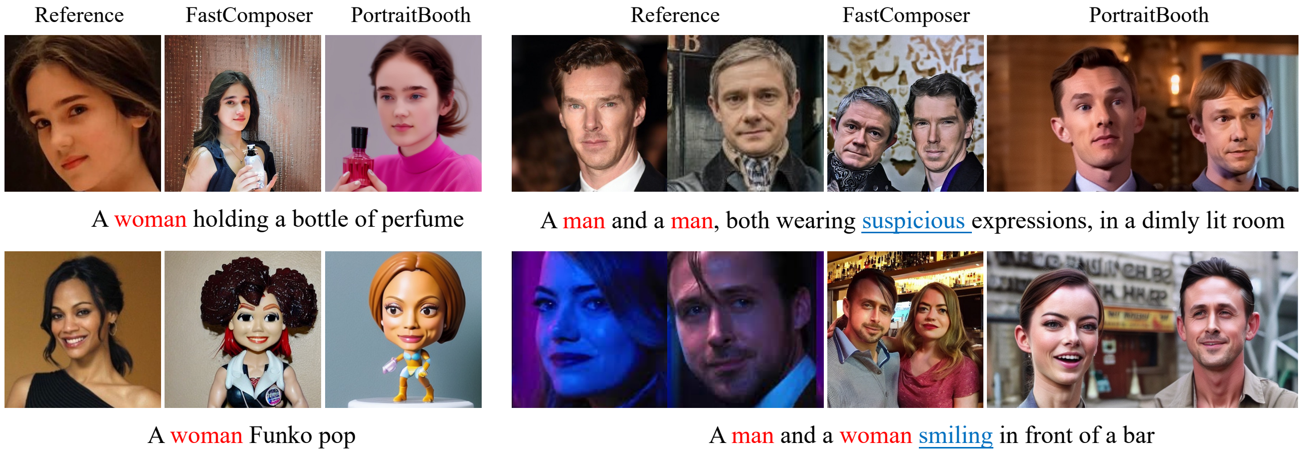
PortraitBooth: A Versatile Portrait Model for Fast Identity-preserved Personalization

Xu Peng[1,2], Junwei Zhu[3], Boyuan Jiang[3], Ying Tai[4], Donghao Luo[3], Jiangning Zhang[3],
Wei Lin[1,2], Taisong Jin[1,2†], Chengjie Wang[3], Rongrong Ji[1,2]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China.
[2]School of Informatics, Xiamen University, China. [3]Tencent Youtu Lab.
[4]School of Intelligence Science and Technology, Nanjing University, Suzhou, China.
{penglingxiao,lvvviolette}@stu.xmu.edu.cn, tyshiwo@gmail.com
{junweizhu,byronjiang,michaelluo,vtzhang,jasoncjwang}@tencent.com
{jintaisong,rrji}@xmu.edu.cn



A woman holding a bottle of perfume

A man and a man, both wearing suspicious expressions, in a dimly lit room

A woman Funko pop

A man and a woman smiling in front of a bar

Figure 1. **Qualitative comparison of PortraitBooth and FastComposer** on action, style, expression editing, multi-subject generation, and identity preservation, all without any test-time tuning.

Abstract

Recent advancements in personalized image generation using diffusion models have been noteworthy. However, existing methods suffer from inefficiencies due to the requirement for subject-specific fine-tuning. This computationally intensive process hinders efficient deployment, limiting practical usability. Moreover, these methods often grapple with identity distortion and limited expression diversity. In light of these challenges, we propose PortraitBooth, an innovative approach designed for high efficiency, robust identity preservation, and expression-editable text-to-image generation, without the need for fine-tuning. PortraitBooth leverages subject embeddings from a face recognition model for personalized image generation without fine-tuning. It eliminates computational overhead and mitigates identity distortion. The introduced dynamic identity preservation strategy further ensures close resemblance to the original image identity. Moreover, PortraitBooth incorporates emotion-aware cross-attention control for diverse facial expressions in generated images, supporting text-driven expression editing. Its scalability enables efficient and high-quality image creation, including multi-subject generation. Extensive results demonstrate superior performance over other state-of-the-art methods in both single and multiple image generation scenarios. Our project page is at https://portraitbooth.github.io.

1. Introduction

Recent years have witnessed remarkable progress in text-to-image synthesis [4, 22, 29, 41], propelled by the emergence of diffusion models [6, 15, 16, 31, 47]. Pre-trained text-to-image generation models have opened up new avenues

† Corresponding Author.

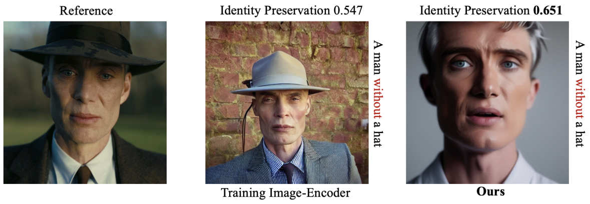


Figure 2. Comparison of identity information obtained based on the trained image encoder and pre-trained face recognition model.

for creative content creation, with personalized generation gaining popularity for its diverse applications.

