MasterWeaver: Taming Editability and Face Identity for Personalized Text-to-Image Generation

Yuxiang Wei[1,2] Zhilong Ji[3] Jinfeng Bai[3] Hongzhi Zhang[1]
Lei Zhang[2(✉)] Wangmeng Zuo[1,4(✉)]

[1]Harbin Institute of Technology [2] The Hong Kong Polytechnic University
[3]Tomorrow Advancing Life [4] Pazhou Lab Huangpu

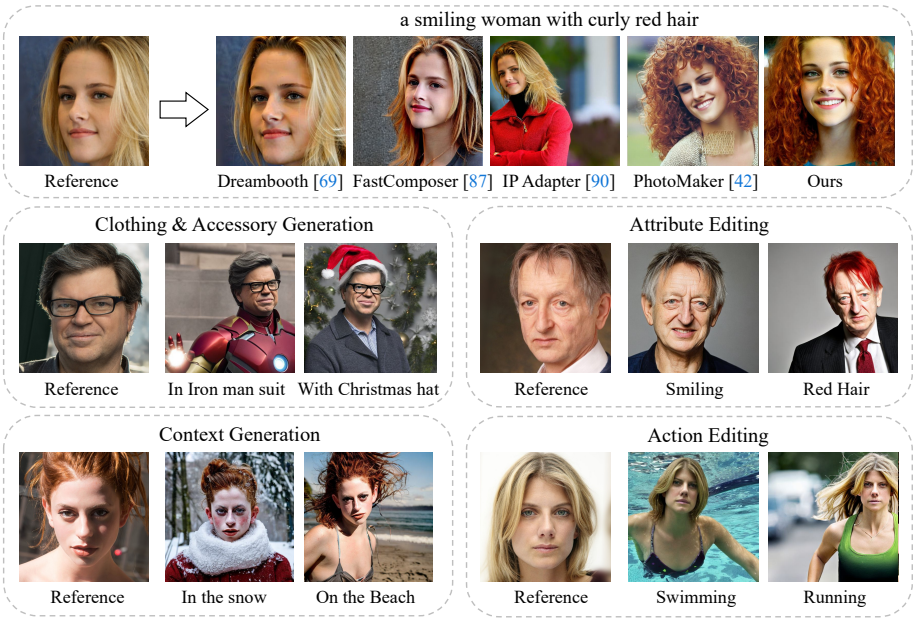

Fig. 1: With one single reference image, our MasterWeaver can generate photo-realistic personalized images with diverse clothing, accessories, facial attributes and actions in various contexts. In comparison with existing methods, our method exhibits superior editability while maintaining high identity fidelity.

Abstract. Text-to-image (T2I) diffusion models have shown significant success in personalized text-to-image generation, which aims to generate novel images with human identities indicated by the reference images. Despite promising identity fidelity has been achieved by several tuning-free methods, they often suffer from overfitting issues. The

learned identity tends to entangle with irrelevant information, resulting in unsatisfied text controllability, especially on faces. In this work, we present **MasterWeaver**, a test-time tuning-free method designed to generate personalized images with both high identity fidelity and flexible editability. Specifically, MasterWeaver adopts an encoder to extract identity features and steers the image generation through additionally introduced cross attention. To improve editability while maintaining identity fidelity, we propose an editing direction loss for training, which aligns the editing directions of our MasterWeaver with those of the original T2I model. Additionally, a face-augmented dataset is constructed to facilitate disentangled identity learning, and further improve the editability. Extensive experiments demonstrate that our MasterWeaver can not only generate personalized images with faithful identity, but also exhibit superiority in text controllability. Our code can be found at https://github.com/csyxwei/MasterWeaver.

Keywords: Personalized Text-to-Image Generation · Identity Preservation · Editability

1 Introduction

Recently, text-to-image diffusion models [68] have demonstrated superior capabilities in generating high-quality and creative images. Building upon these advancements, personalization methods [20,69,85,90] further enable the generation of a specific visual subject (*e.g.*, objects, animals, or people) as indicated by one or several reference images. In this work, we focus on the personalized image generation of human identities, which has wide applications, such as personalized portrait photos [42], art creation [78], and visual try-on [80].

Earlier studies, *e.g.*, Dreambooth [69] and Textual Inversion [20], usually learned the novel identities from reference images by optimizing the word embeddings or model parameters. Albeit the promising results, these methods require per-identity optimization, which is time-consuming and impractical for real applications. Additionally, they often take multiple images for identity learning and suffer from poor editability in limited-data scenarios (see the top of Fig. 1). Recent tuning-free methods, *e.g.*, FastComposer [87] and IP Adapter [90] trained additional visual encoders to extract the identity information, and inject them through model tuning or adapters. After training on human datasets, these methods could produce personalized results in several denoising steps with only one identity image as the reference. Despite these methods having achieved promising identity fidelity, they usually suffer from overfitting issues. Since the reference identity image and the input image used during training are from the same image, these methods tend to directly copy the reference image during generation, resulting in poor editability (see the top of Fig. 1). Photomaker [42] took multiple images of the same identity to learn a stacked face embedding, which can be used to generate images with diverse attributes. However, its identity fidelity is sacrificed for improving editability.

To address the above issues, we propose **MasterWeaver**, a tuning-free method for generating personalized images with both high identity fidelity and flexible text controllability. Following [85, 90], MasterWeaver adopts an encoder to extract the identity features and steers the image generation through additionally introduced cross attention. To improve the editability while keeping the identity fidelity, we propose identity-preserved editability learning. We first propose an editing direction loss to facilitate the model training. Specifically, we identify the editing direction in the feature space of diffusion model by inputting paired text prompts that denote an editing operation, *e.g.*, (a photo of a person, a photo of a person with <attribute>). As the direction's computation solely depends on the attribute difference, such a direction captures the meaningful semantic editing prior and is unrelated to identity. By aligning the editing direction of our MasterWeaver with that of the original T2I model, we can significantly enhance the text controllability without compromising identity fidelity.

