Ano-Face-Fair: Race-Fair Face Anonymization in Text-to-Image Synthesis using Simple Preference Optimization in Diffusion Model

Yeon Gyu Han[1,2] Seung Han Song[3] Dongheon Lee[1,4,5 *]
[1] MODULABS [2] Department of Biomedical Engineering, Chungnam National University
[3] Department of Plastic and Reconstructive Surgery, Chungnam National University Hospital
[4] Department of Radiology, Seoul National University College of Medicine
[5] Department of Radiology, Seoul National University Hospital
dusrb37@gmail.com silverwinebag@gmail.com dhlee.jubilee@gmail.com

Abstract

*Face anonymization requires effectively hiding identities while preserving essential features, yet existing models often show racial bias, particularly in representing Asian faces. We propose "**Ano-Face-Fair**" an approach for race-fair face anonymization based on Stable Diffusion-v2 Inpainting with three key contributions: (1) Focused Feature Enhancement (FFE) loss $\mathcal{L}_{FFE}$, for detailed facial feature generation across diverse racial groups, (2) Difference (DIFF) loss $\mathcal{L}_{DIFF}$, to prevent catastrophic forgetting by maintaining distinct racial characteristics, and (3) Simple Preference Optimization (SimPO) for enhanced synthetic image consistency. Our method enables flexible control through both mask and text-based prompting, achieving robust anonymization while maintaining high image quality and accuracy in Asian face generation. We validate our method's effectiveness through extensive experiments on facial image generation across diverse racial groups. This work advances face anonymization by addressing racial biases in image generation, demonstrating robust and realistic face editing across diverse racial groups through mask and text-based prompting, thus contributing to more ethical generative model.*
Code: https://github.com/i3n7g3/Ano-Face-Fair

1. Introduction

Facial images contain detailed personal information, requiring a trade-off in the face anonymization process, as identities must be effectively anonymized while minimizing data loss. Traditional image processing for facial anonymizing identities typically involves applying masks to facial regions or blurring features; however, these methods often result in significant loss of facial detail. Recently, generative models for face editing have enabled the preservation of certain parts of the face while simultaneously synthesizing the rest to appear realistic. For example, parts of the face, such as the eyes or nose, remain unchanged, while the rest is synthesized using inpainting methods.

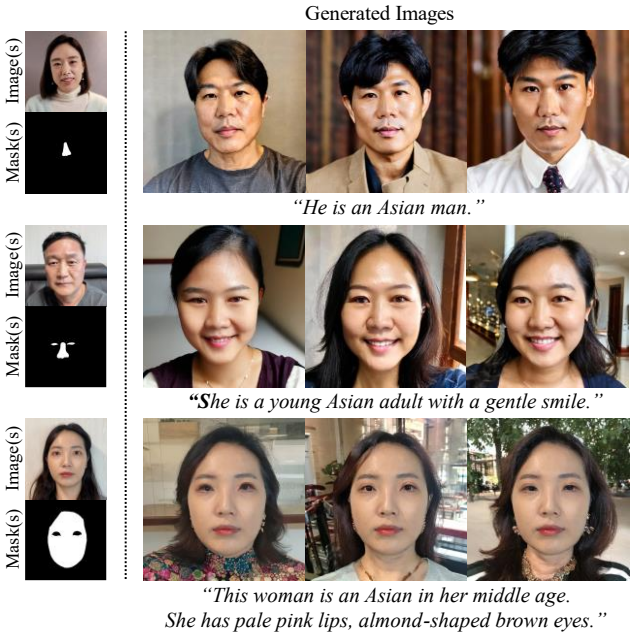


Figure 1: ***Ano-Face-Fair*** represents results of anonymized Asian faces under multiple conditions. The third row shows results generated from mask-based and text-based prompts.

In essence, the key to facial anonymization is achieving **minimal data loss** within facial images while simultaneously enabling the synthesis of **natural-looking faces**, which can be utilized in various fields, such as social or medical domains. This facial preservation allows users and researchers to conduct meaningful analyses without compromising individual privacy. For instance, our method in the medical domain enables the anonymization of preserving disease-specific features while protecting patient identity, thereby achieving a balancing between privacy and clinical utility.

* Corresponding Author.

However, although face generation and editing can be used for anonymization, issues related to AI fairness still remain. Much of the work on face synthesis and editing has relied on public datasets, such as CelebA [1] and FFHQ [2], which include a limited range of Asian faces. Consequently, these models fail to accurately represent Asian facial features, such as generally smaller eyes, among other characteristics.

Therefore, our model, **Ano-Face-Fair**, is designed to achieve both goals. First, it protects privacy by keeping key facial regions intact while editing other parts as needed based on prompts. We used the Stable Diffusion v2 Inpainting [3] model as the baseline for mask-based and text-based prompting to guide the specific regions of the face and set conditions for face editing. To recognize each mask-prompt within the face automatically, we trained the PointRend [4] for facial segmentation using own facial dataset. The results of the segmentation model are used as a mask-prompt to determine which parts to edit, helping to keep certain facial regions unchanged and thereby ensuring anonymity.

Second, we proposed two types of loss functions to ensure the model performs well, even with a small dataset of Asian images for training. We trained the baseline model using the DreamBooth [5] and applied Focused Feature Enhancement (FFE) loss, $\mathcal{L}_{FFE}$, which is designed for precise training across a variety of facial features. This method works particularly well in inpainting applications, helping the model focus on important features, including detailed parts of the face. The $\mathcal{L}_{FFE}$ complements the existing DreamBooth loss by refining details. It is implemented through two key components: a critical feature mask that focuses on regions of high importance for training, and an FFE loss weight that adjusts the impact of $\mathcal{L}_{FFE}$ on the overall loss function.

Next, we proposed a Difference loss, $\mathcal{L}_{DIFF}$, to prevent catastrophic forgetting [6]. When we applied the proposed $\mathcal{L}_{FFE}$, we observed that the model began to forget information about certain racial representations, such as Caucasians, which it had previously synthesized well. By using $\mathcal{L}_{DIFF}$ during the fine-tuning phase with an Asian image dataset, we ensure that the model retains the trained information of specific contrast (preservation) classes, thereby recognizing and preserving the differences between the class being fine-tuned and the contrast classes.

