

EmojiDiff: Advanced Facial Expression Control with High Identity Preservation in Portrait Generation

Liangwei Jiang Ruida Li Zhifeng Zhang Shuo Fang Chenguang Ma

Terminal Technology Department, Alipay, Ant Group

{jiangliangwei.jlw, ruida.lrd, jason.zzf, fangshuo.f, chenguang.mcg}@antgroup.com

Project Page: <https://emojidiff.github.io>

Abstract

This paper aims to bring fine-grained expression control to identity-preserving portrait generation. Existing methods tend to synthesize portraits with either neutral or stereotypical expressions. Even when supplemented with control signals like facial landmarks, these models struggle to generate accurate and vivid expressions following user instructions. To solve this, we introduce **EmojiDiff**, an end-to-end solution to facilitate simultaneous dual control of fine expression and identity. Unlike the conventional methods using coarse control signals, our method directly accepts RGB expression images as input templates to provide extremely accurate and fine-grained expression control in the diffusion process. As its core, an innovative decoupled scheme is proposed to disentangle expression features in the expression template from other extraneous information, such as identity, skin, and style. On one hand, we introduce **ID-irrelevant Data Iteration (IDI)** to synthesize extremely high-quality cross-identity expression pairs for decoupled training, which is the crucial foundation to filter out identity information hidden in the expressions. On the other hand, we meticulously investigate network layer function and select expression-sensitive layers to inject reference expression features, effectively preventing style leakage from expression signals. To further improve identity fidelity, we propose a novel fine-tuning strategy named **ID-enhanced Contrast Alignment (ICA)**, which eliminates the negative impact of expression control on original identity preservation. Experimental results demonstrate that our method remarkably outperforms counterparts, achieves precise expression control with highly maintained identity, and generalizes well to various diffusion models.

1. Introduction

The development of diffusion model [19, 20, 31, 37, 39] has catalyzed significant advancements in identity (ID) cus-

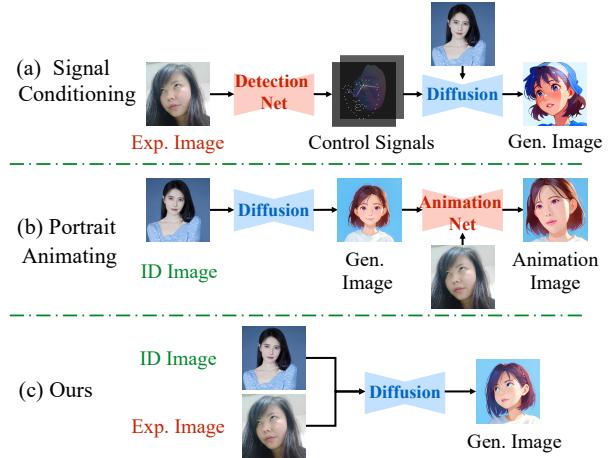


Figure 1. **Expression customization by different methods.** (a) The methods extract the control signals from expressions and then inject them into diffusion models. (b) The methods first generate expressionless images, and manipulate them through animation. (c) The proposed method directly incorporates portrait images and reference expressions into diffusion models, bypassing the limitations of the detection network and post-processing.

tomization [17, 43, 47]. This task involves generating images that align with the specified identity of portrait images, with wide-ranging applications in AI portraits [22, 25, 28], image animation [21, 44], and virtual try-ons [5, 6, 40]. Despite substantial advancement, existing methods reveal a notable limitation: monotonous facial expressions. This shortcoming constrains the expressiveness of the generated portraits, undermining their ability to convey subtle emotions and severely hindering practical applications.

At the same time, researchers seek to use reference images to manipulate portrait expressions during the diffusion process and have devoted many efforts [9, 23, 33]. The representatives of them, including MediaPipe [7], Arc2Face [32], and FineFace [42], focus on extracting control signals from expression images and then injecting them into

the diffusion model. However, fine-grained expression control conflicts with identity preservation. Existing methods use coarse control signals (*e.g.*, landmarks, poses) instead of original RGB expression images to prevent interference from identity-related information (such as skin and lighting). However, these signals lose many facial details (*e.g.* frowning, pouting, cheek twitching), which are also limited by the robustness and accuracy of third-party networks. Besides, the integration of the above methods with portrait generation approaches also affects the vanilla identity-preserving generation. This limitation arises because these methods overlook the relationship between expression control and identity control, resulting in facial muscle changes caused by expression control reducing identity fidelity.

In this paper, we propose EmojiDiff, an innovative end-to-end solution that seamlessly integrates fine-grained expression control with high-fidelity portrait generation. To achieve accurate expression control, we argue that the model should derive the motion directly from the original reference expressions, permitting comprehensive expression representation and robust generalization [16, 45]. However, expression images themselves contain irrelevant information like identity and style. If the model directly replicates the appearance and structure from the expression reference, it compromises the identity and style of generated images (referred to as identity leakage and style leakage). To address these challenges, a decoupled scheme is proposed. On one hand, we innovate ID-irrelevant Data Iteration (IDI), a novel strategy for decoupled training. Unlike previous methods [2, 32, 42] that employ same-identity data pairs, IDI sequentially manipulates the expression and identity of the image to synthesize high-quality data pairs with consistent expressions but differing identities. Training the model with these cross-identity expression pairs facilitates the transfer of detailed expressions to the generated output while implicitly filtering out identity information concealed in the expressions. On the other hand, we discover that the network layers serve distinct functions, with specific layers responsible for incorporating expression features while omitting expression-irrelevant information. By injecting expression features solely into these layers, we effectively address most style leakage issues. To optimize the compatibility between expression and identity control, we develop the ID-enhanced Contrast Alignment (ICA) to further fine-tune the model. ICA efficiently computes the identity loss and expression loss to align identity representations of images generated with and without the expression control, thereby mitigating the negative impacts of expression control on identity and improving identity fidelity. The main contribution of this paper lies in four-fold:

- 1) We propose an end-to-end solution that integrates fine-grained expression control, high-fidelity ID preservation, and strong adaptability to various diffusion models.

- 2) We innovate a decoupled scheme to disentangle expression features from extraneous information in RGB expression images, which effectively prevents ID leakage and style leakage through ID-irrelevant Data Iteration and selective feature injection at expression-sensitive layers.

- 3) We propose ID-enhanced Contrast Alignment, a fine-tuning strategy designed to ensure generated portraits maintain stable identity characteristics across diverse expressions, thereby improving identity fidelity.

- 4) We extensively evaluate the performance of the proposed method on different base models and expressions, showing that EmojiDiff achieves better performance compared to state-of-the-art methods.

2. Related Work

2.1. ID-Preserving Image Generation

The task of identity-preserving generation [14, 20, 39] seeks to empower diffusion models to synthesize images of given identities. Recent studies [3, 17, 22, 25] have explored tuning-free methods, which utilize face embeddings to conditioning generation without altering raw models. For instance, FaceChain [28] introduces an independent adapter before the cross-attention layer, preventing interference between face and text conditions. IP-Adapter [47] introduces a novel decoupled cross-attention mechanism for identity injection. InstantID [43] proposes an IdentityNet to incorporate facial key points with image conditions. Although these methods can quickly generate customized portraits, they severely lack fine control over the expressions.

