1. Appendix

1.1. Detail process of SAA

A diagram of the process of SAA in multi-ID personalization is provided in Fig. 1.

1.2. VariFace-10k dataset details

The training of our IMR necessitates a comprehensive personalized dataset, where each identity is characterized by a diverse collection of facial images exhibiting a wide spectrum of expressions, orientations, and other attributes, complemented by corresponding textual prompts. This comprehensive dataset is crucial for advancing the model's understanding of the disentanglement and entanglement between identity and motion features within the feature space. However, currently, available high-quality facial datasets, including FFHQ [5], SFHQ [1], and CelebA [6], demonstrate significant limitations in terms of intra-individual image diversity, typically constrained to a narrow range of expressions (predominantly neutral and happy) and even representing individuals with only a single image. To overcome these limitations, we have developed the VariFace-10k dataset, which contains 35 distinct facial images per individual, each exhibiting substantial variations across multiple dimensions. This dataset serves as a fundamental resource for training our IMR and addresses the existing gap in personalized dataset availability. Our dataset construction process involved initially curating high-quality facial images from the FFHQ dataset, subsequently augmenting this collection with additional high-quality images sourced from the internet and further expanding the dataset through GAN-based generation of supplementary high-quality facial images. All images were standardized through uniform cropping to 512x512 resolution, resulting in a foundational set of 10k distinct facial images. Building upon this foundation, we employed the Face-Adapter [4] to perform face reenactment using images from the KDEF dataset as driving images, ultimately generating an extensive collection of 350k facial images comprising 10k unique identities, each represented by 35 distinct facial attributes. Recognizing the potential for facial distortion in generating profile views from frontal images, we processed each set of 35 images per individual through the IP-Adapter-FaceID-Portrait model [8] to regenerate 35 refined images. Finally, we implemented [3] for landmark generation and utilized [2] to provide detailed, fine-grained textual prompts for each facial image in our dataset.

1.3. More evaluation details

For quantitative analysis, we randomly selected 500 identities from the CelebA dataset to construct our test set, adhering to the methodology outlined in [7]. We employed 20 prompts encompassing various accessories, clothing, backgrounds, actions, and styles. Table 2 provides the complete list of prompts. Each base prompt was systematically augmented through the injection of supplementary facial attributes, including seven distinct facial expressions (neutral, happy, angry, disgusted, surprised, sad, afraid) and four orientation descriptors (front view, side view, facing up, facing down). Single-ID prompts are structured as: a person with a *happy* expression *in a side view*, *wearing headphones*. Multi-ID prompts are: The person on the left has a neutral expression in a side view, and the person on the right has a *sad* expression *in a front view*, both *wearing headphones*. For the *Pose* metric, if the source prompt includes "facing up" or "facing down," we utilize the pitch angle with a threshold of 10 degrees to categorize the images into 'up,' 'down,' or 'front' classes. Conversely, if the source prompt contains "in front view" or "in side view," we employ the yaw angle with a threshold of 10 degrees to classify the images into 'side' or 'front' categories.

1.4. Detailed implementation of LDC term

The **L**atent **D**iffusion **C**onsistency term can be more precisely expressed as:

$$||\epsilon_\theta(z_t, t, \xi_{pred}, \tau) - \epsilon_\theta(z_t, t, \xi_{tgt}, \tau)||_2^2 \quad (1)$$

where z_t is derived from the target facial image, and τ is the text embedding corresponding to the facial prompt associated with the target image. Within this framework, the Latent Diffusion Consistency term ensures semantic equivalence in the T2I model's latent space, ensuring that the predicted features induce generative behaviors similar to those of the target features.

1.5. More ablation

Effect of parameters α and β on multi-ID personalized generation: We conducted an in-depth investigation into the effects of parameters α and β on multi-ID personalized generation, using the probability of detecting valid faces across all target regions as the evaluation metric. As illustrated in Fig. 2, the experimental results led us to select $\alpha = 0.24$ and $\beta = 2$ as the optimal values.