Lastly, this study has improved the face editing model using Simple Preference Optimization (SimPO) [7], a technique that aligns with user preferences. SimPO enhances computational efficiency by eliminating the need for a separate reward model, instead using a simple preference function to compare two generated samples. This approach enables the model to reflect user preferences more efficiently and in a more personalized manner. By applying SimPO to a diffusion model, we can train on complex data distributions through denoising, thereby better capturing user-preferred image styles and details. While Direct Preference Optimization (DPO) [8] has recently made significant contributions in language models, our comparison of DPO and SimPO showed that SimPO is more suitable for our image-based task.

Our approach to face editing, which employs masks and text prompts, consistently generates robust results across different races, as demonstrated in Figure 1. The main contributions of our paper are summarized as follows:

- We employed the Stable Diffusion v2 Inpainting model as our baseline and for implementing mask-based and text-based prompting. We trained a facial segmentation model on our own dataset for mask-prompting to automatically designate fixed regions of the face, and we controlled the generated facial areas using text-prompting.
- We trained the model using a curated Asian dataset with the proposed two types of loss functions: (1) Focused Feature Enhancement (FFE) loss, $\mathcal{L}_{FFE}$, which effectively reproduces typical Asian facial features. Additionally, we enhanced the training process by incorporating a (2) Difference loss, $\mathcal{L}_{DIFF}$, to preserve the quality of generation for other ethnicities where the model had previously shown good performance.
- By using Simple Preference Optimization (SimPO) on diffusion models, we trained the model to achieve more consistent synthetic images. Using this SimPO in text-to-image tasks, particularly for inpainting, significantly improved the quality of generated images, especially for Asian faces, thereby enhancing the model's performance in terms of race fairness.

2. Related Work

Face Anonymization. In traditional image processing, facial anonymization is mainly categorized into three main methods: pixelation and blurring [9], masking [9], and morphological transformations [10]. Firstly, pixelation and blurring [9] is common techniques, but they often anonymize images, reducing their utility. This heavy editing can obscure important details, making the images less valuable for analysis. Next, Masking [10], which conceals facial regions such as the eyes or nose using opaque overlays, can effectively obscure identities but faces challenges in preserving finer details, thereby complicating customization. Lastly, morphological transformations [10], which change the size or position of facial regions, require manual adjustments. This requirement significantly slows down the standardization of anonymization across different images. On the other hand, generative model-based approaches to facial anonymization [11-14] have emerged due to their ability to produce a wide variety of results. However, accurately training on complex facial features remains challenging, often resulting in generative outputs that lack fidelity. To

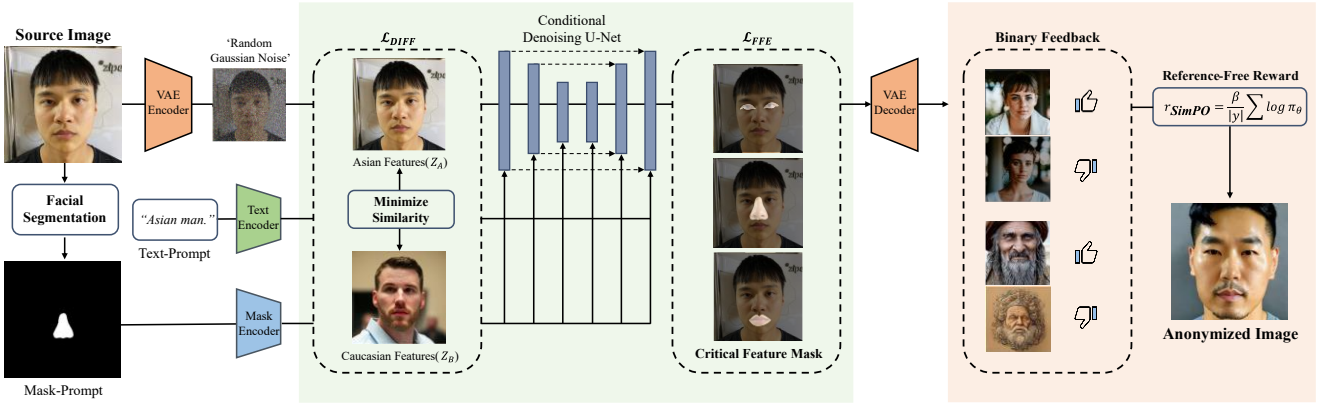


Figure 2: The overview of ***Ano-Face-Fair***. The facial mask from the facial segmentation model was utilized to prompt edits in specific facial regions. The Stable Diffusion v2 Inpainting model used as our baseline, trained on the curated Asian dataset. We applied $\mathcal{L}_{FFE}$ to enhance performance even with limited data and used $\mathcal{L}_{DIFF}$ to address the catastrophic forgetting issue of the pre-trained model. Finally, we employed a model trained with Simple Preference Optimization (**SimPO**), using LoRA Adapter for efficient fine-tuning, to generate more refined and enhanced images.

address this issue, this study presents a method to overcome the fidelity limitations of current generative models.

Face Generation and Editing. Previous studies on facial generation have employed GAN-based methods such as StyleGAN [2], which improve facial synthesis quality and enable style control within an interpretable latent space. Additionally, methods like VQGAN [15], which transform input images into a lowdimensional feature space for codebook training, have also been utilized. More recently, diffusion-based models [4, 16, 17] have been applied to text-to-image generation, producing high quality results. However, they encounter challenges in achieving detailed control over generated images and face issues with fairness across different races. In face editing, various methods have been proposed, including GAN-based image inpainting [18, 19], text-to-image editing [20-22], and conditional facial generation [23, 24]. Recent advancements have introduced more versatile methods using diffusion models, which are categorized into text-based [20-22] and mask-based [27, 28] approaches. While these methods aim to preserve original identity and achieve precise editing, they often lack support multi-condition and seamless blending with surroundings. Textual inversion [29] and DreamBooth have attempted to fine-tune models for new classes or concepts, but they often result in biased outputs and catastrophic forgetting. In contrast, our proposed diffusion-based model supports multi-condition and addresses the forgetting issue, thereby enhancing image editing capabilities.

Preference Optimization in Diffusion Models. Aligning diffusion models with human preferences has been less explored compared to language models. Initially, supervised fine-tuning using curated datasets [27, 30] was the primary approach. The introduction of Reinforcement Learning from Human Feedback (RLHF) [31-33] marked a significant advancement, introducing reward models to fine-tune diffusion models via policy gradient techniques. Notable examples include ReFL [34], DRaFT [35], and AlignProp [36], which specifically align text-to-image diffusion models to human preferences. However, these methods often encounter challenges such as memory constraints and potential biases from the learned reward models. Diffusion-DPO [37] adapted Direct Preference Optimization to diffusion models, eliminating the need for reward models unlike previous RLHF methods. However, its reliance on reference models can limit flexibility, especially when there is a significant difference between the characteristics of the preference data and those of the reference model. Recent trends have shifted towards reference model-free approaches, such as Margin-aware Preference Optimization (MaPO) [38]. MaPO eliminates the need for reference models, offering advantages like memory efficiency and faster training, but it introduces complexity in margin calculations and requires careful hyperparameter tuning. In contrast, our Simple Preference Optimization (SimPO) for diffusion models provides a simple optimization process without the need for reward models, reference models, complex calculations, while maintaining efficiency.