Additionally, we construct a face-augmented dataset to facilitate disentangled identity learning. In particular, we utilize the face editing method [53] to modify the attributes of reference identity images, which are then incorporated into our augmented dataset. In this dataset, the reference identity image and its corresponding input image have the same identity but differ in a single attribute (*e.g.*, hair color, style, or expression). Such a controlled attribute misalignment can effectively facilitate model learning to extract faithful identity features disentangled with attribute details, thereby improving editability. As illustrated in Fig. 1, with only one reference image, our MasterWeaver can generate photo-realistic personalized images with faithful identity and diverse contexts.

Extensive experimental evaluations demonstrate that our MasterWeaver outperforms current state-of-the-art methods. The main contributions of this work are summarized as follows:

- We propose MasterWeaver, a novel method that can generate personalized images with high efficiency, faithful identity, and flexible controllability.
- An editing direction loss is proposed to improve the editability while keeping the identity. A face-augmented dataset is constructed to facilitate disentangled identity learning, further improving edtiability.
- Experimental results show that MasterWeaver can generate the target identity faithfully, while showing more flexible text controllability.

2 Related Work

2.1 Text-to-Image Diffusion Models

Recently, diffusion models [16, 27] have demonstrated remarkable capabilities in generating photo-realistic images and have been widely adopted in text-to-image (T2I) generation [5, 7, 17, 56, 60, 66, 68, 72]. Benefiting from the advance in language model [64, 65, 74] and large-scale text-image datasets [74], these T2I diffusion models are capable of generating textually coherent and high-quality

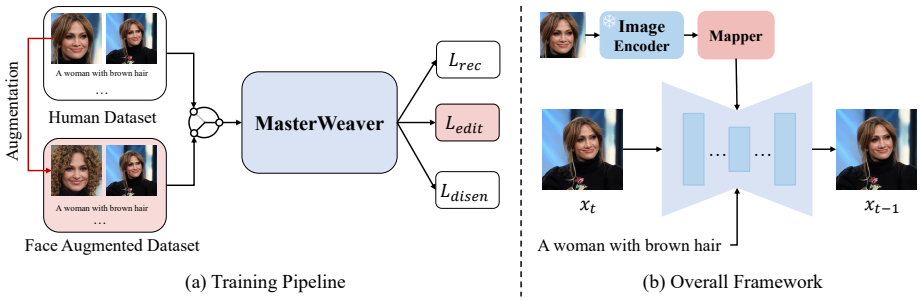

Fig. 2: (a) Training pipeline of our MasterWeaver. Specifically, to improve the editability while maintaining identity fidelity, we propose an editing direction loss $\mathcal{L}_{edit}$ for training. Additionally, we construct a face-augmented dataset to facilitate disentangled identity learning, further improving editability. **(b) Framework of our MasterWeaver.** It adopts an encoder to extract identity features and employ it with text to steer personalized image generation through cross attention.

images. Among them, Stable Diffusion [68] is one of the representative open-sourced models, which performs a diffusion process in the latent space to reduce computational complexity. It has demonstrated the superior capacity in generating high-quality and diverse images, and facilitated a surge of recent advances in downstream tasks [31, 44, 52, 61, 92, 94, 95]. Stable Diffusion XL (SDXL) [60] employed a larger UNet and an additional text encoder to achieve superior image generation quality and enhanced textual fidelity. Stable Diffusion is also employed as the T2I model in our experiments.

2.2 Personalized Text-to-image Generation

Building upon the advances in T2I models, personalization methods [20,69,85,90] further enable the generation of specific visual concepts (*e.g.*, objects, animals, people) as indicated by one or several reference images. Earlier studies [3, 4, 8, 10, 18–20, 25, 28, 30, 36, 38, 48, 49, 51, 55, 58, 63, 69, 71, 77, 79, 84, 86, 93, 97, 99] usually learned the novel concept from reference images by optimizing word embeddings or model parameters. For example, Textual Inversion [20] optimized a new "word" embedding using a few reference images to learn the target concept. DreamBooth [69] finetuned the parameters of the T2I model to align the unique identifier with the high-fidelity new concept. To improve the computation efficiency, Custom Diffusion [36] only updated the key and value mapping parameters in cross attention layers. CelebBasis [91] projected the human identity into a celeb space, extracted from celebrity identities, and showed superior editability. Though the results are promising, these methods usually take multiple images to learn the concept and require substantial time for finetuning. To reduce the tuning time, several methods [21, 22] suggested initial pre-training on

large datasets followed by tuning on smaller datasets to expedite the process. However, they still take tens of steps for finetuning, limiting their practicality.

Recent tuning-free methods [2, 9, 12, 29, 33, 39, 54, 62, 76, 85, 96] train additional visual encoders to extract the concept information, and inject them via model tuning or adapters. By training on personalization datasets, these methods could produce personalized results in tens of denoising steps with only one reference image. For example, ELITE [85] employed a global mapping to encode the CLIP [64] features of concept image as textual embedding, and further improved finer details with a local mapping network. BLIP Diffusion utilized a pre-trained BLIP2 [40] as the feature extractor to produce a visual representation aligned with the text. Additionally, several methods [11, 13, 32, 41, 43, 47, 59, 70, 78, 81, 82, 88, 89] are designed for personalization of human identities. DreamIdentity [13] crafted an editing dataset based on celebrity identity and a pre-trained stable diffusion model for editability learning. Photomaker [42] collected a human dataset that consists of multiple images of the same identity and used them to learn a stacked face embedding, which could generate images with diverse attributes.