2.2. Expression Customization

Most methods for expression animation integrate control signals extracted from expression images into diffusion models. ControlNet [49] is a pioneering effort that enables the incorporation of human poses, depth maps, and other image conditions. Following this framework, MediaPipe [7] specializes in extracting facial landmarks from reference expressions to preserve more expression details. Arc2Face [9] extract 3D normal maps [13] to personalize expressions based on given portraits. Additionally, some studies pursue more flexible control. Collaborative Diffusion [23] allows localized editing of facial parts using semantic maps from expression references. [33] combined valence & arousal, and action units and GANmut [8] to characterize emotion information. DiffSFSR [27] utilizes an emotional dictionary to transfer facial expressions to the generated image. FineFace [42] employs action units to allow localized control by describing the intensity of facial muscle movements.

In practice, portrait animation [4, 10, 21, 30, 48] serves as an indirect tool for expression customization. Notable examples, such as FaceAdapter [18], X-Portrait [45] and LivePortrait [16], transfer expressions from expression tem-

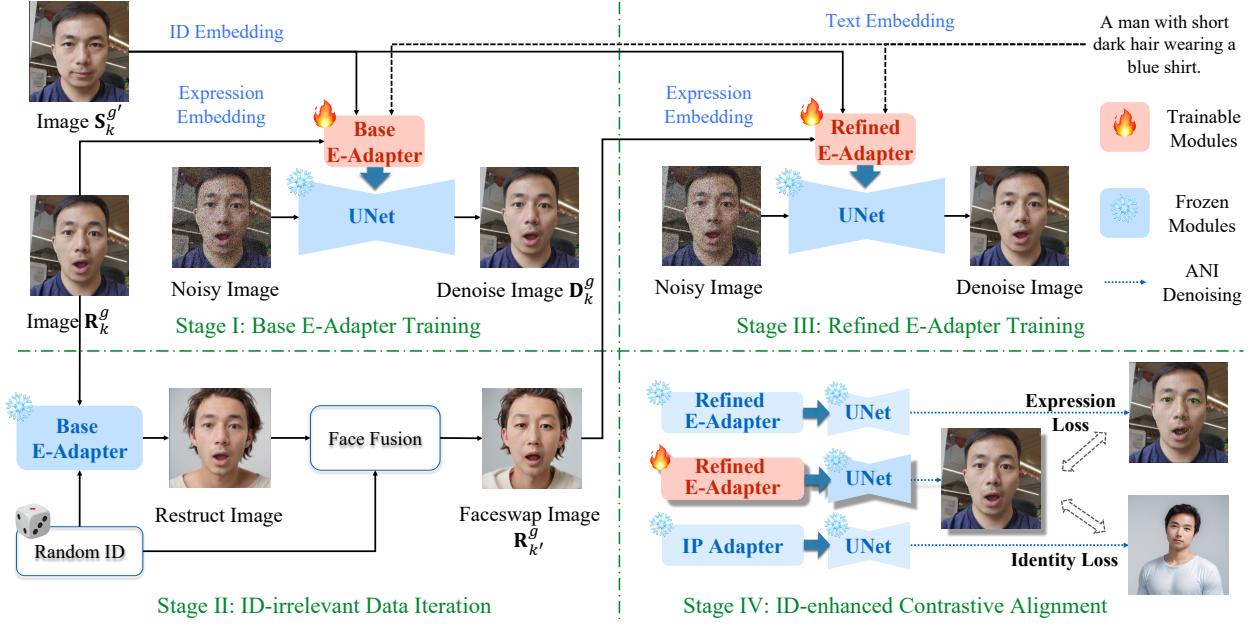


Figure 2. Overview of the proposed method. To integrate expression control into diffusion models, we aim to train the model using cross-identity triplet data $\{\mathbf{S}_k^{g'}, \mathbf{R}_{k'}^g, \mathbf{D}_k^g\}$ and mitigate the negative impact on the original structure through contrastive alignment. The method involves four stages. First, the fundamental expression controller (*i.e.*, Base E-Adapter) is trained with same-identity triplet data (the detailed structure of the E-Adapter is illustrated in Fig. 3). Next, the trained Base E-Adapter and FaceFusion [11] are utilized to alter the identity of portraits while maintaining consistent expressions, thereby creating cross-identity expression pairs $\{\mathbf{R}_{k'}^g, \mathbf{D}_k^g\}$. Subsequently, the Refined E-Adapter is trained using newly synthesized data, facilitating dual control of identity and expression without ID leakage. Finally, the Refined E-Adapter is fine-tuned by introducing expression and identity loss, further minimizing its negative impact on identity.

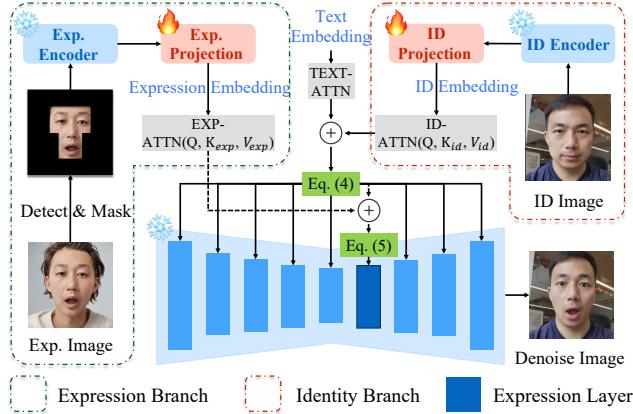


Figure 3. The proposed E-Adapter. The embeddings of identity and expression images are obtained through their respective branches. Subsequently, the ID embedding and the text embedding are incorporated into all network layers through attention, as demonstrated in Eq. (3). In designated expression layers, all embeddings are integrated into the network, as depicted in Eq. (5).

plates to portrait images. However, these methods cannot customize identities and expressions simultaneously, but instead post-process the images generated by the diffusion

model, adversely affecting identity fidelity and image quality. Our method avoids the need for third-party networks and post-processing. Instead, it directly uses RGB expression images and portrait images to generate expressive images. Fig. 1 illustrates the differences among these methods.

3. Methodology

Given a source portrait \mathbf{S} , a textual description \mathbf{T} , and a reference \mathbf{R} , our method aims to generate customized images \mathbf{D} . To enable the model to accept RGB expression inputs, we aim to construct a high-quality ternary data sets $\mathcal{T} = \{\mathbf{S}_k^{g'}, \mathbf{R}_{k'}^g, \mathbf{D}_k^g\}$, where k and g denote the k -th ID and the g -th expression, respectively. During training, the model takes $\mathbf{S}_k^{g'}, \mathbf{R}_{k'}^g$ as inputs and \mathbf{D}_k^g as the learning objective, represented as $\mathbf{D}_k^g = \mathcal{F}(\mathbf{S}_k^{g'}, \mathbf{R}_{k'}^g)$, where \mathcal{F} is the trained expression controller. This setup allows the model to maintain the identity of \mathbf{S} while accurately replicating the expression from \mathbf{R} . However, existing methods [45, 46] fail to synthesize high-quality cross-identity expression pairs $\{\mathbf{R}_{k'}^g, \mathbf{D}_k^g\}$ in \mathcal{T} . Besides, the integration of expression control hinders the original ID control.

