wang, R. Hong, and M. Wang. Rethinking data-free quantization as a zero-sum game. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9489–9497, 2023. 1

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1, 2

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[30] O. Ronneberger, P. Fischer, and T. 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*, pages 234–241. Springer, 2015. 3

[31] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. 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*, pages 22500–22510, 2023. 1, 5

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[33] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 6

[34] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 4

[35] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, and Y. Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 7

[36] S. Wang, X. Li, X. Sun, G. Wang, T. Zhang, J. Li, and E. Hovy. Turn that frown upside down: Faceid customization via cross-training data. *arXiv preprint arXiv:2501.15407*, 2025. 6

[37] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11379–11388, 2022. 2

[38] Y. Wang, B. Qian, H. Liu, Y. Rui, and M. Wang. Unpacking the gap box against data-free knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[39] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 4, 5

[40] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[41] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 4, 6, 7

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 6

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6