3. Method

In this section, we first explain how we prepare semantic facial mask to control multiple condition. Next, we discuss the fine-tuning of the diffusion model using the proposed Focused Feature Enhancement Loss $\mathcal{L}_{FFE}$ and Difference Loss $\mathcal{L}_{DIFF}$. Finally, we detail the optimized of the proposed model for user preferences.

3.1. Facial Segmentation

The facial segmentation model used is PointRend [4],

which employed an efficient point-based approach for high-quality semantic segmentation. PointRend adaptively selects points to render, focusing on challenging regions that require fine detail. We employed the model and trained it on 17,697 facial images from A hospital.

The trained model precisely segments facial features such as the eyes, nose, lips, and facial contours. Its performance was evaluated using the mean Intersection over Union (mIoU) metric, achieving a score of 0.92 on the test set. These accurate facial region segmentations are then used as detailed mask-prompts for the baseline model's input, allowing precise control over areas to edit or preserve during anonymization.

3.2. Fine-tuning Stable Diffusion Models

Focused Feature Enhancement (FFE) Loss. We propose a novel loss function called Focused Feature Enhancement (FFE) loss $\mathcal{L}_{FFE}$, designed to enhance Asian facial features often underrepresented in standard datasets. We fine-tuned the Stable Diffusion v2 Inpainting model using both instance loss $\mathcal{L}_i$ and prior-preserving loss $\mathcal{L}_{pp}$ from DreamBooth.

While $\mathcal{L}_i$ focuses on training set features and $\mathcal{L}_{pp}$ preserves pre-trained knowledge, using only these losses often results in overfitting and fails to capture detailed features. Unlike traditional gradient-based approaches that naturally emphasize regions with large errors, our $\mathcal{L}_{FFE}$ provides more precise control over feature enhancement through a dynamic critical feature mask.

The computation of $\mathcal{L}_{FFE}$ begins with an error mapping function E_x that quantifies the pixel-wise differences between predicted output x_p and target image x_t:

$$E_x = |x_p - x_t| \tag{1}$$

where x_p represents the model's current prediction and x_t is the ground truth target image. Based on this error map, we generate a critical feature mask, M_c that identifies regions requiring focused enhancement:

$$M_c = (x_p - x_t) > \theta * MAX\left((x_p - x_t)\right) \tag{2}$$

where θ is a threshold parameter that determines which regions receive enhanced attention. Regions with differences larger than θ times the maximum error are set to 1, indicating areas requiring focused enhancement.

The final $\mathcal{L}_{FFE}$ is formulated as:

$$\mathcal{L}_{FFE} = \lambda \frac{1}{N} \sum_{i=1}^{N} M_c^i (O^i - T^i)^2 \tag{3}$$

where O^i represents the feature representations at position i in the output image (extracted from intermediate layers of our model), T^i denotes the corresponding target features, M_c^i modulating their contribution based on the importance of each region's feature enhancement needs. Based on initial analysis, we set θ = 0.5 and λ in the range of 0.01 to 0.1 to achieve optimal performance.

Difference Loss. In our experiments, applying $\mathcal{L}_{FFE}$ led to the observation that details about 'Caucasians', which were pre-trained in the baseline model, were being forgotten. This issue, similar to the language drift problem observed in language models, arose when fine-tuning a layer based on text embeddings. We also noted that this forgetting occurred more frequently with longer training epochs.

Traditional contrastive learning loss, $\mathcal{L}_c$ aims to bring similar subjects closer together in the embedding space while pushing dissimilar samples apart. However, this approach does not specifically address the catastrophic forgetting of racial features during fine-tuning. To address this limitation, we propose the Difference Loss $\mathcal{L}_{DIFF}$, which extends $\mathcal{L}_c$ by strategically maximizing the angular separation between different racial class embeddings:

$$\mathcal{L}_{DIFF} = \frac{1}{|P|} \sum_{(i,j)\in P} \left(1 - \frac{z_i \cdot z_j}{|z_i||z_j|}\right) \tag{4}$$

Where z_i and z_j are embedding vectors from different racial classes, $z_i \cdot z_j$ represents their dot product, and $|z_i|, |z_j|$ are their respective magnitudes. P represents the set of all embedding vector pairs from different classes. By minimizing this loss, we maximize the angle between embeddings of different races, preserving distinct racial characteristics through clear boundaries.

To prevent embedding space collapse during optimization, we introduce constraints:

$$||z_i|| = ||z_j|| = 1, \qquad \theta_{ij} \geq \theta_{min} \tag{5}$$

where unit normalization ensures consistent scaling, and minimum angular separation (θ_{min}) maintains distinct representation of different racial characteristics, preventing ambiguous or mixed features.

Integration of $\mathcal{L}_{FFE}$ and $\mathcal{L}_{DIFF}$. These two losses form a complementary optimization framework: $\mathcal{L}_{FFE}$ enhances local facial details, ensuring precise retention of race-specific features, while $\mathcal{L}_{DIFF}$ maintains global distribution separation in the embedding space, preventing overfitting to any particular racial group.

We integrate these losses into a balanced total objective that combines instance-level supervision with distribution-level constraints:

$$\mathcal{L}_{Total} = \mathcal{L}_i + \lambda_{pp} * \mathcal{L}_{pp} + \lambda_{FFE} * \mathcal{L}_{FFE} + \mathcal{L}_{DIFF} \tag{6}$$

This formulation allows our model to simultaneously improve Asian facial feature generation while preserving high-quality generation for other ethnicities, as demonstrated in Figure 3.

Figure 3: The results of applying the proposed $\mathcal{L}_{FFE}$ and $\mathcal{L}_{DIFF}$. (*Top*) When only $\mathcal{L}_{FFE}$ was applied, the generation quality improved for the Asian race, but it declined for other races (e.g., Caucasian), which had previously shown well with the pre-trained model. (*Bottom*) When both $\mathcal{L}_{FFE}$ and $\mathcal{L}_{DIFF}$ were applied, the quality was consistently maintained across all races.