To overcome these issues, our method is divided into four stages: Firstly, we design the expression controller and train a base version (Base E-Adapter) using same-identity

ternary data sets as detailed in Sec. 3.2. The images generated by Base E-Adapter exhibit a very high expression consistency with given expression references, although they also have a similar appearance to the references (ID leakage). Secondly, we develop the ID-irrelevant Data Iteration (IDI) to alter the identity of the images generated by Base E-Adapter, thus promoting the given reference expressions and the altered images form data pairs with the same expression but different identity, as outlined in Sec. 3.3. Using the newly created cross-identity dataset, we retrain the proposed expression controller to develop the Refined E-Adapter. Refined E-Adapter accurately controls expression details while preventing ID leakage when generating portrait images (see Sec. 3.4). Finally, we introduce ID-enhanced Contrast Alignment (ICA) to further fine-tune the Refined E-Adapter, reducing the negative impact of expression manipulation on ID control (Sec. 3.5). Fig. 2 provides an overview of our pipeline.

3.1. Preliminary

Latent Diffusion Model. The latent diffusion models involve a diffusion process and a reverse process in the latent space. During the diffusion process, an image x is first projected to a smaller latent representation by a VAE [24] encoder. Then, random noise is gradually added to the representation z_0 . The noisy representation z_t is derived as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

where t , $\bar{\alpha}_t$ denotes timestep and predefined function of t , respectively. Conversely, during the denoising process, a U-Net [38] is trained to predict the added noise ϵ from the noisy latent representation z_t . The training objective aims to minimize the loss function as follows:

$$\mathcal{L}_{sd} = E_{z_0, \epsilon \sim \mathcal{N}(\mu, \sigma^2)} \left[\|\epsilon - \epsilon_\theta(z_t, t, C)\|_2^2 \right], \quad (2)$$

where C represents the condition.

Image Prompt Adapter. IP-Adapter [47] introduces a novel approach to incorporating an image prompt, working together with a text prompt to control image generation. A decoupled cross-attention is used to separate cross-attention layers for text features and image features, formulated as:

$$Z = \text{Attn}(Q, K_t, V_t) + \lambda \cdot \text{Attn}(Q, K_i, V_i), \quad (3)$$

where λ is a weight factor, and $\{Q, K, V\}_{\text{source}}$ refers to the source of the feature tensor.

3.2. Base E-Adapter Training

Cross-identity expression pairs $\{\mathbf{R}_k^g, \mathbf{D}_k^g\}$ are essential for training expression controllers, as they disentangle expression features in RGB expression images from other irrelevant information. Nonetheless, existing methods [11, 16,

45, 46] struggle to synthesize data pairs that maintain high expression consistency while exhibiting significant identity discrepancies, particularly with extreme expressions. Therefore, we seek to develop a novel strategy for data construction. Reviewing previous works [45, 46], we notice that they avoid using same-identity expression pairs $\{\mathbf{R}_k^g, \mathbf{D}_k^g\}$ for controller training to prevent identity leakage during inference. However, our insights suggest that the controller trained with same-identity pairs exhibits superior expression transfer capabilities, and it is highly suitable for data construction when combined with portrait generation methods. Specifically, we divide the synthesis of cross-identity expression pairs into two sequential steps. Initially, we utilize same-identity data to train an expression controller that accurately transfers reference expressions onto generated images without considering identity leakage. Subsequently, we apply portrait generation techniques to further alter the identity of the controller-generated image. With the powerful generative abilities of the diffusion model, the modified image forms a high-quality cross-identity data pair with the given expression template.

In this section, we first propose the expression controller (E-Adapter) suitable for portrait generation and then train its basic version for subsequent data construction. Inspired by IP-Adapter [47], the designed E-Adapter integrates identity and expression through an attention mechanism, ensuring the controller remains lightweight. As illustrated in Fig. 3, E-Adapter includes two main control branches: the identity and expression branches, which process portrait and expression images, respectively. In the identity branch, we inherited IP-Adapter to encode portrait images into identity embeddings and inject them into the U-Net along with text embeddings using a decoupled attention, formulated as:

$$\begin{aligned} Z_{id} = & \text{Attn}(Q_{noise}, K_{text}, V_{text}) \\ & + \lambda_{id} \cdot \text{Attn}(Q_{noise}, K_{id}, V_{id}), \end{aligned} \quad (4)$$

where λ_{id} refers to the strength of the ID control.

In the expression branch, the local image is preprocessed from the origin expression reference, which removes irrelevant elements such as the background, face contour, etc. We then extract the expression embedding from the local image and inject it into the same cross-attention module with the identity embedding, denoted as:

$$Z_{exp} = Z_{id} + \lambda_{exp} \cdot \text{Attn}(Q_{noise}, K_{exp}, V_{exp}), \quad (5)$$

where λ_{exp} refers to the strength of the expression control.

Noteworthy, we discover that expression images contain some abstract, expression-irrelevant information (e.g., lighting, skin), which may affect the quality and style of the generated image. As shown in Fig. 4, we visualize different cross-attention layers in the adapter. It becomes evident that the cross-attention layers serve distinct responsibilities, some of which are responsible for injecting facial

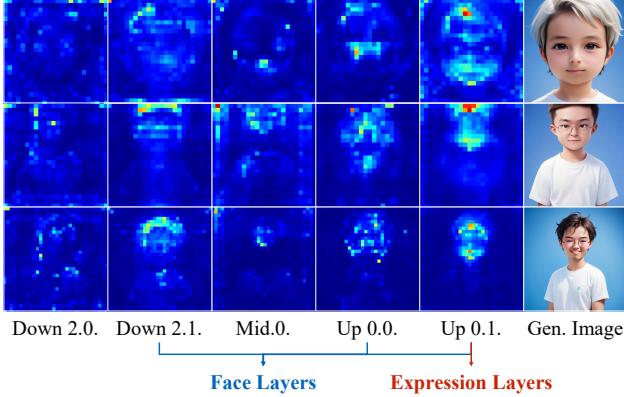


Figure 4. **Visualization of the attention maps**, where the regions colored red/blue represent the most/least salient areas.

information and showing pronounced responses in the facial region. Furthermore, certain layers are sensitive to structural information while neglecting style information, making them ideal for expression control. Consequently, we inject expression embeddings exclusively into these layers (*i.e.*, Expression Layer), effectively avoiding style leakage.

Based on the designed controller, we utilize the same-identity data $\{\mathbf{S}_k^{g'}, \mathbf{R}_k^g, \mathbf{D}_k^g\}$ to train the Base E-Adapter, adopting denoising loss \mathcal{L}_{sd} throughout the process.

3.3. ID-irrelevant Data Iteration

The image identity constructed directly using the Base E-Adapter is greatly influenced by the expression reference’s appearance. The core idea of IDI is to minimize this disruption to further increase the identity distinction between the generated image and the expression reference. Initially, we innovatively reuse the ID control branch of the base E-Adapter to modify the identity of the generated image. Specifically, we utilize expression images \mathbf{R}_k^g inherited from training as the expression condition but randomly selected portraits as the input of the identity branch, combining with the diffusion model to synthesize the reconstruct image. Since the Base E-Adapter has been exposed to all expression images, it effectively transmits expressions to the reconstruct images. Through the identity branch, we also integrate the identity of randomized portraits. Notably, the identity of the generated image does not need to precisely match the given ID condition; it simply needs to differ from the identity of the expression image. Additionally, we employ FaceFusion [11] on the reconstruct image to create the faceswap image, further increasing the identity disparity between the generated portrait and the expression reference. Occasionally, the whole process might modify expressions (*e.g.*, changing from blinking to non-blinking), and we employ facial blendshapes [15] and landmark [29] differences to exclude expression-changed data.