In this work, we propose the tuning-free MasterWeaver. Compared with existing methods, MasterWeaver could generate personalized images with high efficiency, faithful identity preservation, and flexible controllability.

2.3 Face Editing

Face generation and editing are popular topics in computer vision and computer graphics. With the development of Generative Adversarial Networks (GANs) [24], several methods have been proposed for high-quality face generation [34, 35] and flexible face editing [6, 45, 53, 57, 75]. Generally speaking, existing GAN-based face editing methods can be classified into two types. The first type of methods [14, 15, 46] utilize image-to-image translation techniques, where the original face image is fed to the network to produce the edited image. The second type of methods [6, 45, 75] first invert the given face into the latent code of pre-trained GANs [1, 67, 83], and then manipulate it with specific directions. Recently, with the advance of CLIP [64], several methods [23, 53, 57] have been proposed to perform face editing based on text prompts.

In our work, we adopt the face editing method in [53] to construct a face-augmented dataset, which can be used for disentangled identity training.

3 Proposed Method

Given one single reference image x of a specific identity, personalized T2I generation aims to generate photo-realistic images of the given identity according to the text prompts y. Nonetheless, the generated images are expected to keep a faithful identity and exhibit diversity in attributes, actions, contexts, and so on. To achieve this goal, we propose MasterWeaver, a tuning-free method for generating personalized images with promising identity fidelity and flexible editability. As

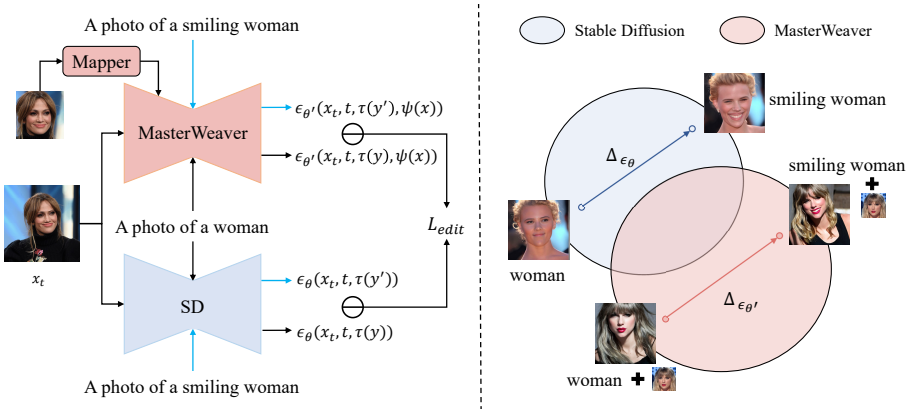

Fig. 3: Illustration of Editing Direction Loss. By inputting paired text prompts that denote an editing operation, *e.g.*, (a photo of a woman, a photo of a smiling woman), we identify the editing direction in the feature space of diffusion model. Then we align the editing direction of MasterWeaver with that of original T2I model to improve the text controllability without affecting the identity.

illustrated in Fig. 2, MasterWeaver first employs an identity mapping network to encode identity feature, and incorporates it with text to guide the image generation. To improve the model editability while keeping identity fidelity, we further propose the id-preserved editability learning, including an editing direction loss and a face-augmented dataset. In the following, we first present an overview of the T2I model utilized in our approach (Sec. 3.1). Then, we introduce the details of the face identity injection (Sec. 3.2) and id-preserved editability learning (Sec. 3.3). Finally, we give our learning objective (Sec. 3.4).

3.1 Preliminary

In this work, we employ the large-scale pre-trained Stable Diffusion (SD) [68] as our T2I model, which can generate diverse and photo-realistic images. It consists of two components, *i.e.*, the autoencoder ($\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$) and the conditional diffusion model $\epsilon_\theta(\cdot)$. Specifically, the encoder $\mathcal{E}(\cdot)$ is trained to map an image x to a lower dimensional latent space $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}(\cdot)$ is trained to map the latent code back to the image so that $D(\mathcal{E}(x)) \approx x$. The conditional diffusion model $\epsilon_\theta(\cdot)$ is trained on the latent space of autoencoder, which can generate latent codes based on text condition y. The mean-squared loss is employed to train the model:

$$\mathcal{L}_{rec} = \mathbb{E}_{z,y,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ denotes the unscaled noise, t is the time step, z_t is the latent noise at time t, and $\tau(\cdot)$ represents the pretrained CLIP text encoder [64]. During inference, it starts from a random Gaussian noise z_T and iteratively denoises it to z_0. The decoder then maps it to the final image $x' = \mathcal{D}(z_0)$.

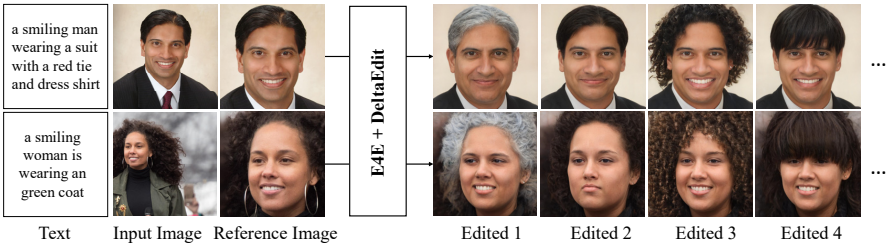

Fig. 4: Construction of Face-Augmented Dataset. We employ E4E [67] and DeltaEdit [53] to edit the attribute of the reference identity image, and construct the face-augmented dataset.

