Towards a Simultaneous and Granular Identity-Expression Control in Personalized Face Generation

Renshuai Liu[1,2], Bowen Ma[2], Wei Zhang[2],
Zhipeng Hu[2], Changjie Fan[2], Tangjie Lv[2], Yu Ding[2]*, Xuan Cheng[1]*
[1]School of Informatics, Xiamen University
[2]Virtual Human Group, Netease Fuxi AI Lab

Figure 1. Our framework takes three inputs: a prompt describing the background, a selfie photo uploaded by the user, and a text related to the fine-grained expression labels. The generated faces well match the inputted triples and exhibit fine-grained expression synthesis.

Abstract

In human-centric content generation, the pre-trained text-to-image models struggle to produce user-wanted portrait images, which retain the identity of individuals while exhibiting diverse expressions. This paper introduces our efforts towards personalized face generation. To this end, we propose a novel multi-modal face generation framework, capable of simultaneous identity-expression control and more fine-grained expression synthesis. Our expression control is so sophisticated that it can be specialized by the fine-grained emotional vocabulary. We devise a novel diffusion model that can undertake the task of simultaneously face swapping and reenactment. Due to the entanglement of identity and expression, separately and precisely controlling them within one framework is a nontrivial task, thus has not been explored yet. To overcome this, we propose several innovative designs in the conditional diffusion model, including balancing identity and expression encoder, improved midpoint sampling, and explicitly background conditioning. Extensive experiments have demonstrated the controllability and scalability of the proposed framework, in comparison with state-of-the-art text-to-image, face swapping, and face reenactment methods.

1. Introduction

The research community has been striving to improve controllability in the generation of face images tailored to user preferences. A common practice in controllable generation and manipulation is to use different modalities as conditioning in a face generator model, such as texts [23, 26, 29, 43, 46], reference images [5, 22, 27, 45], segmentation masks [12, 21, 34] and audios [32, 54, 55].

Although these methods have realized the ability to control the local features and global attributes in a face, the simultaneous control of identity and expression in a specific background has not been fully explored, which involves three important high-level attributes (i.e. identity, expression, and background) to determine a face image. Since identity and expression are highly entangled, it's challeng-

Work done when Renshuai Liu was an intern at Netease Fuxi AI Lab
* Co-corresponding Authors.

ing to separately and precisely control them in a unified framework. Additionally, in existing generation or manipulation methods, the granularity of expression control remains at a coarse level, often limited to the commonly used seven or eight labels, e.g. "surprise", "happiness", "anger" etc. These labels struggle to cover the entire emotional space sufficiently in the open world.

To tackle these issues, this paper proposes a novel framework that can *simultaneously control identity, expression, and background from multi-modal inputs*. As shown in Fig. 1 and 2, the inputs contain three items: 1) a text that describes the scene, 2) a selfie photo uploaded by the user to provide identity, 3) a text related to the expression labels. Human language can conceptually describe expressions and accurately describe scenes but can't describe identity precisely, while images can be naturally used in identity recognition. On the output side, the generated face will have the same identity as the input selfie photo, show the expression specified by the text, and be placed in the background described in the text [1, 9], as shown in Fig. 1. To support *fine-grained expression description*, we employ an expression dictionary of 135 English words [4], e.g. "amazement", "exhilaration", "hysteria" etc., which can more comprehensively describe the emotion domain.

The technical core inside the proposed framework is a novel diffusion model that can conduct *Simultaneous Face Swapping and Reenactment (SFSR)*. Swapping and reenactment, which transfer the identity or expression of the source face to the target face, are two classical face manipulation tasks and have been studied extensively. Meanwhile, SFSR is a relatively new and unexplored task, which aims at separately transferring the identity from the source face, and the expression from another source to one target face, while keeping the background attributes (e.g. face pose, hair, glasses, and surroundings) in the target unchanged. To prepare the two sources and one target for SFSR, the text that describes the scene will be input to a pre-trained text-to-image model (Stable Diffusion) [35] to get the background image, while the text that describes the expression will be used as the search key in the 135-class emotion dataset [4, 17] to retrieve the expression image. Together with the input identity image, the three images will be used as conditioning in a latent diffusion model [35] to generate the result, which has already exhibited high customizability of various conditions on image generation [51].

Based on the foundations of the diffusion model [10, 35], we also propose several elaborate designs in SFSR diffusion model. 1) *Balancing identity and expression encoder*. We develop the identity and expression encoder, which are competitive with each other, to reduce the transfer of residual identity attribute in the expression encoder to the final result. 2) *Improved midpoint sampling*. To achieve both efficiency and accuracy in imposing the identity and expression constraints during training, we propose the improved midpoint sampling, which can generate the face of higher quality with only two times of prediction than the original midpoint sampling [53]. 3) *Explicitly background conditioning*. We provide background condition in the training phase so that the diffusion model can focus on the generation of face but not background, and get more hints from inputs to recover face pose and lighting. This design is different from previous methods [2, 53] that use background image only in inference, and proves to be more effective. We name the newly designed diffusion model for SFSR task as *DiffSFSR*. Finally, the contributions of this paper are summarized as:

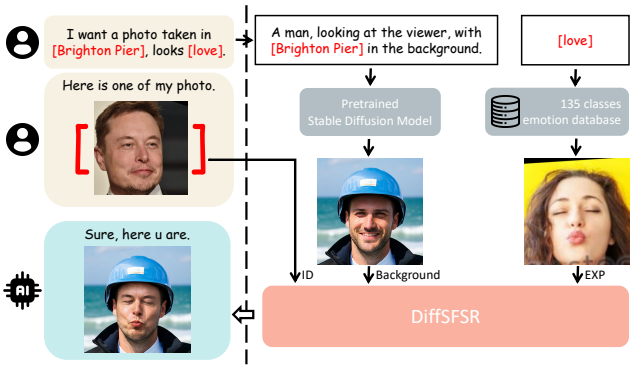

Figure 2. Overview of the proposed face generation framework.