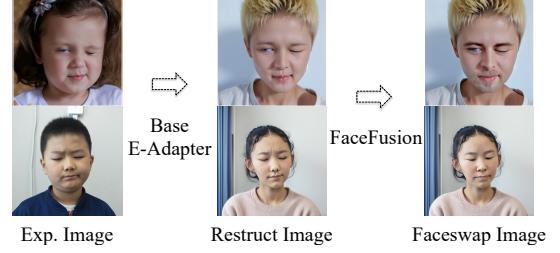


Figure 5. **Examples of Stage II**.

Through these processes, we acquire cross-identity expression pairs $\{\mathbf{R}_{k'}^g, \mathbf{D}_k^g\}$ that maintain highly consistent expressions while exhibiting completely distinct identity, and the example is illustrated in Fig. 5.

3.4. Refined E-Adapter Training

Utilizing the newly developed cross-identity data, we train a new expression controller termed the Refined E-Adapter. The key distinction from the Base E-Adapter training is that the expression and identity are no longer sourced from the same portrait. Consequently, the identity and expression control branches are implicitly decoupled in their functions, each responsible solely for conveying identity and expression, respectively.

3.5. ID-enhanced Contrast Alignment

Both identity control and expression control manipulate the face of the generated image, causing potential interference between the two. Inevitably, incorporating expression control affects the original identity control in portrait generation tasks. Our findings indicate that emphasizing expression control alters muscle details and facial attributes, such as the presence of glasses, due to the generative diversity inherent in the diffusion model.

To this end, we propose ICA to fine-tune the E-Adapter. Based on the same inputs, ICA incorporates the frozen Refined E-Adapter and the original identity controller (IP-Adapter [47]) as supervision to calculate expression loss \mathcal{L}_{exp} and identity loss \mathcal{L}_{id} . \mathcal{L}_{id} encourages the E-Adapter which integrates expression control, and the IP-Adapter to generate similar portraits, thereby minimizing the impact of expression control on identity fidelity. Additionally, \mathcal{L}_{exp} ensures that fine-tuning does not compromise the existing expression control capability of the Refined E-Adapter.

Both losses rely on the denoised clean image \hat{x}_0 . Contrary to standard diffusion inference which obtains clean images in multiple steps, we achieve approximately reconstructed images through one-step decoding. Specifically, we propose the Adaptive Noise Inversion (ANI) to directly

acquire \hat{x}_0 at any given timestep t , denoted by the formula:

$$\begin{aligned}\hat{x}_0 &= \text{Decode}(z_0 + f(t) \cdot (\epsilon - \epsilon_\theta(z_t, t, C))), \\ f(t) &= \begin{cases} \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} & \text{if } t \leq R_{tmax}, \\ f(R_{tmax}) & \text{if } t > R_{tmax}, \end{cases}\end{aligned}\quad (6)$$

where "Decode" is the vae decode function, and R_{tmax} is a constant. When $t > R_{tmax}$, we set $f(t)$ to $f(R_{tmax})$ to prevent reconstructed sample \hat{x}_0 from becoming too noisy, which compromises the accurate calculation of losses. Derivation details can be found in *supplementary materia*.

Building on ANI, we employ the IP-Adapter to generate portrait image \hat{x}_0^{id} , and the currently fine-tuning E-Adapter to generate the image with expressions \hat{x}_0^{cur} . \mathcal{L}_{id} minimizes the identity difference between two images, is defined as:

$$\mathcal{L}_{id} = 1 - \cos(\phi(\hat{x}_0^{id}), \phi(\hat{x}_0^{cur})), \quad (7)$$

where \cos is cosine similarity, and $\phi(\cdot)$ is the pre-trained face embedding extractor [9]. Similarly, we utilize the frozen Refined E-Adapter to generate image \hat{x}_0^{exp} . The expression loss \mathcal{L}_{exp} reduces the difference in facial landmarks between \hat{x}_0^{exp} and \hat{x}_0^{cur} , is defined as follows:

$$\mathcal{L}_{exp} = \text{Wing}(\varphi(\hat{x}_0^{exp}), \varphi(\hat{x}_0^{cur})),$$

where "Wing" denotes the wing loss [12], and $\varphi(\cdot)$ is the pre-trained facial landmarks detector [41].

The loss during fine-tuning is formulated as below:

$$\mathcal{L}_{ft} = \mathcal{L}_{sd} + w_{id} \cdot \mathcal{L}_{id} + w_{exp} \cdot \mathcal{L}_{exp}, \quad (8)$$

where w_{id} and w_{exp} are the balanced hyper-parameters.

4. Experiments

We first give an overview of the dataset, implementation details, baselines, and benchmarks used in the experiments. Then, we present the experimental results on different base models, followed by an ablation study of each module.

Dataset. We first collect 10k indoor and outdoor images of people of different countries and genders, each with 3-12 expressions. Then, we use LIQE [50] to filter out low-quality images and crop out 512×512 face square images. Subsequently, this dataset is employed to train the Base E-Adapter based on SD1.5 [37]. Afterward, we construct about 120k cross-identity triplets through IDI. Through expression consistency filtering, we retain about 100k triplets for training and fine-tuning the Refined E-Adapter. For evaluation, we gather images of 50 celebrities from diverse fields, with 30 different expressions.

Implementation Details. We train the Base E-Adapter on SD1.5 to construct data and the Refined E-Adapter on both SD1.5 [37] and SDXL [35]. In the Base E-Adapter training stage, our model is trained from scratch for 0.5

days. In SD1.5, we train the Refined E-Adapter for 1 day and fine-tune it for 4 hours. In SDXL, the input images are super-resolution using SwinIR [26] to a resolution of 1024×1024 , and we train the Refined E-Adapter for 2 days and fine-tun it for 0.5 days. We select Up 0.1. as the expression layers in SDXL. We do not observe style leakage on SD1.5, so we choose all its layers as expression layers.

Baselines. We compare our model with prior methods including OpenPose [49], MediaPipe [29], Arc2Face [32], and FineFace [42], using the same configurations (base model, text prompt, inference steps, etc.). Although portrait animation methods are fundamentally different from our approach, we assess these methods, including FaceAdapter [18], X-Portrait [45], and LivePortrait [16], to thoroughly evaluate our method. For the diffusion-based methods, we integrate IP-Adapter [47] plugin for identity preservation. For animation-based methods, an expressionless image of the portrait is generated, and then animation is applied. We select LivePortrait [16] to construct data triples \mathcal{T} and train the E-Adapter, which serves as our strong baseline.

Benchmarks. To ensure a fair comparison, all experiments are conducted using the SD1.5 framework across three styles: realistic, anime, and ink painting. Given the identity image, expression reference, and generated image, we evaluate all methods using ID fidelity (ID), image quality (IQ), expression similarity (Exp.), and facial landmarks movements similarity (LMS). Different from the models used in the ID encoder, the Antelopev2 [1] model is utilized to extract identity embeddings from both the portrait and generated image, with the cosine similarity between them used to evaluate identity fidelity. For the assessment of image quality, the pre-trained network LIQE [50] is employed. Expression similarity is measured by computing the L1 difference between facial blendshapes extracted from the expression and generated image using MediaPipe [29]. To evaluate the keypoint similarity, crucial facial landmarks (such as eyes, pupils, and mouth) are detected with Mediapipe to calculate movement amplitude differences. More metric details are provided in *supplementary materia*.