Effect of the two Terms in the IMR training stage: As evidenced in Table 1, both the Direct Feature Matching term and the Latent Diffusion Consistency term play pivotal roles in attaining flexible facial editability while maintaining high identity preservation. Our ablation study demonstrates that the elimination of the Latent Diffusion Consistency term substantially impairs facial editing capability, whereas the removal of the Direct Feature Matching term significantly compromises identity fidelity. These empirical findings underscore the complementary nature and synergistic interplay of these two terms in achieving optimal performance in the personalized generation.

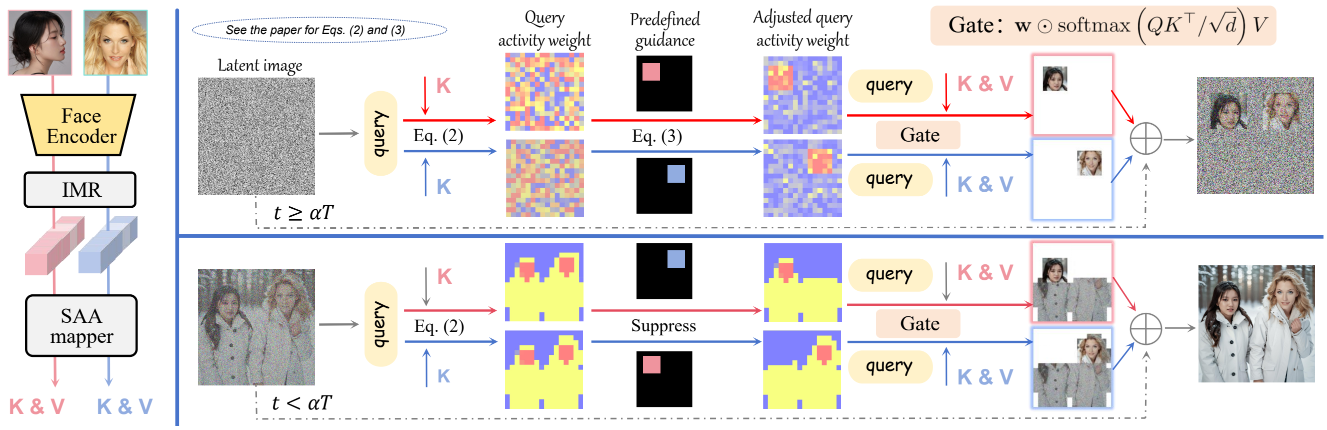

Figure 1. Detail process of SAA in multi-ID personalization.

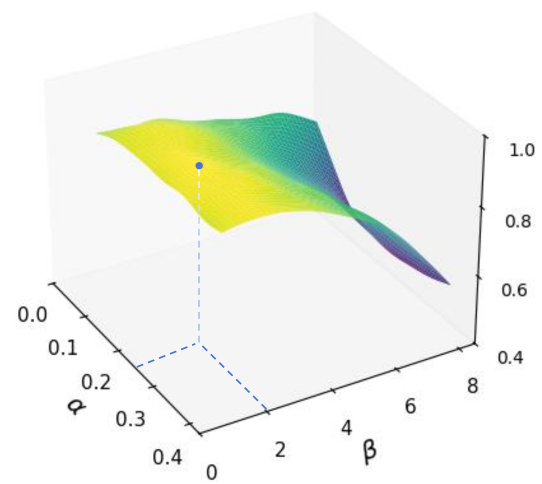

Figure 2. Investigating the impact of varying hyperparameter values of α and β for multi-ID personalized generation, our experimental results led us to select $\alpha = 0.24$ and $\beta = 2$.