- A novel face generation framework that achieves simultaneous control of identity and expression, and more fine-grained expression synthesis than state-of-the-art text-to-image methods.
- A novel face manipulation task, simultaneously face swapping and reenactment, which has never been explored by previous methods. This task is also compatible with the traditional separate swapping and reenactment tasks by re-combination of inputs.
- Three innovative designs in the conditional diffusion model, including balancing identity and expression encoder, improved midpoint sampling, and explicitly background conditioning, which increase the controllability and image quality.

2. Related Works

Conditional Face Generation. Early methods usually use a single modality as conditioning. For example, there has been a surge of text-to-face researches that utilize the pre-trained StyleGAN [14–16] and the text encoder, such as TediGAN [46], StyleCLIP [29] and StyleT2I [23]. Using images as conditioning [5, 22, 27, 45] is also popular in the research community. This kind of methods usually generate the face that shares the same identity or expression with the input face image. The most recent methods begin to use multiple modalities, due to the fact that different modalities

are complementary to each other. For example, the tuple of texts and segmentation masks [8, 11] is very popular to control face generation. Our proposed face generation framework also takes as input multiple modalities, namely text and image.

Our work is closely related to face swapping and face reenactment. The mainstream way to improve visual quality is using GANs [3, 5, 6, 18, 19, 22, 33, 39, 40, 45, 48, 56], which injects the identity or expression features extracted from the source into the generation network, and uses multiple losses to ensure semantic consistency and image fidelity. The most recent method [53] employs a diffusion model, and reformulates the face swapping as a conditional inpainting task. There exist methods [27, 28, 30, 47] that combine the two tasks, namely swapping and reenactment, in a single framework. In their pipeline, a switch operator is usually placed in the facial features transfer stage to switch between swapping and reenactment tasks. The main difference in functionality between these methods and our DiffSFSR is that either identity or expression, but not both of them, is transferred to the result.

Preliminary on Diffusion Models. The diffusion model (DDPM) [10] has been well documented. It contains diffusion and denoising processes. Given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the diffusion process produces a series of intermediate noisy samples $\{\mathbf{x}_t\}$ by continuously adding Gaussian noise $\mathcal{N}$ with variance $\beta_t \in (0,1)$ at timestep t: $q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$ where $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$. x_t can be sampled directly from x_0, without generating intermediate steps:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. When a long increasing sequence $\beta_{1:T}$ is set such that $\bar{\alpha} \approx 0$, the distribution of $\mathbf{x}_T$ will converge to a standard Gaussian.

The denoising process starts from a Gaussian noise sample $\mathbf{x}_T \sim \mathcal{N}(0, I)$, and denoises $\mathbf{x}_T$ to $\mathbf{x}_0$ by sequentially sampling the posteriors $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Based on Bayesian rules, $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ can be derived to:

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}), \\ \mathbf{x}_{t-1} &= \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) + \sqrt{\tilde{\beta}_t}\epsilon, \\ \text{where } \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \\ \text{and } \tilde{\beta}_t &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \end{aligned} \tag{2}$$

Since $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ has no closed-form, a deep neutral network p_θ is trained to approximate it.

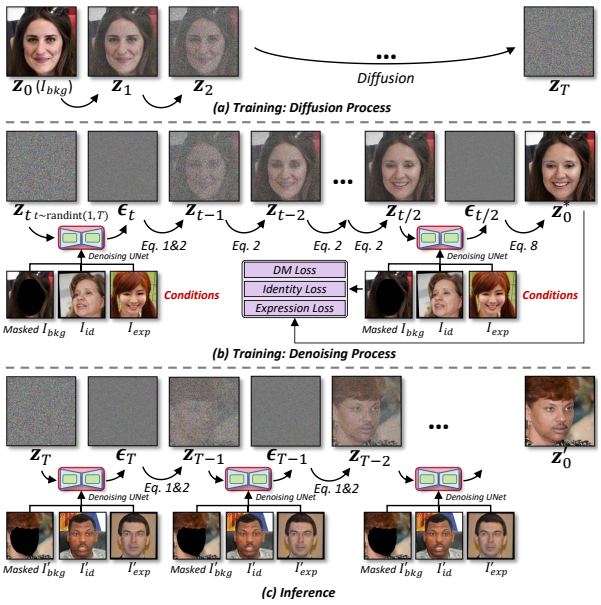

Figure 3. Pipeline of DiffSFSR, including training and inference phases. Although the diffusion model is practically trained and tested in the latent space [35], we illustrate all the processes in the original image space for visualization. The transformations between $\mathbf{x}$ and $\mathbf{z}$ are not illustrated for brevity.