4.1. Comparison

Quantitative results. As summarized in Tab. 1, our method consistently outperforms all competitors by a good margin. Compared to the same type of diffusion conditioning methods, our approach shows significant improvements in ID similarity, image quality, and expression control ability. Even evaluating against portrait animation methods such as FaceAdapter [18] and LivePortrait [16], the proposed method offers superior ID fidelity and more accurate expression control. Notably, X-Portrait, which directly uses RGB expression images, also demonstrates strong expression control. However, X-Portrait shows a noticeable drop in image quality when applied to non-realistic styles, with



Figure 6. **Qualitative Comparisons.** Compared to other methods, our method demonstrates the most accurate and robust transfer of subtle facial expressions (*e.g.*, pouting, single-eye blinks, and pupil movements) while preserving the identity of the source portrait, even when applied in artistic styles (such as anime and ink painting).

Method	Realistic				Anime				Painting			
	ID \uparrow	IQ \uparrow	Exp. \downarrow	LMS \downarrow	ID \uparrow	IQ \uparrow	Exp. \downarrow	LMS \downarrow	ID \uparrow	IQ \uparrow	Exp. \downarrow	LMS \downarrow
Portrait Animation												
FaceAdapter [18]	0.309	4.991	0.059	0.281	0.095	4.853	0.109	0.467	0.146	4.747	0.072	0.348
X-Portrait [45]	0.505	4.902	<u>0.058</u>	<u>0.266</u>	0.188	4.142	0.066	0.335	0.332	3.668	0.062	0.291
LivePortrait [16]	0.609	4.772	<u>0.058</u>	0.319	<u>0.262</u>	4.843	0.116	0.551	0.419	4.749	0.087	0.426
Diffusion Conditioning												
OpenPose [49]	0.628	4.981	0.097	0.447	0.162	4.741	0.142	0.698	0.326	4.484	0.121	0.507
MediaPipe [29]	0.623	4.985	0.104	0.472	0.087	4.713	0.149	0.671	0.287	4.476	0.130	0.524
Arc2Face [9]	0.618	4.946	0.116	0.537	0.097	4.630	0.163	0.713	0.146	4.349	0.141	0.609
FineFace [42]	0.532	4.870	0.120	0.576	0.090	4.423	0.156	0.700	0.297	3.813	0.142	0.624
EmojiDiff (Ours)	0.666	4.995	0.054	0.215	0.304	4.910	<u>0.095</u>	<u>0.359</u>	0.469	4.756	0.078	0.256

Table 1. Quantitative comparisons of our method with SOTA counterparts in various base models.

IQ scores of only 4.142 and 3.668 in the anime and painting styles, respectively. On the contrary, our method produces exquisite stylized images due to the seamless integration with the diffusion model.

Qualitative results. Fig. 6 illustrates the performance of various methods in ID and expression control across three distinct styles. These examples demonstrate that our model effectively conveys reference expressions, including

lip movements and eye gaze while preserving the identity information of the source portrait. The fourth and sixth examples in the figure highlight the ability of our method to capture subtle movements, including lower lip pursing and pouting. Compared to other approaches, our model consistently exhibits robust expression control across different styles.

IDI	ICA	SD1.5 [37]			SDXL [35]		
		ID \uparrow	Exp. \downarrow	LMS \downarrow	ID \uparrow	Exp. \downarrow	LMS \downarrow
		0.265	0.113	0.447	0.432	0.093	0.454
✓		0.295	0.087	0.355	0.465	0.072	0.338
✓	✓	0.304	0.095	0.359	0.493	0.069	0.341

Table 2. Quantitative ablation about IDI and ICA.

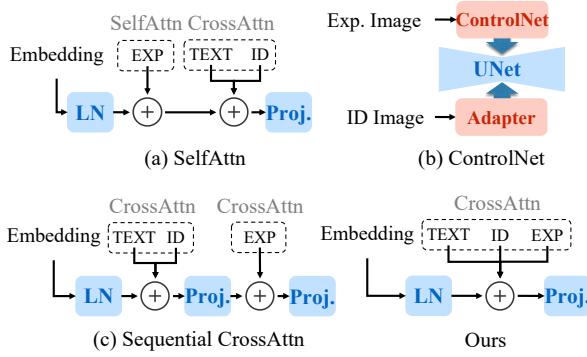


Figure 7. Different expression controller structure.

4.2. Ablation Study

We evaluate the proposed IDI and ICA on SD1.5 and SDXL. As presented in Tab. 2, in comparison to our baseline, IDI substantially improves expression control, reducing the Exp. score of the two base models by 23.0% and 22.6%, respectively. Additionally, ICA further enhances ID fidelity without compromising expression control, increasing the corresponding ID score by 3% and 6%, respectively.

ID-irrelevant Data Iteration. As shown in Tab. 3, we assess the performance of E-Adapter at different data iteration stages. Initially, the Base E-Adapter accurately transfers expressions to generated images compared to the baseline, achieving significant reductions in Exp. and LMS scores. Additionally, using the Base E-Adapter for cross-identity data construction, the newly trained Refined E-Adapter substantially reduces ID leakage, improving the ID score by 0.136. Finally, through further face-swapping, the Refined E-Adapter significantly outperforms the baseline.

Expression Controller Architecture. As illustrated in Fig. 7, we experiment with various expression controller structures. Expression embeddings are injected into the diffusion model using one of the following strategies: (a) self-attention, (b) ControlNet, (c) serial cross-attention with ID embeddings, and (d) parallel cross-attention with ID embeddings. As indicated in Tab. 4, the schemes (a) and (c), which operate without synchronous integration with ID embeddings, are insufficient in both ID fidelity and expression control capability. While the ControlNet achieves high ID fidelity, it lacks satisfactory expression control. Our approach combines identity and expression embed-

Base E-Adapter	Restruct Image	FaceSwap Image	ID \uparrow	Exp. \downarrow	LMS \downarrow
			0.265	0.113	0.447
✓			0.127	0.065	0.283
✓	✓		0.263	0.080	0.362
✓	✓	✓	0.295	0.087	0.355

Table 3. Ablation study of IDI.

Architecture	ID \uparrow	Exp. \downarrow	LMS \downarrow
(a) SelfAttn	0.266	0.109	0.497
(b) ControlNet	0.294	0.121	0.552
(c) Sequential CrossAttn	0.276	0.104	0.435
Ours	0.295	0.087	0.355

Table 4. Ablation study of expression controller architecture.



Figure 8. Ablation study of expression layers. Unlike our method, integrating expression information into either the face layers or the entire U-Net layers results in a degeneration in both image style and quality.

dings within a unified cross-attention block, effectively fostering their interaction and delivering optimal performance.

Expression Layer. As illustrated in Fig. 8, applying expression embeddings across all layers in highly stylized models significantly deteriorates the quality, and restricting expression feature injection to only the face layers does not improve it. By confining the injection of expression embeddings to the expression layers, we effectively address this issue. Notably, compared to employing expression embeddings across all layers or solely on face layers, our approach does not impair the expression control capability.