Method	**CLIP-T**	**FaceSim**	**Expr**	**Pose**
Ours w/o DFM	0.237	0.663	0.433	0.851
Ours w/o LDC	0.238	0.667	0.234	0.644
Ours	**0.239**	**0.671**	**0.456**	**0.878**

Table 1. The proposed **D**irect **F**eature **M**atching term and **L**atent **D**iffusion **C**onsistency term significantly enhance flexible facial editability and maintain identity fidelity.

1.6. More Applications

We provide more applications of our DynamicID, encompassing context decoupling (Fig. 3), layout control (Fig. 4), complex expression editing (Fig. 6), ID mixing (Fig. 5), and transformation from non-photo-realistic domains to photo-realistic ones (Fig. 7).

Category	Prompt
Accessory	wearing headphones with long yellow hair
Clothing	wearing a spacesuit in a chef outfit in a doctor's outfit in a police outfit
Background	standing in front of a lake in the mountains on the street in the snow in the desert on the sofa on the beach
Action	reading books walking on the road playing the guitars holding a bottle of red wine eating lunch
Style	a painting in the style of Ghibli anime a painting in the style of watercolor

Table 2. Evaluation text prompts are categorized into Clothing, Accessories, Background, Action, and Style, which will be incorporated as part of the final prompt.

References

[1] David Beniaguev. Synthetic faces high quality (sfhq) dataset, 2022. 1

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep

face recognition. In *CVPR*, 2019. 1

[4] 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. 1

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1

[7] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, 2024. 1

[8] 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

Figure 3. The application of context decoupling.