Cross attention is adopted in SD to steer the generation process by text prompt. Specifically, cross attention adopts the latent image feature h and text feature $\tau(y)$ as input, and transforms them to query $Q = W_Q \cdot h$, key $K = W_K \cdot \tau(y)$ and value $V = W_V \cdot \tau(y)$ by projection layers. W_Q, W_K, and W_V are weight parameters of query, key, and value projection layers, respectively. Then, attention maps are computed with:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d'}}\right)V, \tag{2}$$

where d' is the output dimension of key and query features.

3.2 Faithful Identity Injection

To leverage the pre-trained SD for generating the personalized images of specified identity, we first introduce a mapping network to encode the identity features, and then integrate the encoded features into the image generation process. As shown in Fig. 2 (b), we employ the pre-trained CLIP image encoder $\psi(\cdot)$ to extract the identity information. To ensure identity fidelity, we utilize the local patch image feature from the penultimate layer of the CLIP model, which contains rich details and is sufficient to represent the target identity faithfully. To bridge the gap between the CLIP features and the SD model, we further employ an identity (ID) mapper $M(\cdot)$, which is trained to project the CLIP features as identity feature f:

$$f = M \circ \psi(x), \tag{3}$$

where $f \in \mathbb{R}^{N \times d}$. N denotes the number of tokens and d is the feature dimension. Here, we also mask the background of x to ease the irrelevant disturbances.

To effectively integrate the identity feature f into the generation, we adopt the dual cross attention mechanism [85, 90]: $\text{Attention}(Q, K^f, V^f)$, where $K^f = W_K^f \cdot f$ and $V^f = W_V^f \cdot f$ represent the projected key and value matrices of the identity information, respectively. W_K^f and W_V^f are two additionally introduced

Fig. 5: Visual comparison of different methods. All images are generated using the single reference image shown on the left. Our MasterWeaver can generate high-quality images with flexible editability and faithful identity. Zoom in for a better view.

projection layers in each cross attention block of SD, which are optimized with our ID mapper simultaneously. To fuse the identity information with the text information, we sum them by:

$$Out = \text{Attention}(Q, K, V) + \lambda \text{Attention}(Q, K^f, V^f), \tag{4}$$

where λ is a trade-off hyperparameter and set as 1 during training.

3.3 ID-Preserved Editability Learning

Although the proposed method in Sec. 3.2 can generate personalized images with faithful identity, its text controllability is limited. Since the reference identity image and input image used in training are from the same image (as shown in Fig. 4 left), when we train the model under the reconstruction paradigm, the learned identity feature is inevitably entangled with facial attributes (*e.g.*, hair, pose and expression). Such an entanglement weakens the control of the text over generated images, and the model tends to directly copy reference images during generation, resulting in poor editability and diversity. To improve editability

while keeping a faithful identity, we propose id-preserved editability learning, which consists of an editing direction loss and a face-augmented dataset.

Editing Direction Loss. To improve the text controllability, we first propose an editing direction loss to facilitate the model learning. The SD model is well-recognized for its superior text controllability and can generate images that closely align with the provided textual descriptions. Intuitively, we can employ its editing capability to regularize MasterWeaver. As shown in Fig. 3, by leveraging paired text prompts that indicate an editing operation, *e.g.*, (y = a photo of a person, y' = a photo of a person with <attribute>), we can identify an editing direction in the feature space of diffusion UNet model:

$$\Delta_{\epsilon_\theta}(y, y') = \epsilon_\theta(z_t, t, \tau(y')) - \epsilon_\theta(z_t, t, \tau(y)). \tag{5}$$

Such an editing direction captures the meaningful prior of SD for semantic editing. Then, we align the editing directions of MasterWeaver with those of SD during training:

$$\mathcal{L}_{edit} = 1 - \text{Sim}(m \cdot \Delta_{\epsilon_\theta}(y, y'), m \cdot \Delta_{\epsilon_{\theta'}}(y, y', x)), \tag{6}$$

where $\epsilon_{\theta'}$ denotes our MasterWeaver model and $\text{Sim}(\cdot, \cdot)$ denotes the calculation of cosine similarity. m is the mask of the facial region to ensure the loss is only calculated on the facial area. Obviously, if MasterWeaver neglects the text prompts during generation, its editing directions will be different from those of SD, resulting in a large loss. Besides, the calculation of editing direction is solely based on the text difference, and the encoded semantic information is unrelated to the identity. Therefore, training with our proposed direction loss can improve the model's editability significantly, with moderate sacrifice of identity fidelity.

Furthermore, our editing direction $\Delta_{\epsilon_\theta}(y, y')$ can be seen as a variant of DDS [26], which is designed to optimize the edited image by minimizing the delta noise: $\mathcal{L}_{dds} = \|\epsilon_\theta(z_t, t, \tau(y)) - \epsilon_\theta(\hat{z}_t, t, \tau(y'))\|_2^2$, where z and $\hat{z}$ are original and edited images, y and y' are source and target prompts. In $\mathcal{L}_{edit}$, editing direction $\Delta_{\epsilon_\theta}(y, y') = \epsilon_\theta(z_t, t, \tau(y')) - \epsilon_\theta(z_t, t, \tau(y))$ is used to represent the editing prior of original model. By aligning the editing direction with that of the original model, MasterWeaver can inherit its editability. In our implementation, we use the same z_t for both source prompt y and target prompt y', and the direction is calculated in feature space instead of noise space. We have collected several attribute-related prompt pairs for training, including hair, age, expression, *etc.* More details can be found in the Suppl.

Face-Augmented Dataset Construction. As we analyzed, one of the reasons behind the insufficient editability is that the learned model entangles the identity with irrelevant attributes and tends to copy the reference image directly, neglecting the text prompt. To address this issue, we build a face-augmented dataset designed for disentangled identity training. As illustrated in Fig. 4, we employ the E4E [67] and DeltaEdit [53] to perform the attribute editing for the reference identity image. Then, we construct our face-augmented dataset by combining the input image, text prompt, and edited face images. In this dataset,

Table 1: Quantitative comparison under single reference setting. We employ CLIP-T metric to measure the text alignment, and utilize DINO, CLIP-I and Face Similarity metrics to measure the identity fidelity. The best result is shown in **bold**, and the second best is underlined.