Personalized generation methods based on diffusion models fall into two main categories: ***1)*** test-time fine-tuning and ***2)*** test-time non-fine-tuning. Some approaches [10, 13, 26, 33, 34] endorse test-time fine-tuning using reference images (typically 3-5) to generate personalized results. However, these methods require specialized network training [33, 37] , making them *inefficient for practical applications*. An alternative to test-time fine-tuning is retraining the base text-to-image model with specially designed strategies, *e.g*. training a distinct image encoder on a massive dataset to capture reference image identity information. However, these approaches [24, 39, 42] face challenges, either *dealing with identity distortion* or *generating images lacking editability*, as depicted in Fig. 2. This is mainly due to the coarse-grained nature of the identity information obtained from the trained image encoder. The better the image encoder is trained, the tighter the identity information with reference image is coupled, severely compromising editability. Additionally, these methods often demand significant GPU resources and high storage, making them impractical for most research institutions. Tab. 1 offers a comprehensive comparison of existing personalized image generation methods across four key aspects.

In this paper, we introduce PortraitBooth, a novel text-to-portrait personalization framework that achieves high efficiency, robust identity preservation, and diverse expression editing. We then describe our main characteristics in detail:

High Efficiency. PortraitBooth stands out as a highly efficient one-stage generation framework, delivering the following advantages: ***1)*** Only a *single* image is required during the inference stage, unlike other schemes such as Dreambooth that need multiple images. ***2)*** *No finetuning or optimization* is conducted during inference, which saves time and avoids delays. ***3)*** *Lower training resource requirement* is needed than Face0 and Subject-Diffusion that demand a lot of high-performance GPU resources.

Robust Identity Preservation. ***1)*** PortraitBooth employs a pre-trained face recognition model (41.5M parameters) to extract a face embedding from a given reference image. This embedding is then projected into the context space of Stable Diffusion using a simple multilayer perceptron, enabling high-fidelity image generation based on the proposed Subject Text Embedding Augmentation (STEA). ***2)*** PortraitBooth Dynamically maintains Identity Preservation (DIP) by incorporating an identity loss during training to facilitate the model to ensure identity preservation.

Methods	Single Image	Test-time None-fine-tuning	Robust ID Preservation	Expression Editing
Textual Inversion [10]	✗	✗	✗	✓
Dreambooth [33]	✗	✗	✗	✓
Custom Diffusion [22]	✗	✗	✗	✓
Break-A-Scene [2]	✓	✗	✓	✗
HyperDreamBooth [34]	✓	✗	✓	✓
FastComposer [42]	✓	✓	✗	✗
Face0 [39]	✓	✓	✓	✗
Subject-Diffusion [24]	✓	✓	✓	✗
PortraitBooth (Ours)	✓	✓	✓	✓

Table 1. **Comparisons of current personalization approaches**.

Diverse Expression Editing. While the discriminative features extracted from a robust face recognition model effectively disentangle identity and attributes, expression editing remains a challenge for existing one-shot methods [24]. To address this, we introduce Emotion-aware Cross-Attention Control (ECAC) via a truncation mechanism. This allows a single area to respond to multiple tokens simultaneously, thereby enabling versatile expression editing (see Fig. 1).

In summary, our contributions are threefold:

- We propose a novel one-shot text-to-portrait generation framework, termed PortraitBooth, which is the first solution to achieve high efficiency, robust identity preservation, and low training cost, simultaneously.
- To address identity distortion, we introduce the STEA and DIP modules for robust identity preservation. Additionally, we propose the ECAC module, achieving diverse expression editing.
- Our method scales effortlessly for single-subject and multi-subject generation, integrating smoothly with multi-object generation methods. Furthermore, our PortraitBooth excels in achieving remarkable fidelity and editability, surpassing other state-of-the-art methods.

2. Related Work

Image Editing with Diffusion Models. Image editing [12, 38] is a fundamental task in computer vision, involving modifications to an input image with auxiliary inputs like audio [50, 52], text [45], masks [12], or reference images [43, 44, 46, 49]. Despite the capabilities of large-scale diffusion models such as Imagen [35], DALL·E2 [30], and Stable Diffusion [31] in text-to-image synthesis, they lack precise control over image generation solely through text guidance. Even a small change in the original prompt can yield significantly different outcomes. Recent research has focused on adapting text-guided diffusion models [1, 8, 14, 17, 20, 21] for real image editing, leveraging their rich and diverse semantic knowledge. One such approach is Prompt-to-Prompt [14], which injects internal cross-attention maps when modifying only the text prompt, preserving the spatial layout and geometry necessary for regenerating an image while modifying it through prompt editing. Existing methods for portrait expression editing based on diffusion models not only focus on designing optimization-free methods [3, 7, 25, 27], but also explore face swapping as an alternative approach. For example, DiffusionRig [9] learns generic facial personalized priors to control face synthesis.

