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

Supplementary Material

In this supplementary material, we provide more experimental details in Sec. A, detailed derivation of Adaptive Noise Inversion (ANI) in Sec. B, additional experimental results in Sec. C.

A. More Experimental Details

A.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 [13]. Then, we employ the proposed IDI to construct cross-identity, same-expression datasets. As mentioned in Sec. 3.2.2 of the main body, facial blendshapes [2] and landmark [5] 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. $\leq$0.05 and LMS $\leq$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 [12]. The learning rates of the Adam optimizer during training and fine-tuning are set to $2\times\ 10^{-5}$ and 5×10^{-6}, respectively, with β_1 = 0.5, and β_2 = 0.999. 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.

In the E-Adapter, we employ CLIP [10] as the encoder for the expression branch to extract expression signals. The image-prompt and FaceID versions of IP-Adapter [12] 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.

A.2. Ablation Experiments Details

In Sec. 4.3 of the main body, we experiment with various expression controller structures. As is shown in the Fig. 1, 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.

A.3. Metric Details

Exp. We employ MediaPipe [5] to extract 52 facial blendshapes [2] 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:

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

Method	ID ↑	Exp. ↓	LMS ↓
DaGAN [3]	0.202	0.135	0.569
Follow-U-Emoji [6]	0.243	0.125	0.428
AvatarArtist [4]	0.239	0.110	0.454
SkyReels-A1 [9]	0.277	0.104	0.387
Ours	**0.304**	**0.095**	**0.359**

Table 1. Comparisons with post-processing animation methods.

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 [5] 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:

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

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

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

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

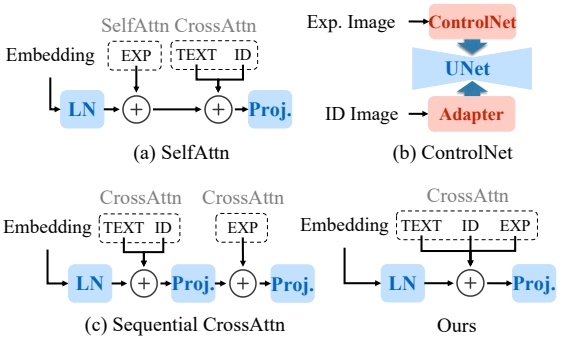

Figure 1. Different expression controller structure.

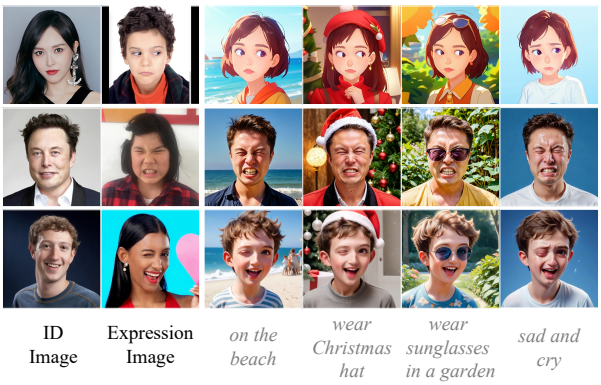

Figure 2. Images generated using different prompts.

B. 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:

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

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 [1, 7]. During the denoising process, we can approximately reconstruct the original latent z_0 by performing single-step sampling, denoted as:

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

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

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

As depicted in Eq. (6), the reconstructed latent 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:

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

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:

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

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. More Quantitative Comparisons

As shown in the Tab. 1, we provide comparisons with more methods (*e.g.*, GAN-based, animation-based, and avatar-based). Benefiting from RGB-level expression input and simultaneous processing with identity information, our method significantly surpasses other methods.

C.2. Combined with Text Prompts

As depicted in Fig. 2, we show how to combine image control and text control to generate images. Actually, prompts can be customized to meet diverse requirements while expressions are stably governed by the expression template. Moreover, an appropriate prompt will further enhance the vividness of expressions (last column).

C.3. More Visualization Results

In Fig. 5 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 **SD1.5** [11] and **SDXL** [8] framework. As shown in Fig. 3 and 4, our method accurately transfers the reference expression to the generated image while maintaining high identity fidelity.

C.4. Data Visualization

In Fig. 4 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. 5, clearly demonstrating the effectiveness of our method and data quality of CIEP100k.

Figure 3. **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** [11] framework.

Figure 4. **More qualitative result**. For the given person (leftmost column), our method generates the corresponding image based on the various expression references, evaluated on **SDXL** [8] framework.