3. The Proposed Framework

As shown in Fig. 2, the proposed face generation framework contains two main modules: firstly converting the multi-modal inputs into three images, and then generating the face image by SFSR diffusion model from the inputted three images. In the converting module, the inputs contain an identity image I_{id}, a text prompt P_{bkg} describing the scene and a text prompt P_{exp} related to the expression label. P_{bkg} is injected to a pre-trained text-to-image diffusion model [35] to obtain the background image I_{bkg}. P_{exp} is used as the search key in the emotion dataset [4], which contains 728,946 face images of 135 emotion categories. According to P_{exp}, an expression image I_{exp} is randomly retrieved from the corresponding category. In the DiffSFSR, I_{bkg}, I_{id} and I_{exp} are used as conditioning to generate the final result I_{out}.

The pipeline of the DiffSFSR is shown in Fig. 3, including the training and inference phases. The latent diffusion model [35] is chosen as the backbone due to its high customizability for various conditions. Similar to DDPM [10], the training of the latent diffusion model also consists of the diffusion process and the denoising process. One sample in the training data is the triplet $[I_{bkg}, I_{id}, I_{exp}]$, excluding the ground truth counterpart of I_{out}. The input I_{bkg} is firstly embedded to a latent $\mathbf{z}_0$, and then be added with the Gaussian noise in the diffusion process. The denoising process denoises the latent $\mathbf{z}_t$ at random t-th timestep to $\mathbf{z}_0^*$, by applying the UNet conditioned on the masked I_{bkg} and the

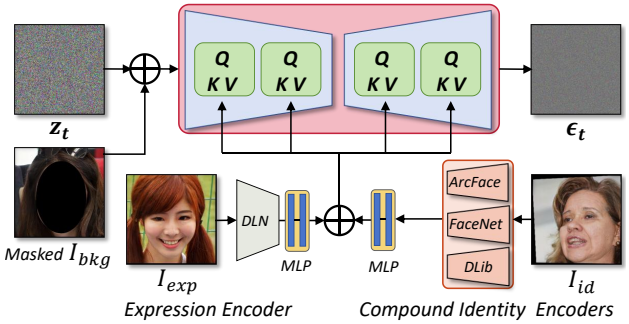

Figure 4. The network architecture of the denoising UNet. QKV denotes the cross-attention layer.

extracted features from I_{id}, I_{exp}. The denoised latent $\mathbf{z}_0^*$ further needs to be transformed back to the image space to generate $\mathbf{x}_0^*$. To better compute the identity loss between $\mathbf{x}_0^*$ and I_{id}, and the expression loss between $\mathbf{x}_0^*$ and I_{exp}, we adopt the improved midpoint sampling, considering both accuracy and efficiency. In the inference phase, the trained denoising UNet is used multiple times to gradually generate $\mathbf{z}_0'$ from a Gaussian noise $\mathbf{z}_T$, which is finally decoded to image $\mathbf{x}_0'$.

3.1. Conditions

The key in the DiffSFSR is how to learn the disentangled representations respectively for I_{bkg}, I_{id}, I_{exp} and then condition the diffusion model on the learned embeddings simultaneously, thus enabling disentangled and precise control of the targeted face. The network architecture of the denoising UNet, together with the three conditioning embeddings, is shown in Fig. 4.

Background. Except for identity and expression, the input I_{bkg} provides all other attributes for the output, e.g. face pose, hair, glasses, lighting and surroundings. We mask the facial region in I_{bkg} with a face parsing method, turning the task from face editing to face inpainting. Then, the masked I_{bkg} is concatenated with the latent $\mathbf{z}_i$ as conditioning for the diffusion model, in both training and inference phases, so that most parts of the background in I_{out} are exactly the same with I_{bkg}.

Our way of preserving background attributes is totally different from the recently proposed diffusion model based face-swapping method, DiffSwap [53]. In DiffSwap, the background pixels (masked I_{bkg}) are not explicitly provided during training, but only during the inference phase, which requires the diffusion model to reconstruct the background pixels in the training phase. To ensure global consistency, the reconstruction loss is computed on the whole image. Since the facial region usually occupies less than 50% of the total image, a significant portion of the network optimization is initially dedicated to the reconstruction of background pixels before the network starts to perform fine-grained generation in the facial region. Hence, the faces generated by DiffSwap may suffer from image blur and low quality, compared with our results.

Another advantage of explicitly providing background pixels in the training is that the diffusion model is forced to learn to estimate the face pose and lighting from the background pixels, as there exists a strong correlation between face pose, lighting, and the background. To summarize, we provide the masked I_{bkg} as conditioning in the hope that, it can make the diffusion model focus on the generation of face but not the generation of background, and provide more hints for recovering important attributes. We are the first to disentangle background attributes from identity and expression and explicitly make them as conditioning, in the task of face manipulation.

Expression. Previous methods prefer to use 2D facial landmarks as the expression representation. Human expression contains complicated and subtle facial movements and is closely related to facial texture, such as facial wrinkles and facial action unit activation, thus the 2D facial landmarks are not enough to represent the accurate expression attribute. In order to get a powerful expression representation, we adopt the identity-disentangled and fine-grained expression representation network named DLN [52] as the encoder. A two-layer MLP is then used for domain transformation and feature shape alignment. After that, the expression embedding is injected into the diffusion model through the cross-attention module.