Personalized Visual Content Generation. Personalized visual content generation aims to create images tailored to individual preferences or characteristics, including new subjects described by one or more images [11]. Textual Inversion (TI) [10] and DreamBooth (DB) [33] are two pioneering works in personalization. They generate different contexts for a single visual concept using multiple images. TI introduces a learnable text token and optimizes it for concept reconstruction using standard diffusion loss, while keeping model weights frozen. DB reuses a rare token and fine-tunes model weights for reconstruction. HyperDreamBooth [34] offers a lightweight, subject-driven personalization for text-to-image diffusion models compared to DB. Custom Diffusion [22] fine-tunes subset layers of the cross-attention in the UNet. However, these tuning-based methods require time-consuming fine-tuning or multiple images, which is inelegant. In contrast, PortraitBooth amortizes costly subject tuning during training, enabling fast personalization with a single image.

Concurrent tuning-free methods include [24, 39, 42], those use an image encoder for accessibility, but Fastcomposer may distort identity due to lack of fine-grained training. Face0 [39] and Subject-Diffusion [24] achieve relatively high identity preservation in personalized generation through massive datasets and expensive hardware resources. However, they require resource-intensive backpropagation. Conversely, PortraitBooth generates personalized portraits with comparable identity preservation in an inference-only manner, requiring fewer hardware resources that most research institutions can afford.

3. Preliminaries

3.1. Stable Diffusion

Stable Diffusion (SD) consists of three components: a Variational AutoEncoder (VAE), a conditional U-Net [32], and a text encoder [28]. Specifically, for an input image x_0, The VAE encoder $\mathcal{E}$ compresses the x_0 to a smaller latent representation z. The diffusion process is then performed on the latent space, where a conditional U-Net denoiser ϵ_θ, denoises the noisy latent representation by predicting the noise θ with current timestep t, t-th noisy latent z_t. This denoising process can be conditioned on textual conditional C through the cross-attention mechanism, Throughout the training process, the network is optimized to minimize the loss function defined as:

$$\mathcal{L}_{noise} = \mathbb{E}_{z \sim \mathcal{E}(x), C, \epsilon \sim \mathcal{N}(0,1), t}\left[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2\right], \tag{1}$$

$$z_t \sim \mathcal{N}(\sqrt{\alpha_t} z_{t-1}, 1 - \alpha_t),$$

where α_t is a predefined sequence of coefficients controlling the variance schedule. The closed form of the distribution $p(z_t|z_0)$ can be easily derived as:

$$\begin{aligned} z_t &= \sqrt{\bar{\alpha}_t} z_0 + (1 - \bar{\alpha}_t)\epsilon, \\ \bar{\alpha}_t &= \prod_{s=1}^{t} \alpha_s, \epsilon \sim \mathcal{N}(0, 1). \end{aligned} \tag{2}$$

3.2. Cross-Attention Mechanism

In the SD model, the U-Net employs a cross-attention mechanism to denoise the noisy latent image conditioned on text prompts. The cross-attention layer accepts the spatial noisy latent image z_t and the text embeddings y as inputs. The embeddings of the visual and textual features are fused to produce spatial attention maps for each textual token. The cross-attention maps are computed with:

$$A = softmax\left(\frac{QK^T}{\sqrt{d}}\right). \tag{3}$$

The query matrix, denoted as $Q = z_t W_Q^{(i)}$, is the projection of the noisy latent image z_t. The key matrix, represented as $K = y W_K^{(i)}$, is the projected textual features. Here, $W_Q^{(i)}$ and $W_K^{(i)}$ represent the weight matrices of the two linear layers in each cross-attention block i of the U-Net, and d is the output dimension of K and Q features.

4. Methodology

4.1. Subject Text Embedding Augmention

From a generative standpoint, our objective is to create a portrait that accurately represents the identity of the source face. To achieve this, we utilize a pre-trained face recognition model called TFace [18] to extract the identity features. In order to better preserve the identity, we incorporate face features as an important input condition and integrate them into the text to enhance its ability to capture the nuances of identity. To elaborate, we first encode the text prompt $P = \{w_1, w_2, ...w_n\}$ and reference face f into embeddings using the pre-trained text encoder and TFace, denoted as ψ and φ respectively. However, as the features generated by the recognition model are primarily designed for recognition purposes and may not be optimal for generation, we choose to extract only the *shallow features* of the recognition model. Subsequently, we concatenate the embedding of the identity token with the facial feature, and then feed the resulting augmented embeddings into the