Simple Preference Optimization for Diffusion model. We propose adapting Simple Preference Optimization (SimPO) to diffusion models for text-to-face editing tasks. SimPO builds upon Direct Preference Optimization (DPO) while addressing its limitations, particularly the need for a reference model and high computational costs.

Unlike DPO, which requires computing gradients through a reference model creating substantial overhead, SimPO uses a reference model-free reward formulation:

$$r_{SimPO}(x, y) = \frac{\beta}{|y|} \sum log\, \pi_\theta\,(y_i \mid x, y_{<i}) \qquad (7)$$

where π_θ represents the policy of the model (the noise prediction network), x is the input, y is the output, and β is a scaling factor. This directly utilizes the model's own likelihood estimates as rewards.

To quantify preferences between outputs, we extend this formulation with a margin parameter γ:

$$logits = r\_SimPO(x, y_w) - r\_SimPO(x, y_l) - \gamma \qquad (8)$$

where y_w and y_l represent preferred and non-preferred outputs respectively. These logits determine probabilistic preference judgments through a Bradley-Terry objective.

For adaptation to diffusion models, which operate in continuous latent space through iterative denoising (unlike discrete language models), we compute preference logits directly from noise prediction losses. The final SimPO loss is expressed as:

$$\mathcal{L}_{SimPO} = -log\, \sigma\left(\beta(\mathcal{L}_l - \mathcal{L}_w - \gamma)\right) \qquad (9)$$

where $\mathcal{L}_l$ and $\mathcal{L}_w$ represent noise prediction losses for less and more preferred outputs.

We implement SimPO using Low-Rank Adaptation (LoRA), enabling efficient fine-tuning by modifying only a small number of parameters. In face anonymization, this approach provides three key advantages: (1) Direct optimization of visual quality without reference constraints, (2) Consistent feature preservation through length normalization, and (3) Efficient training through simplified optimization.

4. Experiments

4.1. Experimental Settings

Baselines. We compare our model with DreamLike Inpainting [40], Stable Diffusion XL Inpainting [41], and Stable Diffusion v2 Inpainting (Baseline model) [42].

Datasets and Models. We demonstrate the effectiveness of our method using a combination of models and datasets. We employ PointRend [4] for facial segmentation, trained on 17,697 facial images from a hospital dataset (divided in 7:2:1 ratio for training, validation, and testing) with institutional review board (IRB) approval. Patient demographics, including age, sex, and history of plastic surgery, were assessed.

We fine-tuned our approach on two primary datasets: (1) an Asian Face Dataset of 4,515 images manually curated from AI Hub [43], selected from six public datasets where facial features are well-exposed: 'Face parsing', 'Facial images with known family relationships', 'Masked Korean facial image data', 'Facial recognition aging image data', 'Language-based image editing data', and 'Composite images for Korean emotion recognition', (2) the Pick-a-Pic dataset [44] for SimPO fine-tuning, comprising 583,747 training pairs with preference labels from 4,375 users. The model used in this study is Stable Diffusion v2 inpainting as the baseline.

For comprehensive evaluation across diverse face distributions, we additionally test on CelebA-HQ [1], containing 30,000 high-quality celebrity images, and FFHQ [2], consisting of 70,000 diverse facial images varying in age, ethnicity, and background.

Hyperparameters for Simple Preference Optimization. We use SimPO for training with a batch size of 8 and gradient accumulation steps of 2, implemented with LoRA (rank=4, alpha=4) requiring 24GB VRAM. Our hyperparameter search revealed critical relationships: (1) optimal learning rate scales inversely with β following $\boldsymbol{\alpha} = c/\boldsymbol{\beta}$ where $c \in$ [1e-3, 1e-2], (2) training destabilizes when $\boldsymbol{\gamma}/\boldsymbol{\beta}$ exceeds 1.5, and (3) gradient explosion occurs at $\boldsymbol{\beta}$ >2000 due to sigmoid function's exponential scaling. After extensive experimentation, we established our final configuration ($\boldsymbol{\beta}$=200, learning rate=1e-5, γ/β=0.5). More detailed results from this configuration are analyzed in our ablation study.

Prompting. There are two methods for conditioning in face editing. First, single or multiple masks can be applied within facial regions, such as the eyes and nose, allowing for the editing of surrounding areas. Second, specified mask regions can be generated using text-based prompts.

Evaluation Metrics. To evaluate face editing, we used FID, SSIM, LPIPS, CLIP-I, and PSNR metrics.

To evaluate anonymization, we used a face recognition model based on InsightFace [45]. This model compared the original images with the generated anonymized images by calculating the cosine distance between their facial embeddings. We defined the Anonymization Success Accuracy as the accuracy of generated images successfully anonymized, determined by a distance greater than a set threshold from the original image. The Average Distance quantifies the overall dissimilarity, while the Min and Max Distances provide the range of anonymization effectiveness.

To evaluate racial fairness, we employed the VGG-Face model [46] across all datasets. We analyzed the racial characteristics of generated images using various prompts related to different ethnicities. The racial accuracy metrics represent the model's ability to maintain appropriate racial characteristics while anonymizing faces. These metrics are crucial for assessing our model's race fairness capabilities across diverse datasets. We compared the baseline models with our proposed method and conducted a comprehensive evaluation, including an ablation study and multi-dataset analysis, to verify both performance and generalization capability.

4.2. Ablation Study of Ano-Face-Fair

In this study, the proposed fine-tuning methods and SimPO were applied, and their performance was compared through an ablation study. Table 1 presents both the racial feature preservation accuracy and image quality metrics across different configurations. While Configuration B shows improved Asian face generation, it exhibits some performance trade-offs for other racial features. Configuration C demonstrates $\mathcal{L}_{DIFF}$ effectively maintains performance across all ethnicities. Finally, Configuration D with SimPO achieves the highest overall performance in both feature preservation and image quality metrics.

Figure 4 illustrates the qualitative results of the ablation study. The effectiveness of our method across different races is further demonstrated through comprehensive visual examples in Figure 5 shows consistent high-quality results across diverse ethnic groups.

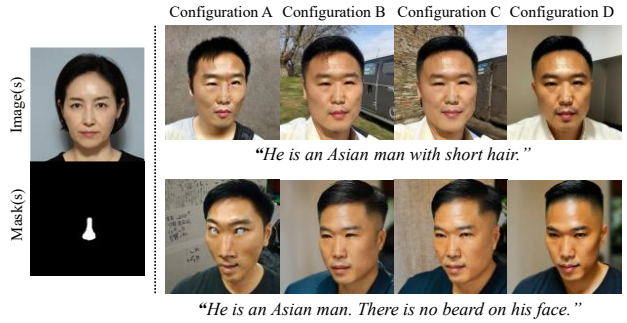


Figure 4: Qualitative results from the ablation study of ***Ano-Face-Fair***.