	CLIP-T (↑)	CLIP-I (↑)	DINO (↑)	Face Sim. (↑)	Speed (s, ↓)
Textual Inversion [20]	0.167	0.669	0.556	0.423	3000
DreamBooth [69]	0.223	0.641	0.509	0.469	908
Custom Diffusion [36]	0.210	0.726	**0.647**	0.626	548
CelebBasis [91]	0.214	0.651	0.501	0.439	490
FastComposer [87]	0.201	**0.747**	0.641	**0.631**	**3**
IPAdapter [90]	0.192	0.714	0.632	0.607	**3**
PhotoMaker [42]	0.231	0.667	0.497	0.544	10
Ours	**0.232**	0.726	0.638	**0.631**	4

the reference identity image and corresponding input image have the same identity but differ in one specific attribute (*e.g.*, hair color, style, or expression). Such a controlled attribute misalignment can effectively facilitate model learning to extract faithful identity features disentangled with attribute details, thereby improving editability. To maintain identity fidelity, we filter to exclude edited face images with lower face similarity to reference images. In our experiments, we utilize a mix of face-augmented and original datasets for balanced editability and identity fidelity. More details of the dataset are provided in the Suppl.

3.4 Learning Objective

Following [41], we employ a background disentanglement loss to encourage the identity feature only controlling the generation of facial region. Specifically, when the identity feature changes, we regularize the background to be unchanged:

$$\mathcal{L}_{disen} = \|(1-m)\cdot[\epsilon_{\theta'}(z_t, t, \tau(y), f) - \epsilon_{\theta'}(z_t, t, \tau(y), A(f))]\|, \tag{7}$$

where $f = M \circ \psi(x)$ is the extracted identity feature and $1-m$ denotes the mask of background region. $A(\cdot)$ denotes the augmentation operation.

The overall learning objective is defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{edit}\cdot\mathcal{L}_{edit} + \lambda_{disen}\cdot\mathcal{L}_{disen}, \tag{8}$$

where λ_{disen} and λ_{edit} denote the trade-off parameters.

4 Experiments

4.1 Experimental Settings

Training dataset. To train our MasterWeaver, we build a dataset including about 160k text-image pairs from the LAION-Face dataset [98]. We have also prepared the corresponding captions and face masks for model training. The

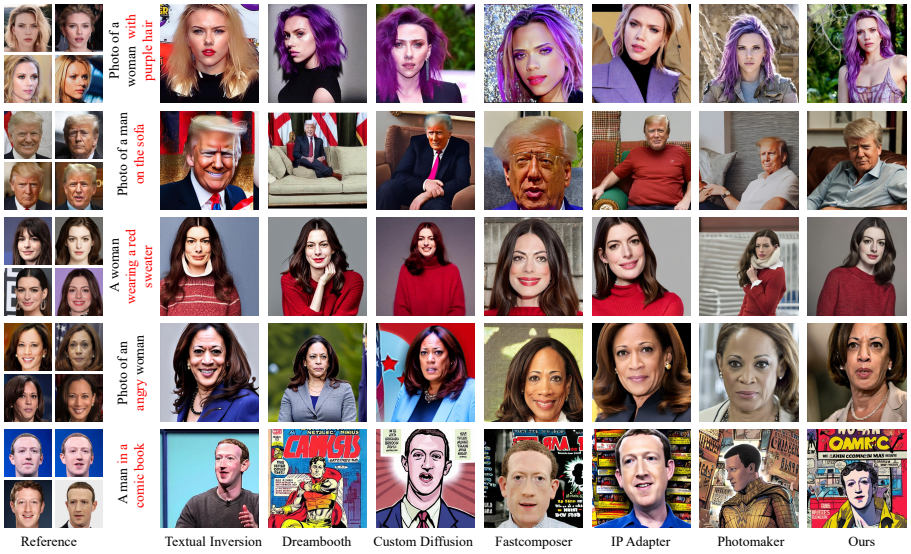

Fig. 6: Visual comparison of different methods. All images are generated using the four reference images shown on the left. Our MasterWeaver consistently shows better editability and identity. Zoom in for a better view.

detailed pipeline of dataset is provided in the Suppl. Moreover, as described in Sec. 3.3, we have built a face-augmented dataset based on filtered images.

Evaluation dataset. To evaluate the one-shot personalization, we randomly select 300 identities from the CelebA-HQ dataset [37], and each identity has one image as the reference. Besides, our method can be extended for personalized generation using multiple reference images without additional training. Here, we also evaluate our method with multiple images as the reference. Following [42], we collect a dataset consisting of 25 identities, and each identity contains four images for evaluation. For quantitative analysis, we employ 50 prompts encompassing a range of clothing, styles, attributes, actions, and backgrounds. We randomly generate five images for each identity-prompt pair. More details are provided in Suppl. For visual comparison, we utilize the identity images collected from existing methods for a convenient comparison. Unless specifically mentioned, all images presented in this paper are produced using a single reference image.

Implementation details. We employ SD V1.5 in our experiments, and our model is trained using a batch size of 16 and a learning rate of 1e-6. To calculate the $\mathcal{L}_{edit}$, we utilize the output features of the cross attention blocks in the first decoder layer. λ_{edit} is set as 0.01 and λ_{disen} is set as 1. To enable classifier-free guidance, we use a probability of 0.05 to drop text and image individually, and a probability of 0.05 to drop text and image simultaneously. All experiments are conducted on 4×A800 GPUs with AdamW [50] optimizer. During image generation, we use 50 steps of the LMS sampler, and set the scale of classifier-

Table 2: Quantitative comparisons under multiple references (*i.e.*, 4 images) setting. The best result is shown in **bold**, and the second best is underlined.