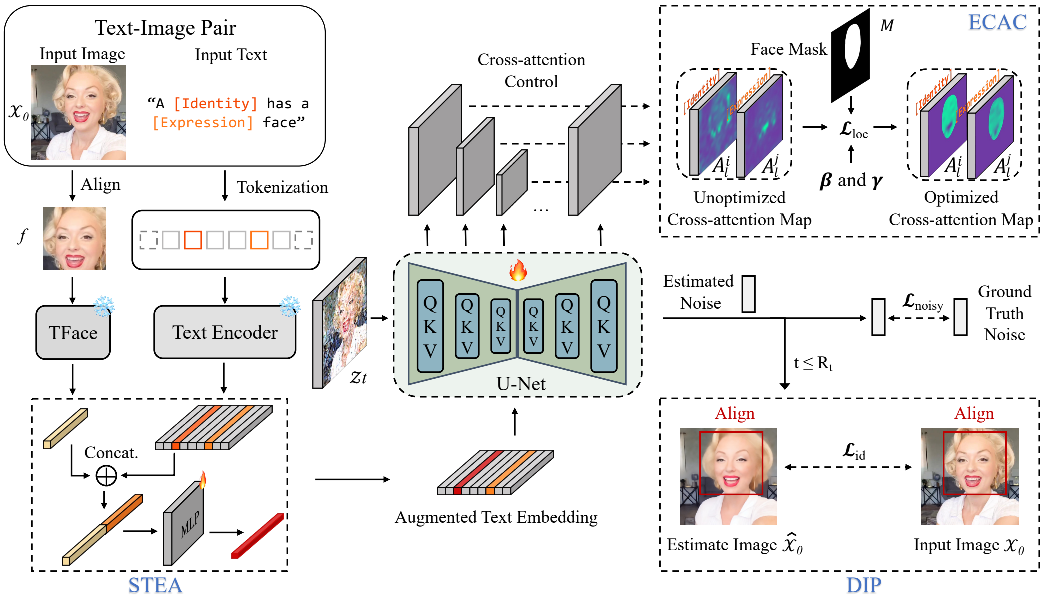


Figure 3. **Overview framework of PortraitBooth**. PortraitBooth extracts the face f from the input image x_0, and augments the subject's features using TFace for improved identity representation. The diffusion model is trained to generate images with enhanced conditioning, incorporating emotion-aware cross-attention for expression editing and dynamic identity preservation to maintain identity. During the testing phase, only the STEA module is required, we just need to input a single image and the corresponding prompt to achieve rapid, robust identity preservation and diverse expression editing capabilities. A_l^i, A_l^j represents the cross-attention map corresponding to the i-th and j-th token at the l-th cross-attention layer, respectively. β and γ represent the maximum values of the cross-attention map for the identity token and expression token respectively, while R_t indicates the timing for identity preservation.

MLP. This process yields the final conditioning embeddings $C = \{c_1, c_2 ... c_n\}$, which are defined as :

$$c_i = \begin{cases} \psi(w_i) & w_i \notin \{identity \quad token\} \\ MLP([\psi(w_i) || \varphi(f)]) & w_i \in \{identity \quad token\}. \end{cases} \tag{4}$$

This approach allows us to generate portraits that not only capture the textual description but also incorporate the identity features extracted from the reference face, resulting in a more accurate representation of the desired identity. Fig. 3 illustrates the STEA module, which provides a concrete example of our augmentation approach.

4.2. Dynamic Identity Preservation

The current SD model achieves image fidelity by relying on accurate prompts, which however poses a significant challenge. When incorporating new image conditions, ensuring the fidelity of the unique reference image becomes necessary. Therefore, it is crucial to incorporate identity loss into the training framework of diffusion models to ensure identity preservation. Let x_0 be the input image, z be its latent space representation, T ($T < 1000$) represents the total number of noise injection steps. For a small value of R_t ($R_t < T$), we can get estimated $\hat{z}_0$ directly from z_t and the predicted noise $\epsilon_\theta(z_t, t, C)$. From Eqn. 2, the one-step reverse formula is defined as :

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}, t \leq R_t, \tag{5}$$

After reverse, the estimated $\hat{z}_0$ is decoded from the latent space using the VAE decoder $\mathcal{D}$ to obtain the estimated input image $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. Then, based on the facial region of the original image, the estimated facial region image $\hat{x}_0^f$ is extracted from the reconstructed image. Finally, the identity loss between the estimated facial image and the reference facial image is defined as:

$$\mathcal{L}_{id} = \begin{cases} 1 - CosSim\left(\varphi(f), \varphi(\hat{x}_0^f)\right) & t \leq R_t \\ 0 & t > R_t. \end{cases} \tag{6}$$

The identity loss is designed to handle noisy images and improve the model's ability to preserve the identity. The DIP module, as illustrated in Fig. 3.

4.3. Emotion-aware Cross-attention Control

For previous one-shot personalized generation works [24, 39, 42] , a common issue is that the generated images always have the same expression as the reference image, regardless of the prompt given. Although we have largely decoupled identity and attributes by utilizing pre-trained facial recognition models to extract discriminative features for subject feature enhancement, the complexity and diversity of facial expressions still pose a challenge in maintaining identity during portrait generation. This issue primarily arises because the cross-attention map is spread across the entire image during image generation. To address this issue and ensure that the cross-attention map corresponding