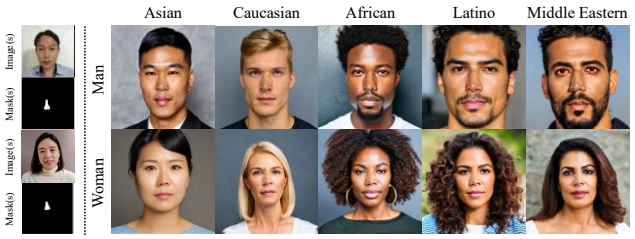


Figure 5: Qualitative results across diverse ethnic groups. The grid shows our method's performance across five racial groups (Asian, Caucasian, African, Latino, and Middle Eastern) for both genders. All images were generated with the prompt "A photo of a [ethnicity] [gender]."

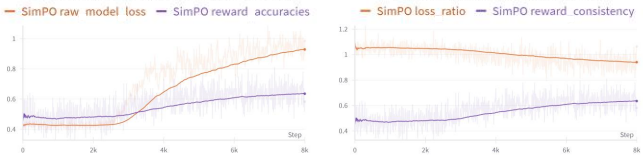


Figure 6: Training dynamics of the final model. The graphs show the progression of raw model loss, reward accuracy, loss ratio, and reward consistency during training.

Figure 6 demonstrates the training dynamics of our final model configuration (β=200, learning rate=1e-5, gamma ratio=0.5), showing balanced optimization across key metrics. Rather than aiming for extreme values, we achieve an optimal trade-off: raw model loss steadily decreases while reward accuracies remain stable (0.6-0.8), indicating effective preference learning without overfitting. The loss ratio and reward consistency stabilize after initial convergence, validating our balanced approach.

Configuration	Racial Feature Preservation Accuracy ↑					Anonymization Accuracy ↑	FID ↓	PSNR ↑
	Asian	Caucasian	African	Latino	Middle Eastern			
A. Baseline	0.862	0.949	0.921	0.901	0.912	1.00	111.5	25.41
B. A. + $\mathcal{L}_{FFE}$	0.958	0.912	0.889	0.878	0.882	1.00	108.783	26.53
C. B. + $\mathcal{L}_{DIFF}$	0.958	0.945	0.919	0.899	0.913	1.00	109.128	26.97
D. C. + SimPO	**0.9583**	**0.945**	**0.922**	**0.899**	**0.915**	**1.00**	**72.60**	**27.91**

Table 1: Ablation study of ***Ano-Face-Fair***.

Comparison of Preference Optimization Methods. We compared SimPO with DPO, SPO [47] and MAPO to evaluate its effectiveness. Table 2 presents the results of this comparison.

Models	PS ↑	HPS ↑	CLIP Aes ↑	IR ↑
DPO	0.5031	0.5031	9.7644	**0.5183**
SPO	0.5034	0.4955	10.199	0.4999
MAPO	0.4963	0.4904	10.157	0.503
SimPO	**0.5047**	**0.5107**	**10.851**	0.5043

Table 2: Comparison of preference optimization methods.

PS: PickScore, HPS: Human Preference Score, CLIP Aes: CLIP Aesthetics, IR: Image Reward

4.3. Comparison with the Baselines

The performance of the proposed method was compared with baseline models. Figure 7 illustrates the qualitative image synthesis results, comparing the proposed method with the baseline models. We also compared the image editing results generated by mask-based and text-based prompting, demonstrating diverse and realistic image generation outcomes across various combinations. These comparisons are illustrated in Figure 8.

Table 3 presents the average quantitative evaluation results, where our model demonstrated the highest performance across key metrics including anonymization accuracy, feature preservation, and image quality.

Figure 7: Qualitative comparison between baseline models.

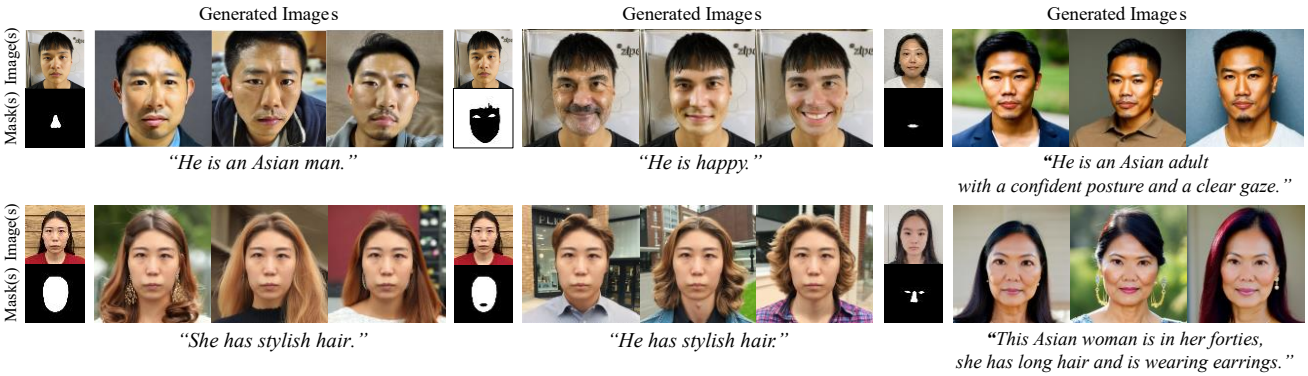


Figure 8: Quantitative results of the proposed method's facial editing using mask-based and text-based prompts. The results show how different combinations of mask-based and text-based prompts affect changes in facial features.

Models	Anonymization ↑				Asian Acc ↑	FID ↓	SSIM ↑	LPIPS ↓	CLIP-I ↑	PSNR ↑
	Acc	Avg. Dist.	Max. Dist.	Min. Dist.						
SD-v2-I	1.00	1.0061	1.1479	0.8509	0.8620	111.5	0.2004	0.7860	0.6521	25.41
DL-I	1.00	1.0029	1.1863	0.7952	0.8932	98.51	0.2910	0.7596	0.6781	27.89
SD-XL-I	1.00	1.0181	1.2045	0.8267	0.9494	75.03	0.2912	0.7596	0.6531	27.91
Ours	**1.00**	**1.0367**	**1.2294**	**0.8703**	**0.9583**	**72.60**	**0.3149**	**0.7341**	**0.7565**	**27.91**

Table 3: Quantitative comparison on the Asian facial dataset.