5. Conclusion

We introduce EmojiDiff, an innovative solution that utilizes both portrait images and RGB expression images to generate expressive portraits. Initially, we propose ID-irrelevant Data Iteration to synthesize high-quality cross-identity data pairs for training, effectively decoupling expression and

identity information in expression images. Additionally, style leakage from expression signals is prevented by injecting expression image features exclusively into expression-sensitive layers. To further harmonize identity and expression control, we propose ID-enhanced Contrast Alignment to mitigate the impact of expression modulation on identity fidelity. Experimental results demonstrate that our approach achieves accurate and nuanced facial expression control while maintaining precise identity resemblance.

References

- [1] Antelopev2. <https://huggingface.co/immich-app/antelopev2>, 2024. 6
- [2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2
- [3] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 2, 1
- [4] Zhiyuan Chen, Jiajiong Cao, Zhiqian Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2
- [5] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. *arXiv preprint arXiv:2403.05139*, 2024. 1
- [6] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 1
- [7] CrucibleAI. Controlnetmediapipeface, 2023. 1, 2
- [8] Stefano d’Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, and Luc Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021. 2
- [9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 2, 6, 7
- [10] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2
- [11] FaceFusion. https://modelscope.cn/models/ic/cv_unet-image-face-fusion_damo, 2023. 3, 4, 5
- [12] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018. 6
- [13] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 2
- [14] 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
- [15] Google. Blendshapev2. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker, 2022. 5, 1
- [16] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 4, 6, 7
- [17] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 1, 2
- [18] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. *arXiv preprint arXiv:2405.12970*, 2024. 2, 6, 7
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1
- [20] 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. 1, 2
- [21] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2
- [22] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024. 1, 2
- [23] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6080–6090, 2023. 1, 2
- [24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4
- [25] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1, 2
- [26] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1833–1844, 2021. 6
- [27] Renshuai Liu, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In *Proceedings of the IEEE Con-*

- ference on Computer Vision and Pattern Recognition, pages 2114–2123, 2024. 2
- [28] Yang Liu, Cheng Yu, Lei Shang, Yongyi He, Ziheng Wu, Xingjun Wang, Chao Xu, Haoyu Xie, Weida Wang, Yuze Zhao, et al. Facechain: A playground for human-centric artificial intelligence generated content. *arXiv preprint arXiv:2308.14256*, 2023. 1, 2
- [29] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris Mc-Clanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 5, 6, 7, 1
- [30] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 2
- [31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1
- [32] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision*, 2024. 1, 2, 6
- [33] Reni Paskaleva, Mykyta Holubakha, Andela Ilic, Saman Mottamed, Luc Van Gool, and Danda Paudel. A unified and interpretable emotion representation and expression generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2447–2456, 2024. 1, 2
- [34] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024. 1
- [35] 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, 8, 4
- [36] 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. 1
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 6, 8, 3
- [38] 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. 4
- [39] 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 Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 2
- [40] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinghui Tang. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*, 2024. 1
- [41] Siyang Song, Yuxin Song, Cheng Luo, Zhiyuan Song, Selim Kuzucu, Xi Jia, Zhijiang Guo, Weicheng Xie, Linlin Shen, and Hatice Gunes. Gratis: Deep learning graph representation with task-specific topology and multi-dimensional edge features. *arXiv preprint arXiv:2211.12482*, 2022. 6
- [42] Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024. 1, 2, 6, 7
- [43] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2
- [44] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animations. *arXiv preprint arXiv:2403.17694*, 2024. 1
- [45] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4, 6, 7
- [46] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024. 3, 4
- [47] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 4, 5, 6
- [48] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023. 2
- [49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 6, 7
- [50] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 6, 1

EmojiDiff: Advanced Facial Expression Control with High Identity Preservation in Portrait Generation

Supplementary Material

In this supplementary material, we provide more metrics details in Sec. A, more implementation details in Sec. B, additional experimental results in Sec. C.

A. More Metrics Details

Exp. We employ MediaPipe [29] to extract 52 facial blendshapes [15] of the expression reference and the generated image, represented as b^{ref} , b^{gen} , respectively. Each blendshape b_i^{gen} in the $[0, 1]$ range indicates the probability of the relevant facial action occurring (*e.g.*, eyeBlinkLeft, mouthClose). The expression difference between reference and generated image can be formulated as:

$$\mathbb{E}_{exp} = \sum_{i=0}^{51} (|b_i^{ref} - b_i^{gen}|) \quad (1)$$

LMS To further assess differences in the expression of critical facial regions (such as eyes, pupils, and mouth), we propose evaluating the landmark movement similarity (LMS) between the reference and generated image. First, we utilize MediaPipe [29] to detect 478 facial landmarks of the expression reference and generated image, denoted as l^{ref} and l^{gen} , respectively. Next, we select representative landmarks to calculate the movement amplitude of key facial actions, including blinking, eye movement, and mouth opening, represented as:

$$\begin{aligned} r_{leye} &= \frac{\|l_{145} - l_{159}\|_2}{\max(10^{-5}, \|l_{133} - l_{33}\|_2)}, \\ r_{lpupil} &= \frac{\|l_{133} - l_{468}\|_2}{\max(10^{-5}, \|l_{133} - l_{33}\|_2)}, \\ r_{reye} &= \frac{\|l_{374} - l_{386}\|_2}{\max(10^{-5}, \|l_{263} - l_{362}\|_2)}, \\ r_{rpupil} &= \frac{\|l_{263} - l_{473}\|_2}{\max(10^{-5}, \|l_{263} - l_{362}\|_2)}, \\ r_{mouth} &= \frac{\|l_{17} - l_0\|_2}{\max(10^{-5}, \|l_{291} - l_{61}\|_2)}, \end{aligned} \quad (2)$$

The LMS metric quantifies the difference in movement amplitude between two images, expressed as:

$$\mathbb{E}_{LMS} = \sum_{k \in S} (|r_k^{ref} - r_k^{gen}|), \quad (3)$$

where $S = \{\text{leye, reye, lpupil, rpupil, mouth}\}$.

B. More Implementation Details

B.1. Detailed Implementation

We gather images of over 10,000 people displaying various expressions from video clips and in-house face databases, removing low-quality images using LIQE [50]. Then, we employ the proposed IDI to construct cross-identity, same-expression datasets. As mentioned in Sec. 3.3 of the main body, facial blendshapes [15] and landmark [29] differences are utilized to filter out expression-changed data during data construction. Specifically, we calculate the Exp. and LMS metrics of the original expression image and the synthesized new image, excluding the data with Exp. ≤ 0.05 and LMS ≤ 0.18 . Based on synthesized data, we train the Refined E-Adapter for SD1.5 at a resolution of 512×512 . On SDXL, we upscale the data to a resolution of 1024×1024 and train it with random scaling following IP-Adapter [47]. The learning rates of the Adam optimizer during training and fine-tuning are set to 2×10^{-5} and 5×10^{-6} , respectively, with $\beta_1 = 0.5$, and $\beta_2 = 0.999$.

In the E-Adapter, we employ CLIP [36] as the encoder for the expression branch to extract expression signals. The image-prompt and FaceID versions of IP-Adapter [47] are utilized to initialize the projection layers of the expression and identity branches, emphasizing structure and identity features, respectively. During training and fine-tuning, the parameters of the expression branch and the projection layer of the identity branch are updated, while other parameters remain fixed.

In all experiments, λ_{id} , λ_{exp} , R_{tmax} , w_{id} , and w_{exp} are consistently set to 1, 1, 600, 0.08, and 10, respectively. For expression loss \mathcal{L}_{exp} , landmarks of the facial contour are excluded from the loss calculation.