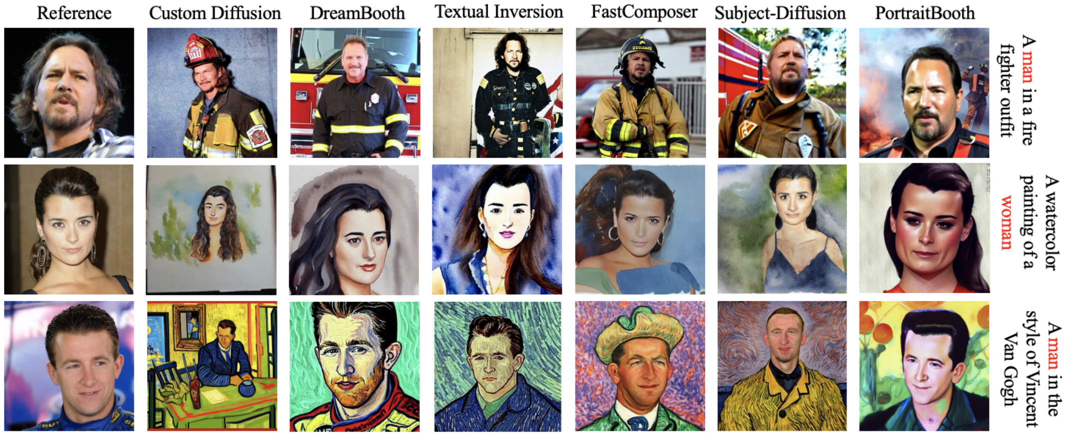
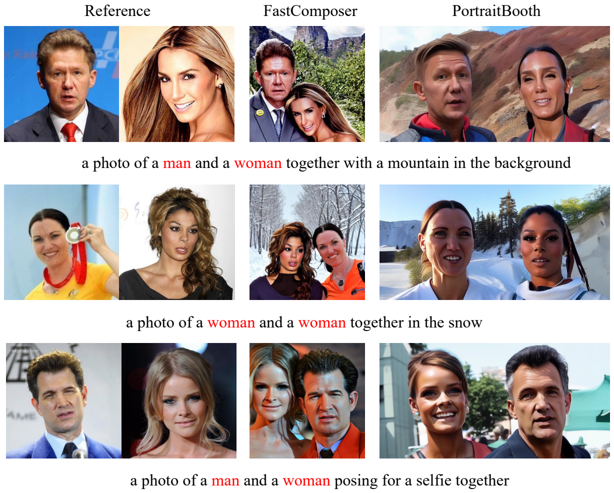

Reference
Custom Diffusion
DreamBooth
Textual Inversion
FastComposer
Subject-Diffusion
PortraitBooth
A man in a fire fighter outfit
A watercolor painting of a woman
A man in the style of Vincent Van Gogh

Figure 4. **Comparison of different methods on single subject image generation** in the testing dataset.



a photo of a man and a woman together with a mountain in the background

a photo of a woman and a woman together in the snow

a photo of a man and a woman posing for a selfie together

Figure 5. **Comparison of different methods on multi-subject image generation** in the testing dataset.

to specific tokens only attends to the image region occupied by the corresponding concept, we propose an emotion-aware cross-attention control mechanism.

Specifically, unlike previous methods [24, 42] that used attention masks to control subject token's attention map solely on the one subject region, we allow attention control of different tokens within the same region by truncating cross-attention mechanism. For instance, when dealing with tokens for facial expressions and identity, we employ a face mask to ensure that the attention maps corresponding to these two tokens are both focused on the face region. However, we observe that when two different tokens' attention maps are both constrained to the same region, one token may learn well while the other may not. To tackle this problem, we propose a complete local control constraint with truncating cross-attention mechanism:

$$\begin{aligned}\mathcal{L}_{loc} = & \frac{1}{N}\sum_{l=1}^{N}\lambda(mean(A_l^i(1-M)) + mean(relu(\beta - A_l^i)M)) \\ & + \frac{1}{N}\sum_{l=1}^{N}\mu(mean(A_l^j(1-M)) + mean(relu(\gamma - A_l^j)M)),\end{aligned} \tag{7}$$

where M is the face mask normalized to [0,1]. $mean$ is the pixel-level averaging. $A_l^i, A_l^j \in [0,1]$ represents the cross-attention map corresponding to the identity and expression token at the l-th cross-attention layer. β and γ are used to constrain the maximum response intensity of the cross-attention map in the facial area corresponding to the identity token and expression token, respectively. We optimize $\mathcal{L}_{loc}$ to ensure that objects' attention map exhibit their respective response in the desired area, which is achieved by maximizing the response of each token's attention map to the face region and minimizing its response to the background, along with the use of a truncated response mechanism in the attention map. λ and μ are localization loss ratios, which are 0.001 and 0.01. Fig. 3 illustrates the ECAC module.

4.4. Objective Function

First, we use TFace φ to extract the face embedding and concatenate it with specific identity token embedding, which is extracted from the text encoder ψ. These are then fed into an MLP for feature enhancement, forming U-Net aware conditional information C. Next, we feed the noisy latent space feature map z_t into a U-Net with conditional guidance to predict noise, while implementing a truncation mechanism for local attention control for specific tokens. To better preserve identity, we employ dynamic identity preservation method to calculate the loss between the estimated face image $\hat{x}_0^f$ and reference face f. The final training objective of PortraitBooth is:

$$\mathcal{L}_{total} = \mathcal{L}_{loc} + \mathcal{L}_{noise} + \mathcal{L}_{id}. \tag{8}$$

5. Experiments

5.1. Experimental Setups

Dataset Description. We constructed a single subject image-text paired dataset based on the CelebV-T dataset [48], which consists of $70,000$ videos. To utilize the additional textual descriptions provided by CelebV-T, we randomly extracted the first or last frames of each video.