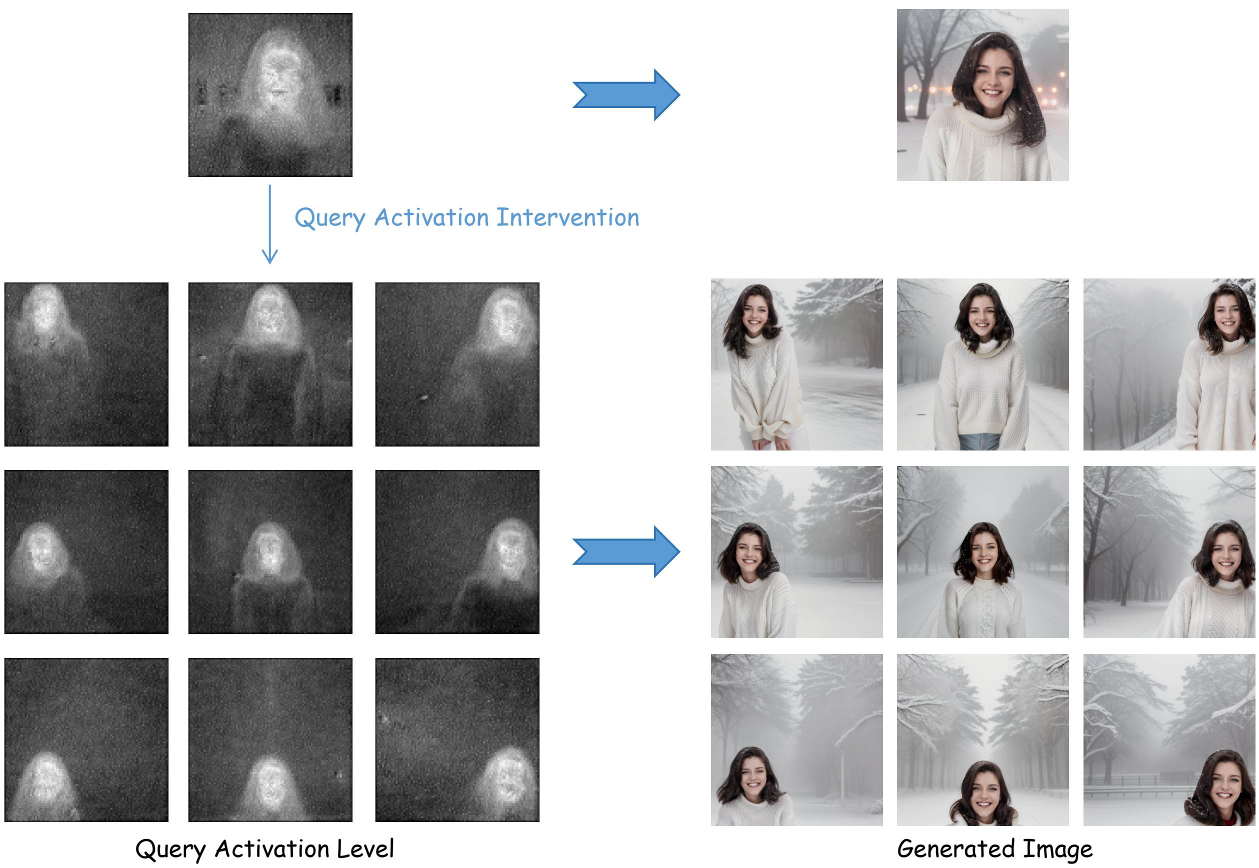

Figure 4. The application of layout control.

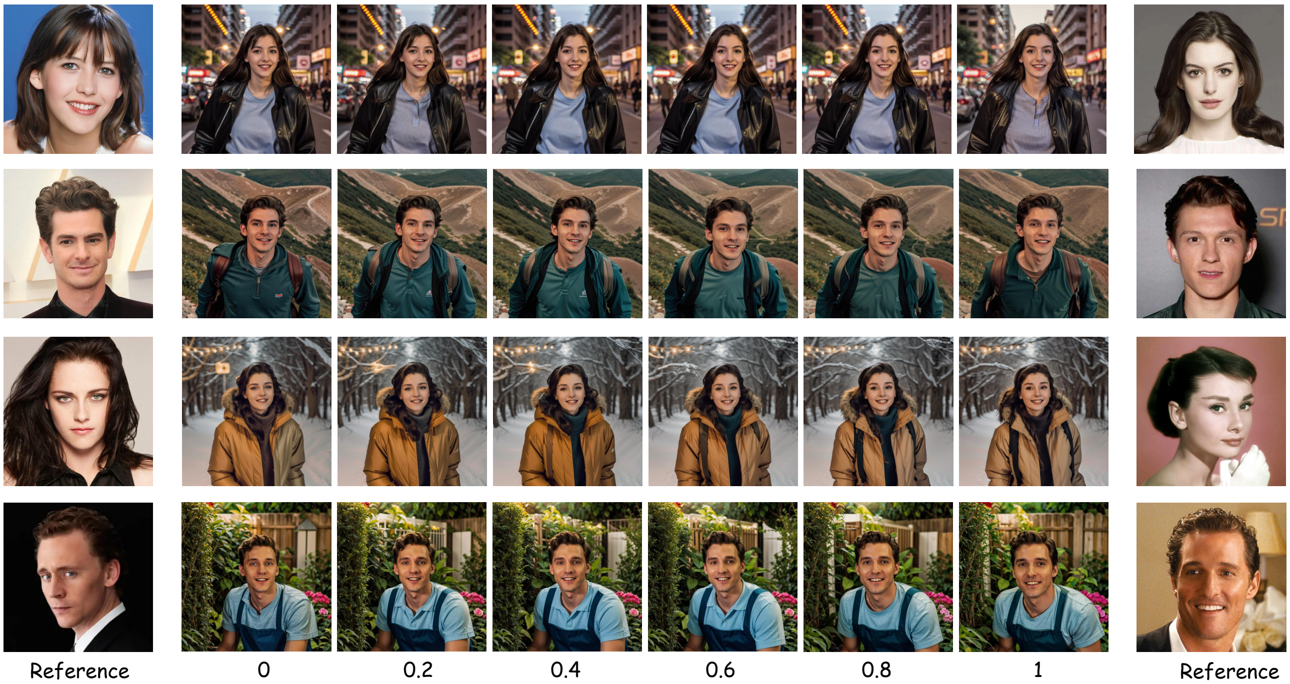

Figure 5. The application of ID mixing.