B.2. Derivation of Adaptive Noise Inversion

During the diffusion process, the noisy representation z_t is derived from the original latent representation z_0 and added noise ϵ , represented as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} * \epsilon, \quad (4)$$

where t , α_t represent timestep and a predefined function of t , respectively. When the noise perturbation is small, the noise $\epsilon_\theta(z_t, t, C)$ predicted by U-Net approximately equals to the added noise ϵ [3, 34]. During the denoising process, we can approximately reconstruct the original latent z_0 by

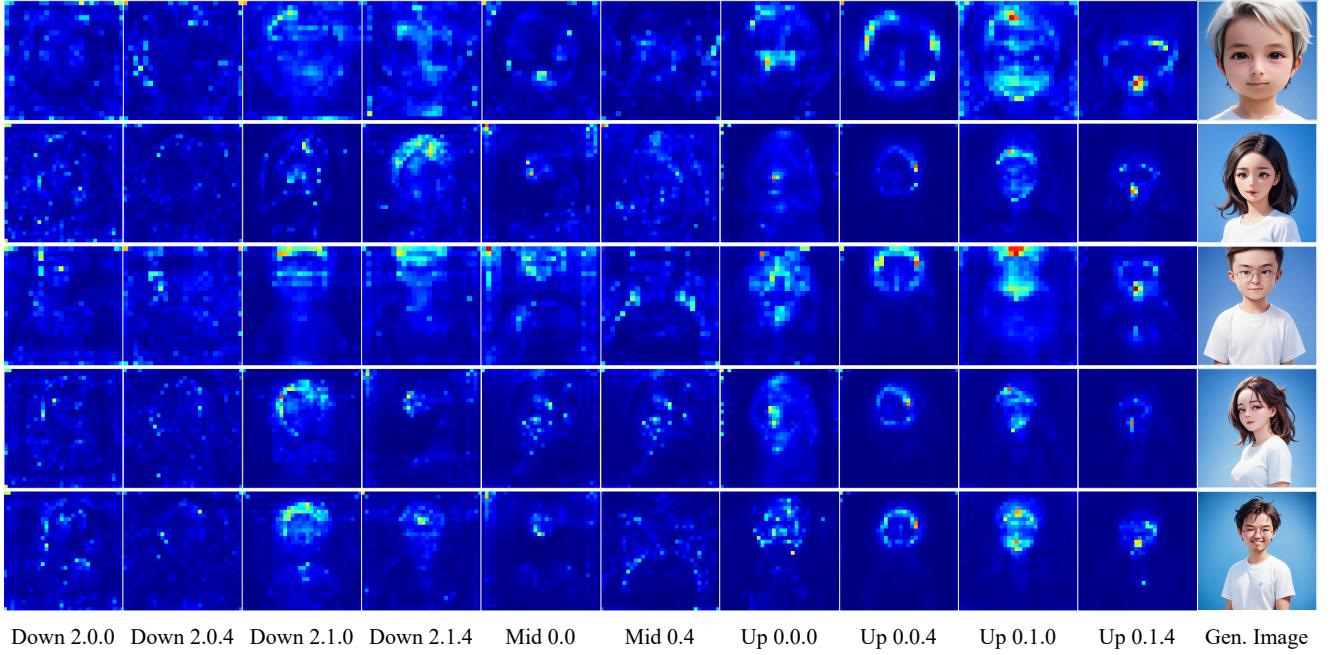


Figure 1. **Visualization of attention maps.** Down.2.0.0, Mid.0.0 represents down_blocks.2.attentions.0.transformer_blocks.0.attn2, mid_blocks.attentions.0.transformer_blocks.0.attn2 layers in the U-Net, respectively.

performing single-step sampling, denoted as:

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} * \epsilon_\theta(z_t, t, C)}{\sqrt{\bar{\alpha}_t}}, \quad (5)$$

By combining the Eq. (4) and Eq. (5), the reconstructed latent \hat{z}_0 can also be expressed as:

$$\hat{z}_0 = z_0 + \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}} * (\epsilon - \epsilon_\theta(z_t, t, C)) \quad (6)$$

As depicted in Eq. (6), the reconstructed latent \hat{z}_0 can be directly derived from the origin latent z_0 and the difference between the added noise ϵ and predicted noise ϵ_θ , and the signal-to-noise ratio is decided by $\sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}}$. The reconstructed sample \hat{x}_0 can be obtained by decoding \hat{z}_0 , formulated as:

$$\hat{x}_0 = \text{Decode} \left(z_0 + \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}} * (\epsilon - \epsilon_\theta(z_t, t, C)) \right), \quad (7)$$

where "Decode" refers to the vae decode function. Notably, in the higher noisy stage (large timestep), the reconstructed sample \hat{x}_0 may become noisy, resulting in inaccurate identity and expression loss calculations. To overcome this, we propose the Adaptive Noise Inversion (ANI), defined as:

$$\begin{aligned} \hat{x}_0 &= \text{Decode}(z_0 + f(t) * (\epsilon - \epsilon_\theta(z_t, t, C))), \\ f(t) &= \begin{cases} \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}} & \text{if } t \leq R_{tmax}, \\ f(R_{tmax}) & \text{if } t > R_{tmax}, \end{cases} \end{aligned} \quad (8)$$

where R_{tmax} is the predefined constant. Building on Eq. (6), ANI directly truncates the model's predictions based on the timestep t . When t exceeds R_{tmax} , $f(t)$ is set to $f(R_{tmax})$ to prevent the reconstructed \hat{x}_0 from being noisy.

C. Additional Experimental Results

C.1. Attention Map Visualization

As shown in Fig. 1, we present additional visualizations of attention maps. Consistent with Fig. 4 in the main text, the face layers exhibit a clear response in the facial area. At the same time, the expression layers (e.g., Up 0.1.0, Up 0.1.4) focus on various facial regions, enabling comprehensive expression control across all face areas.

C.2. More Visualization Results

In Fig. 6 of the main body, we only display a few generated images due to the page limit. In this subsection, we present additional results generated by our method on more styles and expressions. As shown in Fig. 2 and 3, our method accurately transfers the reference expression to the generated image while maintaining high identity fidelity.

C.3. Data Iteration Visualization

In Fig. 5 of the main body, we show examples illustrating the effect of IDI in modifying the identities of individuals while maintaining facial expressions. As a supplement, we provide more visualization results as depicted in Fig. 4, clearly demonstrating the effectiveness of our method.



Figure 2. More qualitative results. For the given person (leftmost column), our method generates the corresponding image based on the various expression references, evaluated on SD1.5 [37] framework.