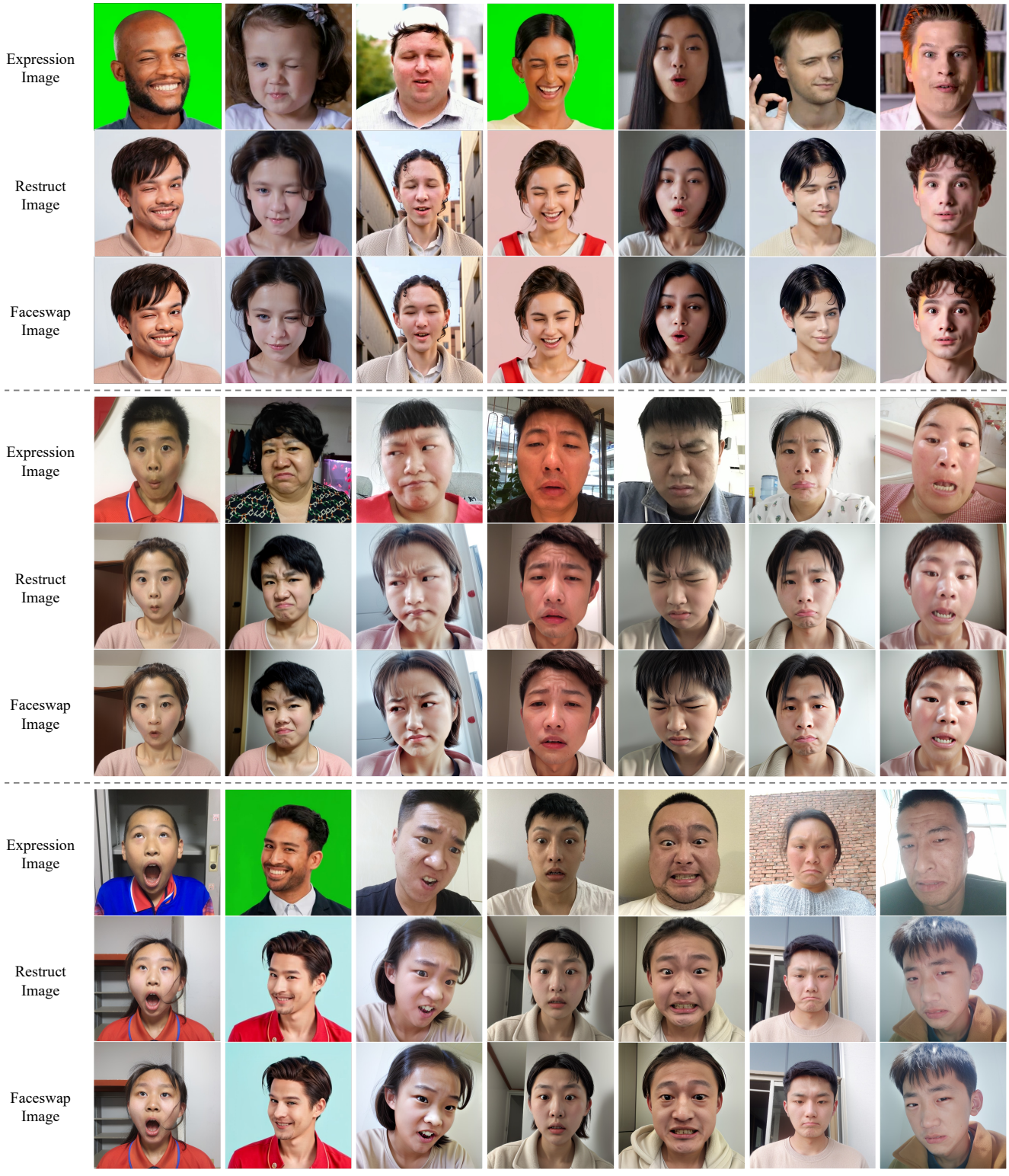

Figure 5. **More examples of IDI and CIEP100k**. ID-irrelevant Data Iteration (IDI) is introduced to transform expression reference images into the restruct and faceswap images, thus synthesizing high-quality data pairs with differing identities and consistent expressions.

References

[1] 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

[2] Google. Blendshapev2. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker, 2022. 1

[3] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[4] Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. Avatarartist: Open-domain 4d avatarization. *arXiv preprint arXiv:2503.19906*, 2025. 1

[5] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, 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. 1

[6] 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. 1

[7] 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. 2

[8] 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. 2, 4

[9] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 1

[10] 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

[11] 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. 2, 3

[12] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[13] 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. 1