	CLIP-T (↑)	CLIP-I (↑)	DINO (↑)	Face Sim. (↑)	Speed (s, ↓)
Textual Inversion [20]	0.172	0.729	0.594	0.618	3000
DreamBooth [69]	0.220	0.704	0.511	0.524	908
Custom Diffusion [36]	0.218	0.754	**0.598**	0.615	548
FastComposer [87]	0.210	0.720	0.582	0.622	**3**
IPAdapter [90]	0.200	**0.757**	0.592	0.632	**3**
PhotoMaker [42]	0.222	0.663	0.520	0.547	10
Ours	**0.230**	0.752	0.596	**0.646**	4

Fig. 7: Ablation Study. The proposed losses and augmented dataset (denotes as aug) significantly improve the editability of the model. Besides, the identity fidelity consistently improves as the number increases.

free guidance as 5. More details of our model and implementation are provided in the Suppl.

Evaluation metrics. Following Dreambooth [69], we employ CLIP-T metric to measure the text alignment of generated images, and utilize DINO and CLIP-I metrics to measure the identity fidelity. We also employ the FaceNet [73] to compute the face similarity in the detected facial regions as another measure of identity fidelity. Moreover, we adopt the inference speed to evaluate the efficiency of each method. For those optimization-based methods, we combine the time required for optimization with the inference time as overall speed. All the inference speeds are tested on a single NVIDIA Tesla V100 GPU.

4.2 Qualitative Comparison

To demonstrate the effectiveness of our MasterWeaver, we conduct the comparisons with existing methods, including Textual Inversion [20], Dreambooth [69], Custom Diffusion [36], CelebBasis [91], Fastcomposer [87], IP Adapter [90], and Photomaker [42]. For a fair comparison, all methods utilize the SD 1.5 version, except for Photomaker, which uses the SDXL model.

Fig. 5 presents visual comparisons with existing methods using a single image as a reference. As depicted, MasterWeaver can generate photo-realistic

Table 3: Ablation Study. The proposed background disentanglement loss L_{disen}, editing direction loss L_{edit}, and the face-augmented dataset (denote as aug) can improve the editability of MasterWeaver, while keeping the identity fidelity.

	CLIP-T (↑)	CLIP-I (↑)	DINO (↑)	Face Sim. (↑)
Ours w/o $\mathcal{L}_{disen}$	0.226	0.724	0.633	0.617
Ours w/o $\mathcal{L}_{edit}$	0.209	**0.739**	**0.643**	**0.637**
Ours w/o aug	0.221	0.730	0.640	0.634
Ours	**0.232**	0.726	0.638	0.631

personalized images in diverse scenarios, including modifications to attributes, clothing, background, and style. Optimization-based methods, *i.e.*, Textual Inversion, Dreambooth, and Custom Diffusion, suffer from the overfitting in this limited-data scenario, leading to limited text alignment and compromised identity fidelity (*e.g.*, rows 1∼3). Tuning-free methods, *i.e.*, Fastcomposer and IP Adapter maintain high identity fidelity but fall short in editing style and facial attributes (*e.g.*, face attributes in rows 1 and 3, and the style in the last two rows). Photomaker shows improved text controllability; however, it sometimes fails to generate faithful identity (*e.g.*, rows 1, 2, and 5). In comparison, our MasterWeaver can generate personalized images that are faithful to both reference identity and text prompts. Fig. 1 illustrates more results generated by our method. MasterWeaver can generate photo-realistic images with diverse clothing, accessories, facial attributes, and actions. Additionally, it allows for simultaneously editing multiple attributes.

MasterWeaver can also be directly extended for personalized image generation with multiple reference images of the same identity. Specifically, during inference, we simply concatenate the identity features of different images along the token dimension. As shown in Fig. 7, using more reference images improves the identity fidelity of the generated personalized images. Furthermore, we compare our method with competing methods in the setting with multiple reference images (*i.e.*, with four reference images). From Fig. 6, one can see that MasterWeaver still outperforms competitors in identity fidelity and text controllability. More qualitative results can be found in Suppl.

4.3 Quantitative Comparison

In addition to the qualitative comparisons, we further conduct the quantitative evaluation to validate the performance of MasterWeaver. Table 1 reports the comparisons conducted with a single reference image. From the table, we see that MasterWeaver outperforms existing state-of-the-art methods regarding text alignment and face similarity, demonstrating its ability to generate personalized images that maintain consistent identity and good alignment with texts. Moreover, MasterWeaver exhibits a competitive inference speed, requiring only 4s to generate an image on a single V100 GPU. Table 2 reports similar findings in a scenario utilizing four reference images, further demonstrating the effectiveness

of MasterWeaver. While MasterWeaver's scores for CLIP-I and DINO are not the highest, it should be noted that these metrics are not specifically designed for facial analysis and can be affected by extraneous factors such as pose. Additionally, we have conducted a user study to compare MasterWeaver with other methods. The detailed results are provided in Suppl.

4.4 Ablation Study

We have conducted ablation studies to evaluate the roles of various components in our method, including the background disentanglement loss $\mathcal{L}_{disen}$, editing direction loss $\mathcal{L}_{edit}$, and the face-augmented dataset.

Effect of the disentanglement loss $\mathcal{L}_{disen}$. We first evaluate the impact of $\mathcal{L}_{disen}$. As illustrated in Fig. 7, without $\mathcal{L}_{disen}$, the model fails to generate images with desired clothing. This suggests that $\mathcal{L}_{disen}$ could ease the influence of identity information on the background, thereby enhancing the editability of the generated images. From Table 3, both text alignment and face similarity metrics decrease without $\mathcal{L}_{disen}$, demonstrating its effectiveness.