Method	**Type**	**Reference Image** ↓	**Id Pres.** ↑	**CLIP-TI** ↑	**Test Time** ↓	**Training Cost**
Stable Diffusion [31]	Zero Shot	0	0.039	0.268	≈2s	-
Face0 [39]	One Shot	1	-	-	≈2s	64 TPU
Textual-Inversion [10]	Finetune	5	0.293	0.219	≈2500s	1 A100
DreamBooth [33]	Finetune	5	0.273	0.239	≈1084s	1 A100
Custom Diffusion [22]	Finetune	5	0.434	0.233	≈789s	1 A100
FastComposer [42]	One Shot	**1**	0.514	0.243	≈**2s**	8 A6000
Subject-Diffusion [24]	One Shot	**1**	0.605	0.228	≈**2s**	24 A100
PortraitBooth (ours)	One Shot	**1**	**0.657**	**0.245**	≈**2s**	3 A100

Table 2. **Comparison between our method and baseline approaches on single-subject image generation.** Our approach achieves highly satisfactory results with the utilization of relatively limited resources under the one-shot setting.

Method	**Type**	**Reference Image** ↓	**Id Pres.** ↑	**CLIP-TI** ↑	**Test Time** ↓	**Training Cost**
Stable Diffusion [31]	Zero Shot	0	0.019	0.284	≈2s	-
Textual-Inversion [10]	Finetune	5	0.135	0.211	≈4998s	1 A100
Custom Diffusion [22]	Finetune	5	0.054	**0.258**	≈789s	1 A100
FastComposer [42]	One Shot	**2**	0.431	0.243	≈**2s**	8 A6000
PortraitBooth (ours)	One Shot	**2**	**0.647**	0.239	≈18s	3 A100

Table 3. **The comparison between our method and the baseline approaches that support multiple-subject image generation.** StableDiffusion was used as the text-only baseline without any subject conditioning.

Additionally, we used the Recognize Anything model [53] to generate captions describing the main subject for all images. To enhance the robustness of our models, we randomly selected a frame from the middle section of each video and used the facial region as our reference face image. We employ the pre-train face parsing model [23] to generate subject face segmentation masks for each image.

Training Details. We start training from the Stable Diffusion $v1-5$ [31] model. To encode the identity inputs, we use TFace model. During training, we only train the U-Net, the MLP module. We train our models for 150k steps on 6 NVIDIA V100 GPUs (For the sake of easy and intuitive comparison later, we roughly convert 6 NVIDIA V100 GPUs into 3 NVIDIA A100 GPUs.), with a constant learning rate of $1e-5$ and a batch size of 2. We train the model solely on text conditioning with 10% of the samples to maintain the model's capability for text-only generation. To facilitate classifier-free guidance sampling [15], we train the model without any conditions on 10% of the instances. During training, we apply the loss only in the subject's face region to half of the training samples to enhance generation quality in the subject area. There are 11 emotion words involved in truncating cross-attention control, such as happy, angry, sad, *etc*. We select a value of 250 for R_t to obtain $\hat{z_0}$ through reverse. The selected identity label is from the categories {"man","woman"}. During inference, We use Euler [19] sampling with 50 steps and a classifier-free guidance scale of 5 across all methods.

Evaluation Metric. We evaluate the quality of image generation based on identity preservation (Id Pres.) and CLIP text-image consistency (CLIP-TI). Identity preservation is determined by detecting faces in the reference and generated images using MTCNN [51], and then calculating pairwise identity similarity using FaceNet [36]. For multi-subject evaluation, we identify all faces within the generated images and use a greedy matching procedure between the generated faces and reference subjects. For the evaluation of expression editing, we calculate the text-image consistency between the emotion words in each prompt and the corresponding generated images as our expression coefficient metric. For efficiency evaluation, we consider the total time for customization, including fine-tuning (for tuning-based methods) and inference. We also take into consideration the total number of GPUs required throughout the entire procedure. All baselines, by default, are run with the standard set of hyperparameters as mentioned in their paper.

5.2. Personalized Image Generation

To evaluate our model's effectiveness in this area, we use the single-entity evaluation method employed in FastComposer [42] and compare our model's performance to that of other existing methods including DreamBooth [33], Textual-Inversion [10], Custom Diffusion [22], and Subject-Diffusion [24]. Methods [10, 22, 33] were used the implementation from diffusers library [40]. Considering that Face0 [39] does not provide open-source code, we can only list the hardware resources mentioned in their paper as a point of comparison. Stable Diffusion [31] was used as the

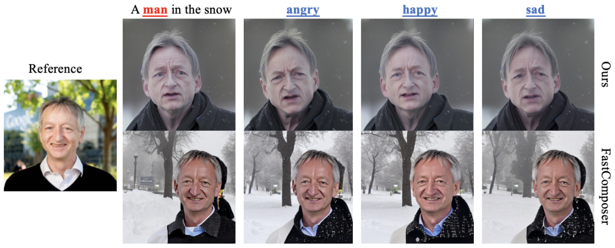


Figure 6. **Comparison chart of expression editing between our method and FastComposer,** focusing on the three most distinct expression terms.