Identity. The identity encoder should be competitive with the expression encoder [52] for balanced conditioning, otherwise, the residual identity attribute in the expression encoder will be accidentally transferred to the result. As the expression is closely entangled with identity, designing a completely identity-ignored expression encoder is still an open problem in the field of face analysis. In the aspect of identity encoder, we apply an identity compound embedding, since a single identity embedding is usually biased [50] and insufficient to balance the expression embedding. Three state-of-the-art face recognition models [7, 20, 38] are selected to construct the identity compound embedding. A single embedding can't compensate for the impact of the residual identity in the expression encoder, while the compound embedding can meet the requirement. Similar to the expression encoder, we use a two-layer MLP to map the identity embedding of different shapes to a uniform dimension and condition them through cross-attention.

3.2. Training Objective

The training objective of the diffusion model can be formulated as the Mean Squared Error (MSE) loss:

$$L_{DM} = \mathbb{E}_{\mathbf{z}_t, \mathbf{C}, \epsilon, t}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{C}, t)||_2^2], \tag{3}$$

where $\mathbf{z}_t$ denotes the noisy latent obtained by adding noise ϵ to $\mathbf{z}_0$ at t-th timestep, ϵ_θ denotes the denoising UNet learned

to predict ϵ, and $\mathbf{C}$ denotes the conditions. The conditions are defined as:

$$\mathbf{C} = [M \odot I_{bkg}, \mathcal{E}_{id}(I_{id}), \mathcal{E}_{exp}(I_{exp})], \tag{4}$$

where M denotes the binary mask of the facial region, $\mathcal{E}_{id}(\cdot)$ denotes the identity encoder, and $\mathcal{E}_{exp}(\cdot)$ denotes the expression encoder.

We also use identity loss and expression loss, following the common practice. The losses are defined as:

$$L_{id} = 1 - \text{CosSim}(\mathcal{E}_{id}(I_{id}), \mathcal{E}_{id}(\mathcal{D}_{DM}(\mathbf{z}_0^*))), \tag{5}$$

$$L_{exp} = \text{MSE}(\mathcal{E}_{exp}(I_{exp}), \mathcal{E}_{exp}(\mathcal{D}_{DM}(\mathbf{z}_0^*))), \tag{6}$$

where $\text{CosSim}(\cdot,\cdot)$ denotes the cosine similarity function and $\text{MSE}(\cdot,\cdot)$ denotes the MSE function. The final loss function can be formulated as:

$$L = L_{DM} + \lambda_1 L_{id} + \lambda_2 L_{exp}, \tag{7}$$

where λ_1, λ_2 denote the hyper-parameters.

3.3. Improved Midpoint Sampling

To compute the identity and expression losses at t-th timestep during training, the denoised latent $\mathbf{z}_0^*$ firstly needs to be generated from the noisy latent $\mathbf{z}_t$. In the original diffusion model, e.g. DDPM [10], the generation of $\mathbf{z}_0^*$ requires multiple times sampling on different timesteps, which is unacceptable in the training phase of high-fidelity image generation task. To tackle this issue, DiffSwap [53] proposes a midpoint sampling method, which can get a coarse $\mathbf{z}_0^*$ with only two steps of sampling. Specifically, in timestep t it firstly estimates $\mathbf{z}_{t_1}$, $t_1 = \lfloor \frac{t}{2} \rfloor$ by using the formula:

$$\mathbf{z}_{t_1} = \frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t_1}}\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})}{\sqrt{\bar{\alpha}_t / \bar{\alpha}_{t_1}}}. \tag{8}$$

Then, starting from the estimated $\mathbf{z}_{t_1}$, it predicts the final $\mathbf{z}_0^*$ by using the formula:

$$\mathbf{z}_0^* = \frac{\mathbf{z}_{t_1} - \sqrt{1 - \bar{\alpha}_{t_1}}\epsilon_\theta(\mathbf{z}_{t_1}, t_1, \mathbf{C})}{\sqrt{\bar{\alpha}_{t_1}}} \tag{9}$$

It seems to be an appealing solution to compute the identity and expression losses with only two steps of sampling. We refer readers to the literature [53] for the detailed derivation of the two formulas in Eq. 8 and Eq. 9.

However, we find that there is an issue in Eq. 8. The estimated noise here should be the noise that can convert $\mathbf{z}_{t_1}$ into $\mathbf{z}_t$, but the noise estimated by $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$ is actually the noise that convert $\mathbf{z}_0$ to get $\mathbf{z}_t$. For a noisy latent $\mathbf{z}_t$, we can get $\mathbf{z}_{t-1}$ using the following process: firstly predicts noise ϵ using the Denoising UNet $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$, then calculates $\mathbf{z}_0$ with the inverse process of Eq. 1, finally gets $\mathbf{z}_{t-1}$ using Eq. 2. In other words, strictly following the formulas in DDPM, we are only allowed to move to $\mathbf{z}_0$ and then $\mathbf{z}_{t-1}$, starting from $\mathbf{z}_t$. The direct moving from $\mathbf{z}_t$ to $\mathbf{z}_{t_1}$ in DiffSwap is suboptimal, which will degrade the overall performance.

To be more in line with the formulas in DDPM, we propose an improved midpoint sampling method, which also samples within two steps but can reduce the information loss, compared to the original midpoint sampling [53]. Specially, starting from $\mathbf{z}_t$, we can obtain $\mathbf{z}_0$ and $\mathbf{z}_{t-1}$ using the process introduced above. Then, we can obtain $\mathbf{z}_{t-2}$ using Eq. 2 again, which is an efficient linear transformation without using the Denoising UNet. Through repeating $t - t_1$ times linear transformation, we can get $\mathbf{z}_{t_1}$ in a more accurate and graceful way than DiffSwap. Finally, $\mathbf{z}_0^*$ is obtained by using Eq. 9 on $\mathbf{z}_{t_1}$.