Effect of the editing direction loss $\mathcal{L}_{edit}$. We then assess the effect of $\mathcal{L}_{edit}$. As shown in Fig. 7, without $\mathcal{L}_{edit}$, MasteWeaver fails to generate the identities with proper facial attributes (*i.e.*, old woman). From Table 3, we see that the CLIP-T score drops significantly without $\mathcal{L}_{edit}$. Though the preservation of identity drops slightly, it is acceptable.

Effect of the face-augmented dataset. We further evaluate the effect of our constructed face-augmented dataset. Fig. 7 demonstrates that the model trained with the face-augmented dataset exhibits improved editability (*e.g.*, with the augmented dataset, MasterWeaver successfully generates the personalized image with a smiling expression). The results in Table 3 further confirm its effectiveness, as the text alignment score increases when using the augmented dataset. More ablations are provided in Suppl.

5 Conclusion

In this work, we proposed MasterWeaver, a novel tuning-free method capable of generating personalized images with high efficiency, faithful identity, and flexible editability. The proposed editing direction loss and face-augmented dataset significantly improved the model's editability while maintaining identity fidelity. Extensive experiments have demonstrated that MasterWeaver outperformed the state-of-the-art methods and generated photo-realistic images that are faithful to both identity and texts. Our method was versatile for various applications, including personalized digital content creation and artistic endeavors. Moreover, the proposed editing direction loss had the potential to be applied to other domains (*e.g.*, animals and objects), enlarging its applicability.

Acknowledgement. This work was supported in part by the National Key R&D Program of China (2021YFF0900500), and the National Natural Science Foundation of China (NSFC) under Grant No. 62441202, and the Hong Kong RGC RIF grant (R5001-18).

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019) 5
2. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., H. Bermano, A.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 5
3. Arar, M., Voynov, A., Hertz, A., Avrahami, O., Fruchter, S., Pritch, Y., Cohen-Or, D., Shamir, A.: Palp: Prompt aligned personalization of text-to-image models. arXiv preprint arXiv:2401.06105 (2024) 4
4. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311 (2023) 4
5. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022) 3
6. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:1811.10597 (2018) 5
7. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 3
8. Cai, Y., Wei, Y., Ji, Z., Bai, J., Han, H., Zuo, W.: Decoupled textual embeddings for customized image generation. arXiv preprint arXiv:2312.11826 (2023) 4
9. Chae, D., Park, N., Kim, J., Lee, K.: Instructbooth: Instruction-following personalized text-to-image generation. arXiv preprint arXiv:2312.03011 (2023) 5
10. Chen, H., Zhang, Y., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023) 4
11. Chen, L., Zhao, M., Liu, Y., Ding, M., Song, Y., Wang, S., Wang, X., Yang, H., Liu, J., Du, K., et al.: Photoverse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793 (2023) 5
12. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024) 5
13. Chen, Z., Fang, S., Liu, W., He, Q., Huang, M., Zhang, Y., Mao, Z.: Dreamidentity: Improved editability for efficient face-identity preserved image generation. arXiv preprint arXiv:2307.00300 (2023) 5
14. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018) 5
15. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020) 5
16. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 3

17. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217 (2022) 3
18. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. arXiv preprint arXiv:2211.11337 (2022) 4
19. Fei, Z., Fan, M., Huang, J.: Gradient-free textual inversion. arXiv preprint arXiv:2304.05818 (2023) 4
20. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 2, 4, 10, 12
21. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Designing an encoder for fast personalization of text-to-image models. arXiv preprint arXiv:2302.12228 (2023) 4
22. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023) 4
23. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022) 5
24. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 5
25. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Detail-preserving visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023) 4
26. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023) 9
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
28. Hong, Y., Zhang, J.: Comfusion: Personalized subject generation in multiple specific scenes from single image. arXiv preprint arXiv:2402.11849 (2024) 4
29. Hu, H., Chan, K.C., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al.: Instruct-imagen: Image generation with multi-modal instruction. arXiv preprint arXiv:2401.01952 (2024) 5
30. Hua, M., Liu, J., Ding, F., Liu, W., Wu, J., He, Q.: Dreamtuner: Single image is enough for subject-driven generation. arXiv preprint arXiv:2312.13691 (2023) 4
31. Huang, T., Zeng, Y., Zhang, Z., Xu, W., Xu, H., Xu, S., Lau, R.W., Zuo, W.: Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. arXiv preprint arXiv:2312.06439 (2023) 4
32. Hyung, J., Shin, J., Choo, J.: Magicapture: High-resolution multi-concept portrait customization. arXiv preprint arXiv:2309.06895 (2023) 5
33. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023) 5
34. Karras, T., Laine, S., Aila, T.: 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) 5

35. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: 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) 5
36. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. arXiv preprint arXiv:2212.04488 (2022) 4, 10, 12
37. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 11
38. Lee, K., Kwak, S., Sohn, K., Shin, J.: Direct consistency optimization for compositional text-to-image personalization. arXiv preprint arXiv:2402.12004 (2024) 4
39. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36** (2024) 5
40. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 5
41. Li, X., Hou, X., Loy, C.C.: When stylegan meets stable diffusion: a w+ adapter for personalized image generation. arXiv preprint arXiv:2311.17461 (2023) 5, 10
42. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. arXiv preprint arXiv:2312.04461 (2023) 2, 5, 10, 11, 12
43. Liang, C., Ma, F., Zhu, L., Deng, Y., Yang, Y.: Caphuman: Capture your moments in parallel universes. arXiv preprint arXiv:2402.00627 (2024) 5
44. Lin, J., Zhang, Z., Wei, Y., Ren, D., Jiang, D., Zuo, W.: Improving image restoration through removing degradations in textual representations. arXiv preprint arXiv:2312.17334 (2023) 4
45. Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: Editgan: High-precision semantic image editing. Advances in Neural Information Processing Systems **34**, 16331–16345 (2021) 5
46. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: Stgan: A unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3673–3682 (2019) 5
47. Liu, R., Ma, B., Zhang, W., Hu, Z., Fan, C., Lv, T., Ding, Y., Cheng, X.: Towards a simultaneous and granular identity-expression control in personalized face generation. arXiv preprint arXiv:2401.01207 (2024) 5
48. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125 (2023) 4
49. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects. arXiv preprint arXiv:2305.19327 (2023) 4
50. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 11
51. Lu, J., Xie, C., Guo, H.: Object-driven one-shot fine-tuning of text-to-image diffusion with prototypical embedding. arXiv preprint arXiv:2401.15708 (2024) 4
52. Lv, Z., Wei, Y., Zuo, W., Wong, K.Y.K.: Place: Adaptive layout-semantic fusion for semantic image synthesis. IEEE Conference on Computer Vision and Pattern Recognition (2024) 4

53. Lyu, Y., Lin, T., Li, F., He, D., Dong, J., Tan, T.: Deltaedit: Exploring text-free training for text-driven image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6894–6903 (2023) 3, 5, 7, 9
54. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023) 5
55. Nam, J., Kim, H., Lee, D., Jin, S., Kim, S., Chang, S.: Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. arXiv preprint arXiv:2402.09812 (2024) 4
56. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 3
57. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021) 5
58. Patel, M., Jung, S., Baral, C., Yang, Y.: λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. arXiv preprint arXiv:2402.05195 (2024) 4
59. Peng, X., Zhu, J., Jiang, B., Tai, Y., Luo, D., Zhang, J., Lin, W., Jin, T., Wang, C., Ji, R.: Portraitbooth: A versatile portrait model for fast identity-preserved personalization. arXiv preprint arXiv:2312.06354 (2023) 5
60. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 3, 4
61. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 4
62. Purushwalkam, S., Gokul, A., Joty, S., Naik, N.: Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. arXiv preprint arXiv:2401.13974 (2024) 5
63. Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., Schölkopf, B.: Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems **36** (2024) 4
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3, 5, 6
65. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 3
66. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 3
67. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2287–2296 (2021) 5, 7, 9
68. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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) 2, 3, 4, 6

69. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022) 2, 4, 10, 12
70. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023) 5
71. Ryu, H., Lim, S., Shim, H.: Memory-efficient personalization using quantized diffusion model. arXiv preprint arXiv:2401.04339 (2024) 4
72. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022) 3
73. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015) 12
74. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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) 3
75. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence **44**(4), 2004–2018 (2020) 5
76. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023) 5
77. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 4
78. Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: Face0: Instantaneously conditioning a text-to-image model on a face. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 2, 5
79. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023) 4
80. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV). pp. 589–604 (2018) 2
81. Wang, Q., Jia, X., Li, X., Li, T., Ma, L., Zhuge, Y., Lu, H.: Stableidentity: Inserting anybody into anywhere at first sight. arXiv preprint arXiv:2401.15975 (2024) 5
82. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024) 5
83. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11379–11388 (2022) 5
84. Wang, Z., Wei, W., Zhao, Y., Xiao, Z., Hasegawa-Johnson, M., Shi, H., Hou, T.: Hifi tuner: High-fidelity subject-driven fine-tuning for diffusion models. arXiv preprint arXiv:2312.00079 (2023) 4
85. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023) 2, 3, 4, 5, 7
86. Wu, Z., Yu, C., Zhu, Z., Wang, F., Bai, X.: Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. arXiv preprint arXiv:2310.08094 (2023) 4

87. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023) 2, 10, 12
88. Yan, Y., Zhang, C., Wang, R., Zhou, Y., Zhang, G., Cheng, P., Yu, G., Fu, B.: Facestudio: Put your face everywhere in seconds. arXiv preprint arXiv:2312.02663 (2023) 5
89. Yang, Y., Wang, R., Qian, Z., Zhu, Y., Wu, Y.: Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. arXiv preprint arXiv:2306.08247 (2023) 5
90. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 2, 3, 4, 7, 10, 12
91. Yuan, G., Cun, X., Zhang, Y., Li, M., Qi, C., Wang, X., Shan, Y., Zheng, H.: Inserting anybody in diffusion models via celeb basis. arXiv preprint arXiv:2306.00926 (2023) 4, 10, 12
92. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 4
93. Zhang, X.L., Wei, X.Y., Wu, J.L., Zhang, T.Y., Zhang, Z.X., Lei, Z., Li, Q.: Compositional inversion for stable diffusion models. arXiv preprint arXiv:2312.08048 (2023) 4
94. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023) 4
95. Zhang, Y., Wei, Y., Lin, X., Hui, Z., Ren, P., Xie, X., Ji, X., Zuo, W.: Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. arXiv preprint arXiv:2403.05438 (2024) 4
96. Zhang, Y., Liu, J., Song, Y., Wang, R., Tang, H., Yu, J., Li, H., Tang, X., Hu, Y., Pan, H., et al.: Ssr-encoder: Encoding selective subject representation for subject-driven generation. arXiv preprint arXiv:2312.16272 (2023) 5
97. Zhao, R., Zhu, M., Dong, S., Wang, N., Gao, X.: Catversion: Concatenating embeddings for diffusion-based text-to-image personalization. arXiv preprint arXiv:2311.14631 (2023) 4
98. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18697–18709 (2022) 10
99. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv preprint arXiv:2305.13579 (2023) 4