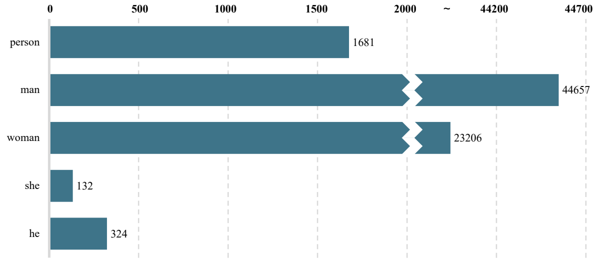


Figure 7. **The number of main subject words** occurrences in the generated 70, 000 captions.

text-only baseline. The entire test set comprises 15 subjects, and 30 texts. The evaluation benchmark developed a broad range of text prompts encapsulating a wide spectrum of scenarios, such as recontextualization, stylization, accessorization, and diverse actions. Five images were utilized per subject to fine-tune the optimization-based methods. For the one-shot method, a single randomly selected image was employed for each subject. As shown in Tab. 2, PortraitBooth significantly outperforms all baseline approaches in identity preservation. Fig. 4 shows the qualitative results of single-subject personalization comparisons, employing different approaches across an array of prompts.

5.3. Multi-Subject Image Generation

We then delve into a more intricate scenario: multi-subject, subject-driven image generation. We scrutinize the quality of multi-subject generation by utilizing all possible combinations (a total of 105 pairs) formed from the 15 subjects described in Section §5.2, allocating 21 prompts to each pair for evaluation. Considering that PortraitBooth was trained on a single-subject dataset, we incorporated the MultiDiffusion [5] generation method, which combines multiple reference diffusion generation processes with shared parameters, to generate images in different regions during inference. Tab. 3 shows a quantitative analysis contrasting PortraitBooth with the baseline methods. The results demonstrate that PortraitBooth significantly improves the identity preservation score. Moreover, our prompt consistency is comparable to tuning-based approaches [10, 22], but weaker than FastComposer and Custom Diffusion. We attribute this vulnerability may stem from our method's inclination to give precedence to subject fidelity. The longer test time, compared to FastComposer, is a result of current multi-subject generation method limitations. We anticipate a significant reduction in our multi-subject generation time as these methods evolve. Fig. 5 shows the qualitative results of multi-subject personalization comparisons.

Method	Type	Expression Coefficients ↑
Textual-Inversion [10]	FineTune	0.158
Custom Diffusion [22]	FineTune	0.182
DreamBooth [33]	FineTune	0.153
FastComposer [42]	One Shot	0.133
PortraitBooth w/o expression control	One Shot	0.177
PortraitBooth (Ours)	One Shot	**0.193**

Table 4. **Comparison of facial expression coefficients** between PortraitBooth and other methods.

Combination Type	Id Pres. ↑	CLIP-TI ↑
{*"person"*}	0.623	0.229
{*"he"*, *"she"*}	0.606	0.208
{*"man"*, *"woman"*}	0.657	0.245

Table 5. **The impact of embedding enhancement of subject tokens** from different categories.

Item	Method	Id Pres. ↑	CLIP-TI ↑
	PortraitBooth	0.657	0.245
(a)	w/o STEA	0.563	0.244
(b)	w/o DIP	0.638	0.239
(c)	w/o ECAC	0.632	0.235

Table 6. **Ablation results of three components.**

5.4. Expression Editing

To demonstrate the effectiveness of our approach in terms of facial expression editing, we conduct a series of comparisons against both test-time fine-tuning methods capable of expression editing and those that are not. The entire test set comprises 15 subjects, as mentioned in Section §5.2, with each subject assigned 11 prompts containing emotion-related words. The comprehensive results in Tab. 4 clearly show that our method significantly outperforms the others. Fig. 6 presents the experimental comparison results for expression editing, showcasing the versatility of our method.

5.5. Ablation Study

Impact of Identity Token. After creating prompts for 70, 000 training images, we analyzed the subject identity token for each image. The results, shown in Fig. 7, revealed three categories of subject words: {"man","woman"}, {"person"}, and {"he","she"}. We tested each category's effectiveness after feature enhancement by converting other identity tokens in each prompt to each experiment token. When converting the "person" token, we manually classified gender correctly for alignment. Our findings, presented in Tab. 5, showed that the {"man", "woman"} category, being more specific, improved subject fidelity and text-image

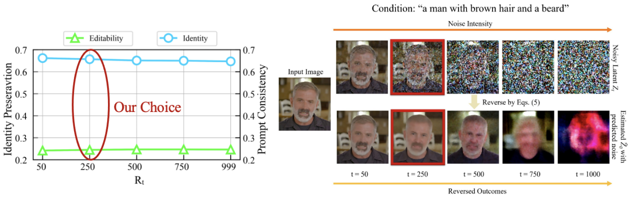


Figure 8. **Effects of using different upper limit of timesteps** for one-step reverse (left), **visualization of noise addition** at different timesteps t and denoising (right).

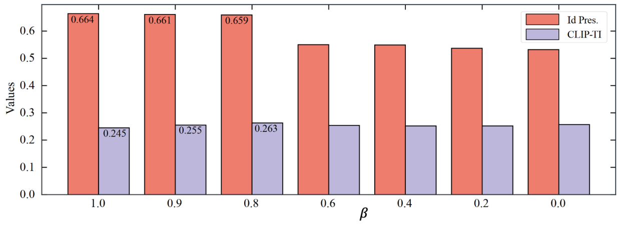


Figure 9. **The impact of truncating cross-attention** only with different values of β.

consistency. The category of {“he”, “she”}, {“person”} was less descriptive and consistent.