4. Experiments

Dataset. We split the CelebA-HQ dataset [13] into a training set of 29,000 images and a test set of 1,000 images, by random selection. Our diffusion model is trained on the training set of CelebA-HQ and FFHQ [14], and evaluated in the test set of CelebA-HQ and FF++ [36]. The competitors are evaluated by using their public pre-trained networks or other open-source projects.

Metrics. The quantitative evaluations are performed in terms of four metrics: identity retrieval accuracy (ID.), expression error (Exp.), pose error (Pose.), and mean squared error (MSE.). For ID., we employ CosFace [42] to perform identity retrieval. For Exp., we adopt the expression embedding model [52] to compute the Euclidean distance between I_{out} and I_{exp}. For Pose., we use a pose estimator [37] to estimate head pose and compute the Euclidean distance between I_{out} and I_{bkg}. MSE. is used to measure the pixel difference between the predicted image and ground truth. It is noteworthy that in the calculation of the metrics in different tasks, the reference images are accordingly changed with the source images.

Implementation Details. The network architecture of our DiffSFSR follows the latent diffusion model [35], which has a $4 \times 64 \times 64$ latent space. DiffSFSR is trained from SD-1.4 in 512×512 resolution with an AdamW optimizer. The hyper-parameters are set as $\lambda_1 = 0.003$, $\lambda_2 = 0.01$. In the first 100k steps, the learning rate is set to $1e-5$ which decays linearly in the following 100k steps. 8 NVIDIA Tesla A100 GPUs are used to train our diffusion model with a global batch size of 64. In inference time, we apply a PNDM [24] sampler with 50 steps, which takes roughly 1 second to generate an image.

4.1. Fine-grained Expression Controlling Results

Fig. 5 shows 5 samples of the fine-grained expression synthesis. More synthesis results from the 135 expression labels [4] are available in supplemental material. To the best

Methods		CelebA-HQ			FF++		
		ID.↑	Exp.↓	Pose.↓	ID.↑	Exp.↓	Pose.↓
Swap EXP	Face2Face	87.9	3.25	-	96.8	2.83	-
	StyleHEAT	98.2	2.43	-	97.7	2.16	-
	DiffSFSR (ours)	**99.9**	**0.58**	-	**98.9**	**0.68**	-
Swap ID	FaceShifter	94.5	0.65	**2.13**	95.4	1.10	**1.62**
	SimSwap	**98.8**	0.93	2.89	**98.0**	1.46	2.87
	HifiFace	85.1	1.11	3.38	92.4	1.80	3.32
	E4S	81.6	2.77	6.99	91.5	2.34	3.96
	DiffSFSR (ours)	90.8	**0.32**	2.59	91.0	**0.49**	3.89
Swap All	Hybrid Method	76.9	2.24	7.07	83.5	2.14	5.67
	DiffSFSR (ours)	**90.2**	**0.55**	**6.00**	**90.7**	**0.73**	**5.19**

Table 1. The quantitative results in the three tasks: "Swap All" denotes SFSR task, "Swap ID" denotes face swapping task, and "Swap EXP" denotes face reenactment task. The scores in ID. and Exp. are scaled up by a factor of 100 for simplicity.

of our knowledge, *there is currently no face generation or manipulation method in academic or industry, that can reach this level of fine-grained expression control.* As mentioned above, the dataset [4] provides 728,946 facial expression images labeled with 135 categories. In our work, a reference expression image is randomly selected from the corresponding category according to the input expression text. As observed in Fig. 5 and supplemental material, the expression of the synthesized facial image is similar to that of the reference image. Readers can zoom in for more details.

User Study. We conduct a user study with Fuxi Youling Crowdsourcing[1] to evaluate the quality of the fine-grained expression synthesis, in terms of expression consistency and identity consistency. For each sample, 27 participants were recruited to answer two questions: whether the synthesized face has a consistent identity and a consistent expression with the reference images? The results of user study are that, the identity consistency is 95.6% (with a variance of 2.6%) and the expression consistency is 90.4% (with a variance of 3.4%). The statistics clearly demonstrate that our method has the ability of simultaneous identity and expression preserving.

We conduct another user study to further evaluate the expression consistency of the proposed method. For each sample, 30 participants were recruited to score the expression consistency between the synthesized face and the reference expression image (with a min score of 1 and a max score of 5; 1 refers to very inconsistent and 5 refers to very consistent). Finally, our method achieves a score of 4.08 (with a variance of 0.81), indicating its ability to generate very consistent expressions.

In summary, our method can achieve fine-grained expression control while maintaining identity consistency.

4.2. Comparisons

Comparison with Text-to-Image Method. We compare our method with the SOTA open-source text-to-image model Stable Diffusion XL (SDXL) [31]. As shown in Fig. 6, the two methods take as input the same prompt and the same fine-grained expression labels. The additional input to our framework is a portrait of "Melinda May". Compared with our results, SDXL can't synthesize the accurate expressions corresponding to "enjoyment", "anxiety" or "grief". The dilemma faced by most text-to-image methods is that they can only recognize a few limited expression labels. Models like ControlNet [51] need a large amount of training data and couple the background to the face attributes. Our framework also supports expression travel by interpolating between embeddings, so that more fine-grained expressions can be generated.