Figure 6. The application of complex expression editing. Zoom in for a better view.

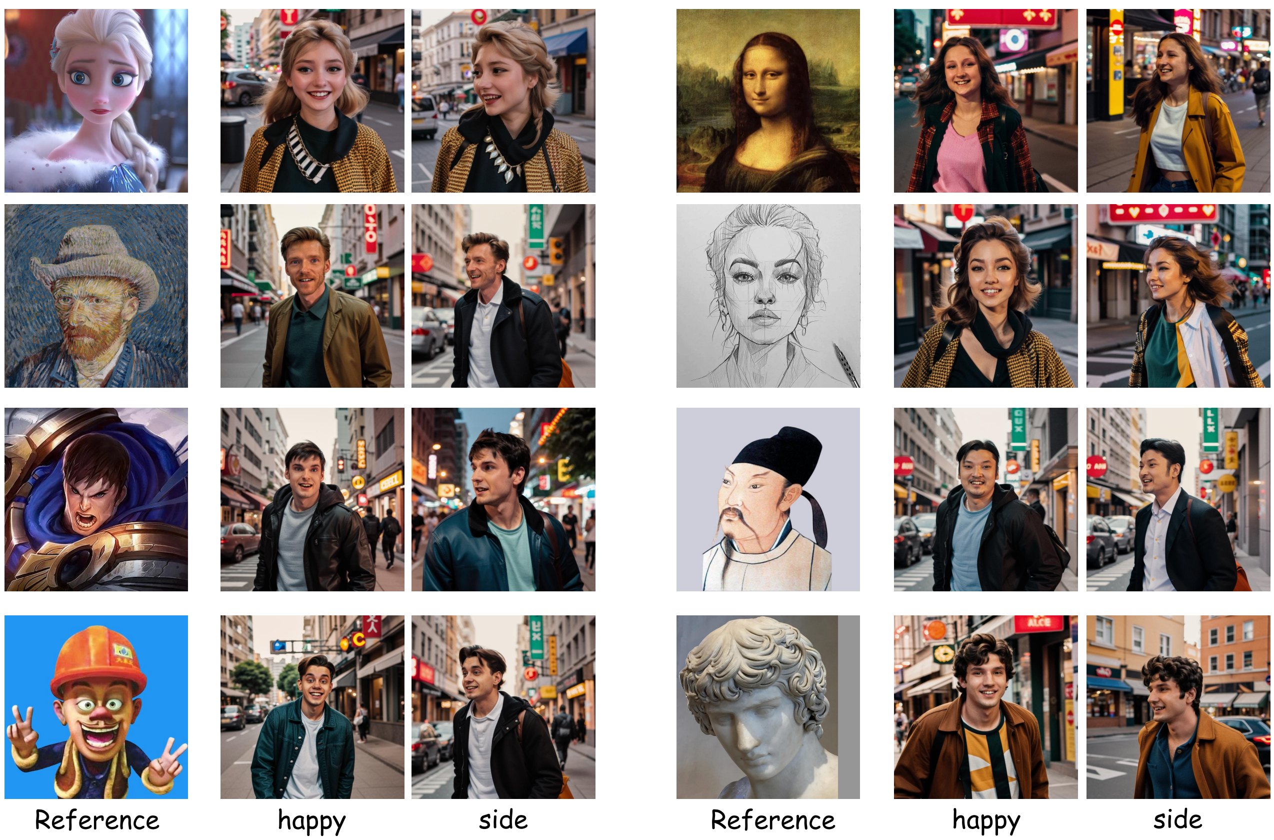

Figure 7. The application of transformation from non-photo-realistic domains to photo-realistic ones.