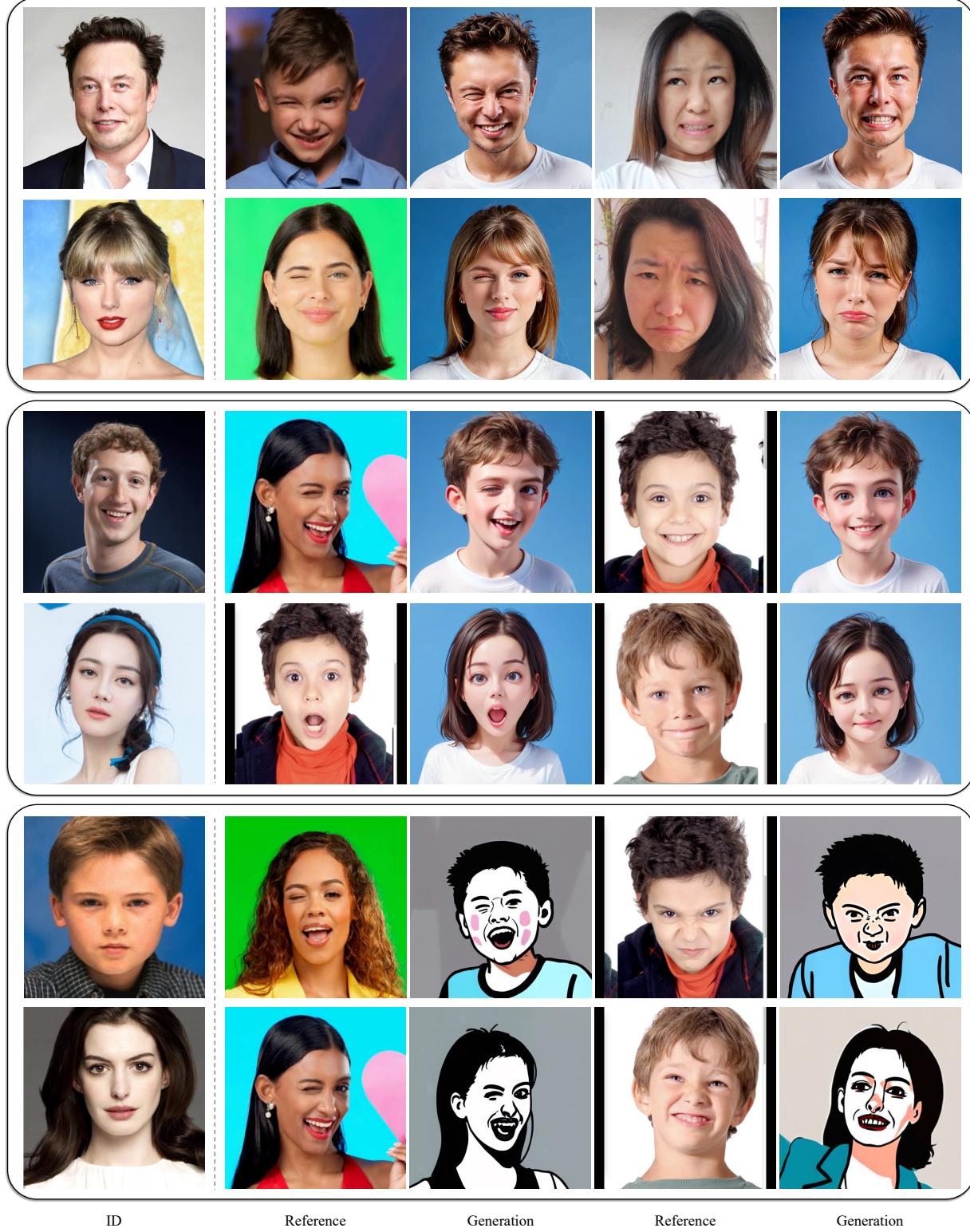


Figure 3. **More qualitative result.** For the given person (leftmost column), our method generates the corresponding image based on the various expression references, evaluated on SDXL [35] framework.

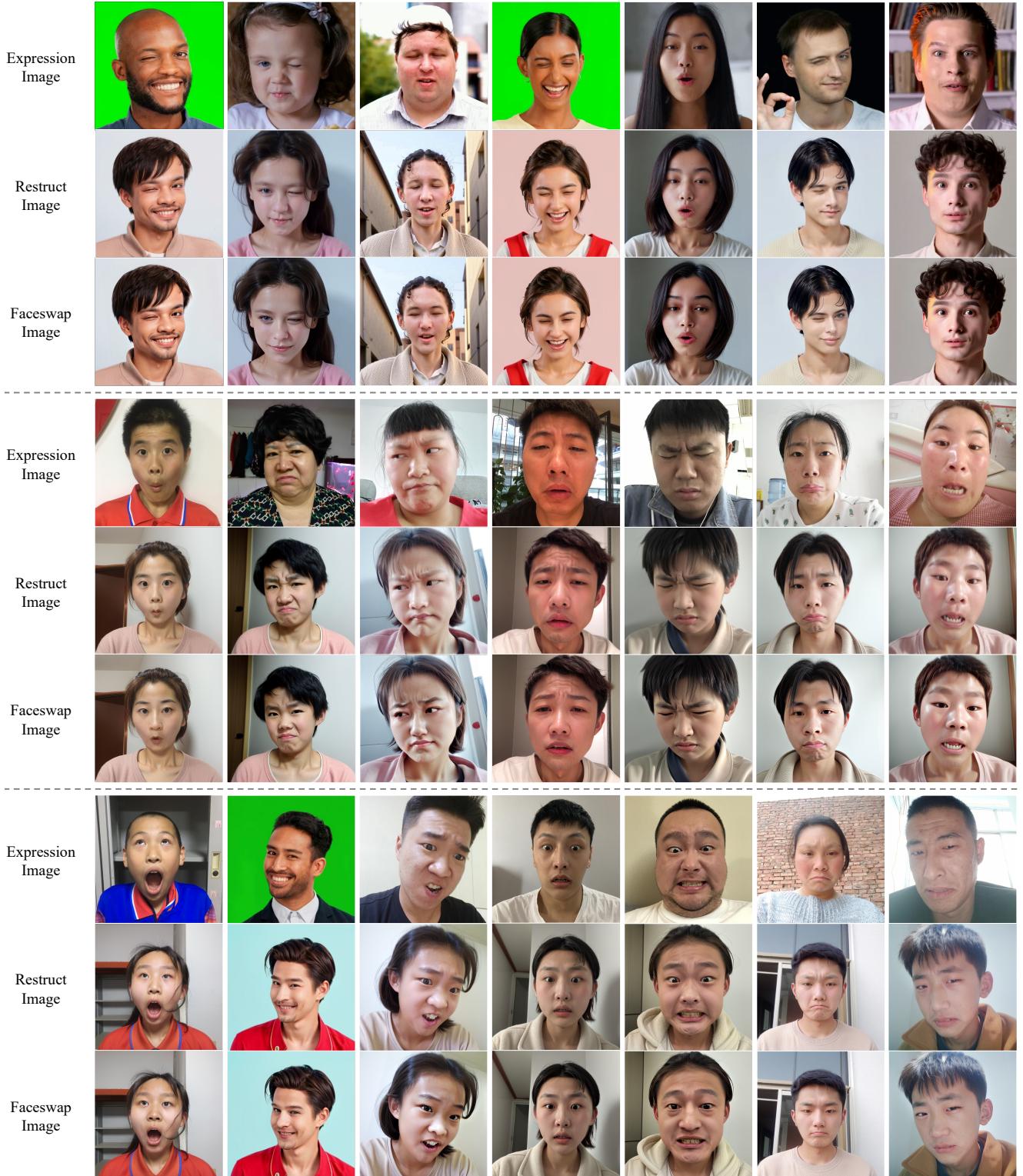


Figure 4. **More examples of Stage II.** ID-irrelevant Data Iteration (IDI) is introduced to transform expression reference images into the reconstruct and faceswap images, thus synthesizing high-quality data pairs with differing identities and consistent expressions.