Figure 5. A subset of 135 classes expression synthesis samples. Please zoom in for more details.

Comparison with Hybrid Methods. As there is currently no SFSR method, we construct a hybrid method as the potential competitor by directly combining the best face-swapping method in terms of ID. score, SimSwap [5], and the best face reenactment method in terms of Exp. score, StyleHEAT [49], based on the statistics reported in the parts about Swap ID and Swap EXP in Tab. 1.

As shown in the part about Swap All in Tab. 1, our method outperforms the hybrid method with a significant margin in all metrics. From the visual results shown in Fig. 7, our method can produce more accurate expressions and poses than the hybrid method, due to the powerful ability of expression embedding [52] and latent diffusion model [35].

Comparison with Face Reenactment Methods. Our DiffSFSR can also conduct the face reenactment task by setting the inputs I_{bkg} and I_{id} as the same image. We make

[1] https://fuxi.163.com/solution/data

Figure 6. Comparison with text-to-image method SDXL, and illustration of expression travel.

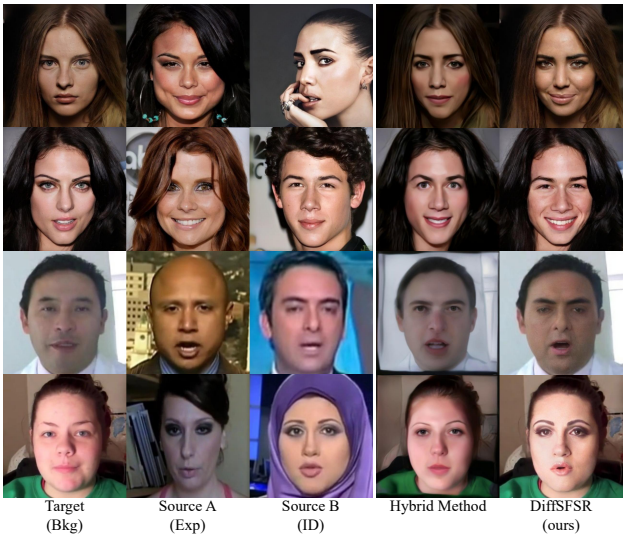

Figure 7. Qualitative comparison in SFSR task.

a comparison with two SOTA face reenactment methods including Face2Face[41] and StyleHEAT[49], and use their pre-trained networks.

From the statistics in the part about Swap EXP in Tab. 1, our method outperforms the two competitors in all metrics. The advantage is more obvious in the qualitative comparison shown in Fig. 8, where the expression in our results is more similar to the source, and the identity and pose are more similar to the target.

Comparison with Face Swapping Methods. Similar to face reenactment, our DiffSFSR can also conduct the face swapping task by setting the inputs I_{bkg} and I_{exp} as the same. We make comparisons with five SOTA face

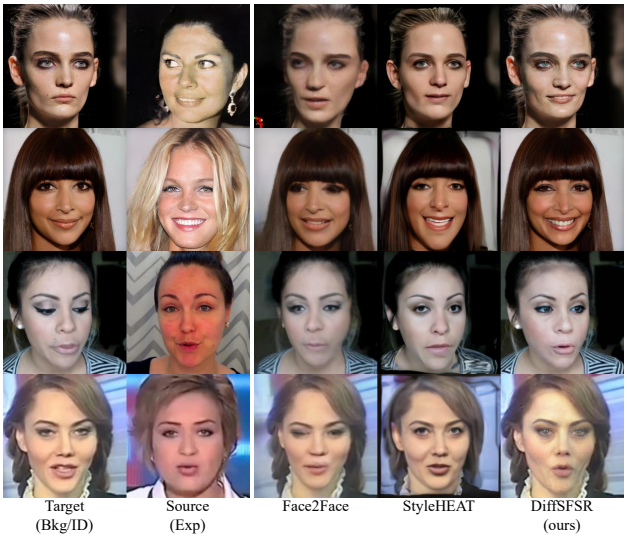

Figure 8. Qualitative comparison in face reenactment task.

Methods	Identity.	Expression.	Realism.	Image Quality.
SimSwap	2.84 (p=0.124)	3.44 (p<0.001)	2.71 (p<0.001)	2.69 (p<0.001)
DiffSwap	2.58 (p<0.001)	3.59 (p<0.05)	3.19 (p<0.005)	3.04 (p<0.001)
E4S	2.87 (p=0.232)	2.72 (p<0.001)	3.04 (p<0.001)	3.03 (p<0.001)
DiffSFSR (ours)	**3.01**	**3.79**	**3.72**	**3.60**

Table 2. Uers study in face swapping methods. The best values are highlighted in bold. The ANOVA tests are conducted, in which a p-value less than 0.05 is considered to indicate a statistically significant difference from the performance of our method.

swapping methods including FaceShifter[22], SimSwap[5], HifiFace[44], E4S[25] and DiffSwap[53]. For SimSwap, E4S and DiffSwap, we directly use their public pre-trained networks. As FaceShifter and HifiFace do not make their codes publicly available, we use the implementations from the open-source community[2][3].