SD-v2-I: Stable Diffusion-v2 Inpainting, DL-I: DreamLike Inpainting, SD-XL-I: Stable Diffusion XL Inpainting

4.4. Multi dataset Evaluation

To verify the generalization capability of our approach across diverse face distributions, we applied our proposed method ($\mathcal{L}_{FFE}$+$\mathcal{L}_{DIFF}$+SimPO) to two standard benchmark datasets: CelebA-HQ [1] and FFHQ [2]. This evaluation allows us to assess whether our race-fair anonymization approach is effective beyond our Asian Face Dataset.

Following the same training protocol described in Section 4.2, we independently trained our method on Asian Dataset, CelebA-HQ and FFHQ. We maintained identical hyperparameters and loss configurations across all experiments to ensure fair comparison. We compared our trained models with the pretrained Stable Diffusion XL Inpainting model, which serves as a strong baseline. Tables 4-5 present the detailed results.

As shown in Table 4, our method demonstrates consistent race fairness improvements across all datasets. Most notably, our approach significantly enhances Asian facial feature preservation accuracy regardless of training data while maintaining excellent preservation of other racial characteristics. Table 5 shows that our method achieves optimal FID scores on the Asian Dataset, while maintaining competitive image quality on CelebA-HQ and FFHQ.

The consistent performance across these diverse datasets confirms that our method's race fairness improvements are not specific to our training distribution but generalize well to standard benchmark datasets. This demonstrates the robustness of our combined approach in addressing racial bias in face anonymization.

5. Conclusion

In this paper, we propose ***Ano-Face-Fair***, a race-fair text-to-face synthesis model, particularly effective for Asian face. The proposed method utilizes mask-based and text-based prompting to generate natural-looking faces while anonymizing them, regardless of race. Our approach introduces two types of loss functions to enhance the Stable Diffusion v2 Inpainting model: (1) the Focused Feature Enhancement (FFE) loss, designed to achieve high performance with a limited training set of Asian face images, and (2) the Difference (DIFF) loss, which prevents catastrophic forgetting across races. Additionally, we applied Simple Preference Optimization (SimPO) to diffusion models for the first time, significantly enhancing image quality and racial fairness while reducing computational costs, outperforming previous methods such as DPO. Experimental results demonstrate the robust generation of anonymized facial images across diverse racial groups, advancing the development of ethical and fair AI for facial image generation and editing.

Models	Training Datasets	Anonymization ↑				Racial Feature Preservation Accuracy ↑				
		Acc	Avg. Dist.	Max. Dist.	Min. Dist.	Asian	Caucasian	African	Latino	Middle Eastern
SD-XL-I (pretrained)		1.00	1.0181	1.2045	0.8267	0.949	0.940	0.921	0.901	0.912
Ours	FFHQ	1.00	1.0215	1.2037	0.8426	0.905	0.941	0.922	0.900	0.912
	CelebA-HQ	1.00	1.0249	1.2103	0.8501	0.913	**0.949**	**0.923**	0.897	0.910
	Asian Dataset	**1.00**	**1.0367**	**1.2294**	**0.8703**	**0.958**	0.945	0.922	**0.912**	**0.915**

Table 4: Quantitative comparison on Multi dataset. Anonymization and Racial Feature Preservation

Models	Training Datasets	FID ↓	SSIM ↑	LPIPS ↓	CLIP-I ↑	PSNR ↑
SD-XL-I (pretrained)		75.03	0.2912	0.7596	0.6531	27.91
Ours	FFHQ	73.13	0.3082	0.7385	0.7295	28.05
	CelebA-HQ	73.34	0.3047	0.7402	0.7328	**28.10**
	Asian Dataset	**72.60**	**0.3149**	**0.7341**	**0.7565**	27.91

Table 5: Quantitative comparison on Multi dataset. Image Quality

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

[1] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiuping Tang. Large-scale celebfaces attributes (celeba) dataset. In *Proceedings of ICCV*, pages 11–15, 2018.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of CVPR*, pages 4401–4410, 2018.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pages 10684–10695, 2022.

[4] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of CVPR*, pages 9799–9808, 2020.

[5] Nataniel Ruiz, Yu Li, Varun Jampani, Yair Pritch, Michael Rubinstein, and Ohad Aberman. Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of CVPR*, pages 22500–22510, 2022.

[6] Robert French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[7] Yunfan Meng, Mingdao Xia, and Dian Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

[8] Rafael Rafailov, Abhishek Sharma, Mitchell Eisner, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, 2024.

[9] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(7):101–116, 2001.

[10] Sreedhar, K., & Panlal, B. Enhancement of images using morphological transformation. *arXiv preprint arXiv:1203.2514*, 2012.

[11] Emmanouil Chatzikyriakidis, Christos Papaioannidis, and Ioannis Pitas. Adversarial face de-identification. In *Proceedings of ICIP*, pages 684–688, 2019.

[12] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Proceedings of ISVC*, pages 565–578, 2019.

[13] Minh-Ha Le, Muhammad Saad Naeem Khan, Georgios Tsaloli, Niclas Carlsson, and Sonja Buchegger. Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security. In *Proceedings of WPES*, pages 87–100, 2020.

[14] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1812.00137*, 2018.

[15] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of CVPR*, pages 12873–12883, 2021.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *In NeurIPS*, volume 33, pages 6840–6851, 2020.

[17] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Chris Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, and T. Salimans. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, volume 35, pages 36479–36494, 2022.

[18] Yize Jiang, Jing Xu, Bing Yang, Jian Xu, and Jing Zhu. Image inpainting based on generative adversarial networks. *IEEE Access, 8:22884–22892*, 2020.

[19] Hongyu Liu, Ziyi Wan, Wei Huang, Yizhou Song, Xiaojun Han, and Jianping Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of CVPR*, pages 9371–9381, 2021.

[20] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jaehong Kim. Expressive text-to-image generation with rich text. In *Proceedings of ICCV*, pages 7545–7556, 2023.

[21] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proceedings of CVPR*, pages 6007–6017, 2023.

[22] Zhihao Zhang, Jiashi Zhao, Feiyang Han, Jianfeng Huang, and Bryan Plummer. Text-to-image editing by image information removal. In *Proceedings of WACV*, pages 5232–5241, 2024.

[23] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *arXiv preprint arXiv:1404.2334*, 2014.