Impact of STEA. To investigate the influence of target features obtained from a pre-trained face recognition model, we conducted an ablation. When removing the STEA module, we employed CLIP-image-encoder for training and extracting target features to enhance subject text embeddings. The experimental results, as depicted in Tab. 6(a), clearly indicate that utilizing a face feature extractor trained on a large-scale dataset is significantly more effective compared to training the image encoder.

Impact of DIP. Tab. 6(b) presents the ablation studies on our proposed DIP. As the results show, the DIP module has proven beneficial for identity preservation.

Impact of ECAC. To let the model focus on semantically relevant subject regions within the cross-attention module, we incorporate the attention map control. Tab. 6(c) indicates this operation delivers a substantial performance improvement for identity preservation as well as prompt consistency. Besides, as shown in Tab. 4, even when our cross-attention control mechanism does not constrain the expression terms, we still achieve satisfactory results in facial expression editing. This further demonstrates the effectiveness of our method in decoupling identity and attributes.

Hyperparameter R_t. As shown on the left side in Fig. 8, when R_t grows, the model trades off identity preservation for improved editability. We select 250 as the optimal R_t value, as it provides a good balance. The right side of the figure illustrates the visual results.

Hyperparameter β and γ. We studied the balance between identity preservation and editability by solely adjusting the β in the truncation process, keeping γ at 0, to minimize their impact. As shown in the Fig. 9, when β is in the range of [0.8, 1], the difference in identity preservation is not significant, but there is a noticeable change in editability. However, when β is less than 0.8, there is a sudden jump in identity preservation. We believe this is because the enhanced face embeddings have a significant effect only on the facial region. Tab. 7 confirms our hypothesis. Therefore, we chose β as 0.8 as our hyperparameter. According to Eqn. 3, the sum of cross-attention values for all tokens in each region is 1, meaning for the facial region, $\beta + \gamma <= 1$, so we conducted experiments with γ values of 0.1 and 0.2. In the Tab. 8, we found that while the difference in identity preservation is not significant between the two values, there is a substantial difference in editability. This is because facial responses include not only expressions but also features like facial hair and accessories, *etc*. Hence, we select γ as 0.1 as our hyperparameter.

Mask Type	Id Pres. ↑	CLIP-TI ↑
Face Mask	0.657	0.245
Person Mask	0.623	0.229

Table 7. **Impact of different types of masks.** “Face Mask” refers to the segmentation of only the facial area, while “Person Mask” refers to the segmentation of the entire person’s body.

γ and β combination	Id Pres. ↑	CLIP-TI ↑
$\beta = 0.8, \gamma = 0.1$	0.657	0.245
$\beta = 0.8, \gamma = 0.2$	0.652	0.223

Table 8. **The influence of different combinations of β and γ.**

6. Conclusion

In the portrait personalization field, we face the core challenge of proposing an efficient, low training cost, and high identity preserving portrait personalization framework. In this paper, we propose PortraitBooth, an efficient one-shot text-to-portrait generation framework, that leverages Subject Text Embedding Augmentation and Dynamic Identity Preservation to achieve robust identity preservation, using Emotion-aware Cross-Attention Control to achieve expression editing, respectively. Experimental results demonstrate the superiority of PortraitBooth over the state-of-the-art methods, both quantitatively and qualitatively.

7. Acknowledgments

This work was supported by National Science and Technology Major Project of China (No. 2022ZD0118201) and the National Natural Science Foundation of China (No. 62072386), supported by Henan Province key research and development project (No.231111212000), Henan Center for Outstanding Overseas Scientists (No. GZS2022011) and the open foundation of Henan Key Laboratory of General Aviation Technology (No. ZHKF-230212). We also thank Digital Culture Lab of Tencent SSV for their invaluable contribution during our research.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 3

[4] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 1

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 7

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 3

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[9] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 6, 7

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 3

[12] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 2

[13] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[18] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 3

[19] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 6

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[21] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3, 6, 7

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 6

[24] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2, 3, 4, 5, 6

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[26] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 2

[27] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 10619–10629, 2022. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

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

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 6

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 6, 7

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2, 3

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[37] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 2

[38] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[39] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *arXiv preprint arXiv:2306.06638*, 2023. 2, 3, 4, 6

[40] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 6

[41] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1

[42] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 3, 4, 5, 6, 7

[43] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *European Conference on Computer Vision*, pages 54–71. Springer, 2022. 2

[44] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7632–7641, 2022. 2

[45] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6619, 2023. 2

[46] Zhiliang Xu, Hang Zhou, Zhibin Hong, Ziwei Liu, Jiaming Liu, Zhizhi Guo, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Styleswap: Style-based generator empowers robust face swapping. In *European Conference on Computer Vision*, pages 661–677. Springer, 2022. 2

[47] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 1

[48] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 5

[49] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5326–5335, 2020. 2

[50] Jiangning Zhang, Xianfang Zeng, Chao Xu, and Yong Liu. Real-time audio-guided multi-face reenactment. *IEEE Signal Processing Letters*, 29:1–5, 2021. 2

[51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 6

[52] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2

[53] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 6