From the statistics in the part about Swap ID in Tab. 1, since we focus on preserving both ID and expression, our method outperforms all the competitors in Exp. and achieves promising performance in ID. and Pose. As observed in Fig. 9, except for SimSwap, our results are more similar to the source faces in terms of inner facial features, e.g., beard. There are obvious artifacts in the results of HifiFace. And the faces generated by E4S do not blend well into the background, leading to less natural results. In addition to expression preserving, our advantage over SimSwap is that our method can generate faces with better image quality, with less blur and artifacts, due to the powerful image generation capability of the latent diffusion model [35].

User Study with Face Swapping Methods. To comprehensively compare our method with other face-swapping methods, we implemented another human evaluation experiment. For simplicity, we compare our DiffSFSR with two most recent methods, DiffSwap[53] and E4S[25], as well as

[2] https://github.com/richarduuz/Research_Project
[3] https://github.com/xuehy/HiFiFace-pytorch

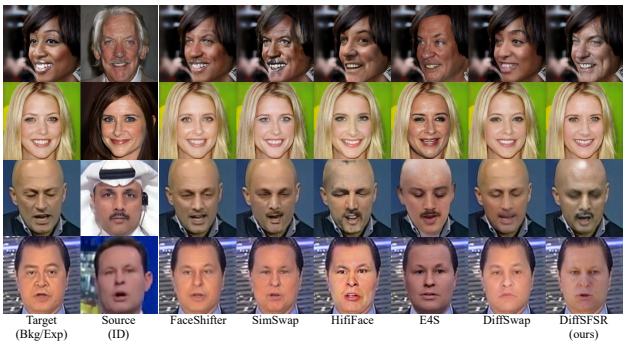

Figure 9. Qualitative comparison in face swapping task.

SimSwap[5], which shows superior performance in terms of ID consistency (see Tab. 1). 50 participants were recruited to score the results of all methods in terms of ID consistency, expression consistency, realism, and image quality (with a min score of 1 and a max score of 5; 1 refers to the worst and 5 refers to the best). As shown in Tab. 2, our method registers the topmost scores across all measured metrics. Moreover, in terms of statistical significance, our method is competitive with other methods in ID consistency while considerably surpassing its competitors on expression consistency, realism, and image quality under a p-value of 0.05. In summary, our method is comparable to other methods in terms of ID consistency but can produce more accurate expressions and more realistic, high-quality images.

4.3. Ablation Study

We conduct ablation study to demonstrate the effectiveness of background conditioning and compound identity embedding, by removing them individually during training. The study on improved midpoint sampling is introduced in the supplemental material. In the method without background conditioning, we compensate with a segmentation map to specify face region in training and provide background only in inference. CelebA-HQ is used as the test set.

Effect of background conditioning. Without background conditioning, the diffusion model can't learn the accurate lighting and face pose, can't generate faces of higher image quality and can't be consistent with the background. These arguments are well supported by the results shown in Fig. 10. The generated faces without background conditioning suffer from inaccurate lighting, exhibit more face pose errors, lack seamless blending with the background, and are comparatively more blurry.

Effect of compound identity embedding. Compound identity embedding can significantly improve identity similarity. As shown in the 1st, 2nd and 3rd rows in Fig. 10, each time we remove an identity embedding, the generated faces become more similar to source A in terms of identity, and less similar to source B which provides the expected identity. This phenomenon indicates that the residual identity attribute in the expression embedding will be unexpectedly transferred to the result when the identity embedding is weaker than the expression embedding. Notably in the fourth row, the inadequate identity representation even incorrectly puts glasses on the results. As shown in Tab. 3, the identity similarity in our results surpasses all the ablation methods.

Methods	ID.↑	Exp.↓	Pose.↓	MSE.↓
w./o. Bkg Condi.	81.5	0.66	8.97	51.40
w./o. ID Emb.3	74.9	0.63	7.54	52.09
w./o. ID Emb.2	37.9	**0.56**	7.90	52.98
Full Model	**87.0**	0.63	**7.36**	**51.23**

Table 3. Quantitative results in ablation study. The score in ID. and Exp. are scaled up by a factor of 100 for simplicity.

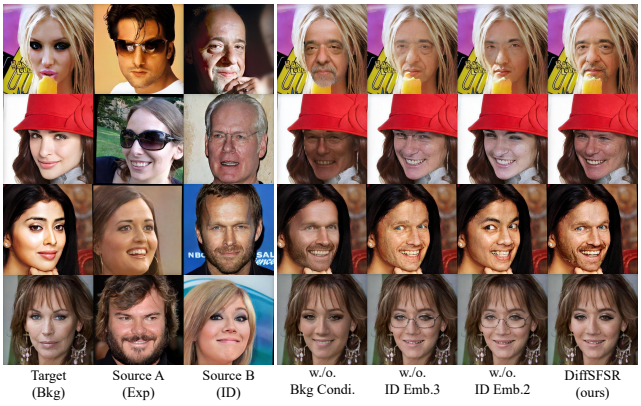

Figure 10. Qualitative results in ablation study.

5. Conclusion

Given a text prompt, an expression label, and a selfie photo, our personalized face generation framework can produce high-fidelity and identity-expression-preserving portraits. To realize the framework, we propose a new diffusion model that can conduct simultaneous face swapping and reenactment tasks. Extensive experiments have demonstrated the controllability and scalability of the proposed framework. We hope our efforts can inspire future work in personalized generation frameworks to explore the use of more modalities as conditioning to achieve higher controllability and image quality.