[24] Lu, Y., Tai, Y.-W., and Tang, C.-K. Attribute-guided face generation using conditional cyclegan. In *Proceedings of ECCV*, pages 282–297, 2018.

[25] Kim, G., Kwon, T., and Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of CVPR*, pages 2426–2435, 2022.

[26] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of ICCV*, pages 2085–2094, 2021.

[27] Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of CVPR*, pages 18208–18218, 2022.

[28] Park, D.H., Luo, G., Toste, C., Azadi, S., Liu, X., Karalashvili, M., Rohrbach, A., and Darrell, T. Shape-guided diffusion with inside-outside attention. In *Proceedings of WACV*, pages 4198–4207, 2024.

[29] Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., and Xu, C. Inversion-based style transfer with diffusion models. In *Proceedings of CVPR*, pages 10146–10156, 2023.

[30] Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

[31] Dai, X., Hou, J., Ma, C. Y., Tsai, S., Wang, J., Wang, R., and Parikh, D. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[32] Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., and de Freitas, N. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

[33] Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., and Rastogi, A. Rlaif: Scaling reinforcement

learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[34] Abdelmoniem, A. M., Sahu, A. N., Canini, M., and Fahmy, S. A. Refl: Resource-efficient federated learning. In *Proceedings of EuroSys*, pages 215–232, 2023.

[35] Clark, K., Vicol, P., Swersky, K., and Fleet, D. J. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

[36] Prabhudesai, M., Goyal, A., Pathak, D., and Fragkiadaki, K. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

[37] Bryan Wallace, Michael Dang, Rafael Rafailov, Lan Zhou, Amanda Lou, Satwik Purushwalkam, and Neeraj Naik et al. Diffusion model alignment using direct preference optimization. In *Proceedings of CVPR*, pages 8228–8238, 2024.

[38] Hong, J., Paul, S., Lee, N., Rasul, K., Thorne, J., and Jeong, J. Margin-aware Preference Optimization for Aligning Diffusion Models without Reference. *arXiv preprint arXiv:2406.06424*, 2024.

[39] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of ICLR*, 2022.

[40] Dreamlike Tech Ltd. Dreamlike Diffusion 1.0 License. https://huggingface.co/dreamlike-art/dreamlike-diffusion-1.0. Accessed: 2023-08-30.

[41] Diffusers Team. SD-XL Inpainting 0.1 Model Card. https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1. Accessed: 2023-08-30.

[42] Diffusers Team. SD-Inpainting 2 Model Card. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting. Accessed: 2023-08-30.

[43] AI Hub. AI Hub Website. https://www.aihub.or.kr/. Accessed: 2022-12-05.

[44] Yoni Kirstain, Ariel Polyak, Uriel Singer, Shai Matiana, Jo Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, volume 36, 2024.

[45] InsightFace Team. InsightFace: 2D and 3D Face Analysis Project. https://insightface.ai/. Accessed: 2023-09-03.

[46] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Proceedings of *BMVC*, pages 1–12, 2015.

[47] Zhanhao Liang, Ziyu Wang, Wei Chen, and Qing Zhang. Step-aware Preference Optimization: Aligning Preference with Denoising Performance at Each Step. *arXiv preprint arXiv:2406.04314*, 2024.

Ano-Face-Fair: Race-Fair Face Anonymization in Text-to-Image Synthesis using Simple Preference Optimization in Diffusion Model

Supplementary Material

We propose a novel framework, ***Ano-Face-Fair***, a text-to-face synthesis using a Stable Diffusion Model that emphasizes race fairness. The main script details the editing mechanism and presents comparisons with state-of-the-art methods for anonymization, specifically race fairness. In this section, we provide the practical details of the proposed model. The code is available at https://github.com/i3n7g3/Ano-Face-Fair

A. Implementation Details

Focused Feature Enhancement (FFE) Loss. The error mapping function E_x measures the difference between predicted output, x_p and target image, x_t:

$$E_x = |x_p - x_t| \tag{1}$$

Based on this error map, we generate a critical feature mask, M_c that identifies regions requiring focused enhancement:

$$M_c = (x_p - x_t) > \theta * MAX\left((x_p - x_t)\right) \tag{2}$$

We optimize θ through a constrained optimization problem:

$$\theta = argmin_\theta \left(\mathcal{L}_{FFE}(\theta) + \lambda R(\theta)\right) \tag{3}$$

To maintain a balance between local feature enhancement and global image coherence, we incorporate a regularization term $R(\theta)$ that controls feature density through entropy regularization:

$$R(\theta) = -\sum\left(\frac{M_c}{|M_c|} \cdot log\frac{M_c}{|M_c|}\right) \tag{4}$$

Algorithm 1 Focused Feature Enhancement (FFE) loss, $\mathcal{L}_{FFE}$

Input: output, target, θ, $\mathcal{L}_{FFE}$ weight
Output: $\mathcal{L}_{total}$

function generate M_c (output, target, θ):
    error map = compute absolute difference (output, target)
    max error = find maximum (error map)
    critical areas = error map > max error * θ
    M_c = convert to float (critical areas)
    return M_c

function $\mathcal{L}_{FFE}$(output, target, generate random mask, M_c , $\mathcal{L}_{FFE}$ weight):
    $\mathcal{L}_{base}$= $\mathcal{L}_{mse}$ (output * random mask, target * M_c)
    $\mathcal{L}_{FFE}$= $\mathcal{L}_{mse}$ (output * M_c, target * M_c)
    $\mathcal{L}_{total}$ = $\mathcal{L}_{base}$ + $\mathcal{L}_{FFE}$ weight * $\mathcal{L}_{FFE}$
    return $\mathcal{L}_{total}$

Difference Loss. To address catastrophic forgetting, we maximize angular separation between different racial class embeddings:

$$\mathcal{L}_{DIFF} = \frac{1}{|P|}\sum_{(i,j)\in P}\left(1 - \frac{z_i \cdot z_j}{|z_i||z_j|}\right) \tag{5}$$

where P represents embedding vector pairs from different classes within a batch. We enforce unit normalization constraints:

$$||z_i|| = \left||z_j|\right| = 1, \qquad \theta_{ij} \geq \theta_{min} \tag{6}$$

The minimum angular separation θ_{min} prevents embedding collapse and ensures distinct representation of racial characteristics, maintaining clear boundaries between features in the embedding space.