6. Acknowledge

Renshuai Liu and Xuan Cheng were partially supported by Natural Science Foundation of Fujian Province of China (No. 2023J05001), Natural Science Foundation of Xiamen, China (No. 3502Z20227012), NSFC (No. 62077039) and the Fundamental Research Funds for the Central Universities, China (No. 20720230106). Yu Ding was partially supported by Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054).

References

[1] 7whitefire7. https://civitai.com/images/3486388. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proc. of CVPR*, pages 18208–18218, 2022. 2

[3] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, pages 122–138, 2018. 3

[4] Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Trans. Affect. Comput.*, 13(4):1906–1916, 2022. 2, 3, 5, 6

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proc. of ACM MM*, pages 2003–2011, 2020. 1, 2, 3, 6, 7, 8

[6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of CVPR*, pages 8789–8797, 2018. 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. of CVPR*, pages 4690–4699, 2019. 4

[8] Xiaoxiong Du, Jun Peng, Yiyi Zhou, Jinlu Zhang, Siting Chen, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Pixelface+: Towards controllable face generation and manipulation with text descriptions and segmentation masks. In *Proc. of ACM MM*, pages 4666–4677, 2023. 3

[9] FrenzyX. https://civitai.com/images/3491907. 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NIPS*, 2020. 2, 3, 5

[11] Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proc. of CVPR*, pages 6080–6090, 2023. 3

[12] Youngjoo Jo and Jongyoul Park. SC-FEGAN: face editing generative adversarial network with user's sketch and color. In *Proc. of ICCV*, pages 1745–1753, 2019. 1

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. of ICLR*, 2018. 5

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. of CVPR*, pages 4401–4410, 2019. 2, 5

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proc. of CVPR*, pages 8107–8116, 2020.

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. of NIPS*, pages 852–863, 2021. 2

[17] Wei Zhang Keyu Chen, Changjie Fan and Yu Ding. 135-class emotional facial expression dataset. https://ieee-dataport.org/documents/135-class-emotional-facial-expression-dataset, 2023. 2

[18] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163, 2018. 3

[19] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-swap: A simple enhancement for face-swapping with smoothness. In *Proc. of CVPR*, pages 10769–10778, 2022. 3

[20] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, 2009. 4

[21] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proc. of CVPR*, pages 5548–5557, 2020. 1

[22] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1, 2, 3, 7

[23] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proc. of CVPR*, pages 18176–18186, 2022. 1, 2

[24] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2021. 5

[25] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8578–8587, 2023. 7

[26] Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. Text2facegan: Face generation from fine grained textual descriptions. In *International Conference on Multimedia Big Data*, pages 58–67, 2019. 1

[27] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: subject agnostic face swapping and reenactment. In *Proc. of ICCV*, pages 7183–7192, 2019. 1, 2, 3

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsganv2: Improved subject agnostic face swapping and reenactment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):560–575, 2023. 3

[29] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proc. of ICCV*, pages 2065–2074, 2021. 1, 2

[30] Bo Peng, Hongxing Fan, Wei Wang, Jing Dong, and Siwei Lyu. A unified framework for high fidelity face swap and expression reenactment. *IEEE Trans. Circuits Syst. Video Technol.*, 32(6):3673–3684, 2022. 3

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6

[32] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proc. of ACM MM*, pages 484–492, 2020. 1

[33] Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proc. of ECCV*, pages 835–851, 2018. 3

[34] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *Proc. of CVPR*, pages 2287–2296, 2021. 1

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*, pages 10674–10685, 2022. 2, 3, 5, 6, 7

[36] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proc. of ICCV*, pages 1–11, 2019. 5

[37] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *Proc. of CVPR Workshops*, pages 2074–2083, 2018. 5

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. of CVPR*, pages 815–823, 2015. 4

[39] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proc. of ICCV*, 2023. 3

[40] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot head swapping in the wild. In *Proc. of CVPR*, pages 10779–10788, 2022. 3

[41] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *Proc. of CVPR*, pages 2387–2395, 2016. 7

[42] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proc. of CVPR*, pages 5265–5274, 2018. 5

[43] Tianren Wang, Teng Zhang, and Brian C. Lovell. Faces à la carte: Text-to-face generation via attribute disentanglement. In *Proc. of WACV*, pages 3379–3387, 2021. 1

[44] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *Proc. of IJCAI*, pages 1136–1142, 2021. 7

[45] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proc. of ECCV*, pages 622–638, 2018. 1, 2, 3

[46] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proc. of CVPR*, pages 2256–2265, 2021. 1, 2

[47] 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 *Proc. of ECCV*, pages 54–71, 2022. 3

[48] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proc. of CVPR*, pages 7622–7631, 2022. 3

[49] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 6, 7

[50] Hao Zeng, Wei Zhang, Keyu Chen, Zhimeng Zhang, Lincheng Li, and Yu Ding. Paste you into game: Towards expression and identity consistency face swapping. In *IEEE Conference on Games*, pages 1–8, 2022. 4

[51] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of ICCV*, pages 10674–10685, 2023. 2, 6

[52] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *Proc. of CVPR*, pages 6759–6768, 2021. 4, 5, 6

[53] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proc. of CVPR*, pages 8568–8577, 2023. 2, 3, 4, 5, 7

[54] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proc. of CVPR*, pages 4176–4186, 2021. 1

[55] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Trans. Graph.*, 39(6):221:1–221:15, 2020. 1

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of ICCV*, pages 2242–2251, 2017. 3