Algorithm 2 Difference loss, $\mathcal{L}_{DIFF}$

Input: latents, latents for $\mathcal{L}_{DIFF}$
Output: $\mathcal{L}_{DIFF}$

function $\mathcal{L}_{diff}$ (latents, latents for $\mathcal{L}_{DIFF}$):
similarity = compute cosine similarity (latents, latents for $\mathcal{L}_{DIFF}$)
$\mathcal{L}_{DIFF}$ = 1 - similarity
$\mathcal{L}_{DIFF}$ = compute mean ($\mathcal{L}_{DIFF}$)
return $\mathcal{L}_{DIFF}$

Integration of $\mathcal{L}_{FFE}$ and $\mathcal{L}_{DIFF}$. The overall loss function combines instance-level supervision with distribution-level constraints:

$$\mathcal{L}_{Total} = \mathcal{L}_i + \lambda_{pp} * \mathcal{L}_{pp} + \lambda_{FFE} * \mathcal{L}_{FFE} + \mathcal{L}_{DIFF} \quad (7)$$

The prior-preserving weight λ_{pp} is maintained between 0.1 and 1.0, empirically determined to ensure essential knowledge retention while allowing adaptation to new features. This formulation enables $\mathcal{L}_{FFE}$ to enhance critical facial features locally while $\mathcal{L}_{DIFF}$ prevents overfitting to any particular racial group.

Simple Preference Optimization for Diffusion model. Our SimPO approach eliminates the reference model through a direct reward formulation:

$$r_{SimPO}(x, y) = \frac{\beta}{|y|} \sum log\, \pi_\theta\,(\, y_i \mid x, y_{<i}\,) \quad (8)$$

where π_θ represents the policy of the model (the noise prediction network), β is a scaling factor, and $|y|$ normalizes for output length. For preference optimization, logits are calculated as:

$$logits = r\_SimPO(x, y_w) - r\_SimPO(x, y_l) - \gamma \quad (9)$$

For diffusion models specifically, we adapt this formulation to noise prediction tasks:

$$\pi_{logratios} = \mathcal{L}_{model_l} - \mathcal{L}_{model_w} \quad (10)$$

$$logits = \pi_{logratios} - \frac{\gamma}{\beta} \quad (11)$$

Model losses are computed as the Mean Squared Error between predicted and target noise:

$$MSE_k = |\epsilon_\theta(x_{k_t}, t) - \epsilon|^2 \quad (12)$$

Algorithm 3 SimPO Loss Calculation

Input: model π_θ, input x, output y, scaling factor β, margin γ
Output: $\mathcal{L}_{SimPO}$

function L_{SimPO} (π_θ, x, y, β, γ):
1. Calculate model loss functions:
$\mathcal{L}_{model_w}$ = MSE $(\pi_\theta(y_w), target_w)$
$\mathcal{L}_{model_l}$ = MSE $(\pi_\theta(y_l), target_l)$

2. Compute log ratios:
$\pi_{log\ ratios} = \mathcal{L}_{model_l} - \mathcal{L}_{model_w}$

3. Calculate logits:
Logits = $\pi_{log\ ratios} - \gamma$

4. Compute SimPO Loss:
if loss type == "sigmoid":
$\mathcal{L}_{SimPO} = -mean(log(sigmoid(\beta * logits)))$
else if loss type == "hinge":
$\mathcal{L}_{SimPO} = mean(max(0, 1 - \beta * logits))$
return $\mathcal{L}_{SimPO}$

B. Experiments Details

Dataset. We trained the model using a public Asian facial dataset from AI Hub, consisting of 4,515 datasets, based on the Stable Diffusion v2 inpainting model. For Simple Preference Optimization (SimPO) fine-tuning, we utilized the Pick-a-Pic dataset, which includes 583,747 training sets and 500 validation-test sets, curated from user interactions with the Pick-a-Pic web application for text-to-image pairs.

Focused Feature Enhancement (FFE) Loss. The core of $\mathcal{L}_{FFE}$ is the M_c, which identifies regions where the difference between x_p and x_t exceeds a specific θ. In our initial experiments, we found θ = 0.5 to be the optimal balance between capturing important features and preventing overfitting to noise. This loss function assigns greater weight to critical regions identified by M_c. Through further experimentation, we refined our parameters, setting θ to 1.0 and $\mathcal{L}_{FFE}$ weight to 0.01. To select the optimal $\mathcal{L}_{FFE}$ weight, we tested values ranging from 0.001 to 0.1, discovering that 0.01 provides the best balance between enhancing detailed features and maintaining overall image coherence. This value significantly improved the model's ability to capture fine facial features without compromising the global image structure.

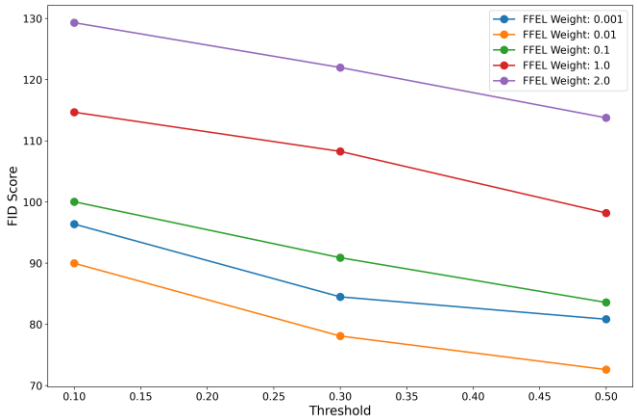


Figure A1: Experiments of $\mathcal{L}_{FFE}$ weight and threshold effects on FID

Difference Loss. To implement $\mathcal{L}_{DIFF}$, we created a set of images using a specific Difference prompt. we used "a photo of the white man" as the Difference prompt, generating 300 images. These generated images were then combined with the training data to compute $\mathcal{L}_{DIFF}$.

Hyperparameters for Simple Preference Optimization. We conducted extensive hyperparameter tuning experiments to optimize SimPO implementation for text-to-face synthesis. Our experiments demonstrated that SimPO's performance is highly dependent on hyperparameter configurations, particularly the interactions between reward scaling parameter (β), margin ratio (γ/β), and learning rate in the preference optimization process. These interactions substantially influence gradient flow, optimization landscape, and training stability during the model training process.

For SimPO fine-tuning process, we set the batch size of 8 with gradient accumulation steps of 2, requiring 24GB of VRAM. We implemented Low-Rank Adaptation (LoRA) with rank 4 and alpha value 4 to optimize computational efficiency while preserving model performance. The complete hyperparameter configuration for our experiments is presented in Table A1.

Our experimental methodology proceeded through four sequential stages. First, we established baseline performance using conservative ranges (β: 200-250, learning rate: 1e-5 to 3e-5) with a fixed γ/β ratio of 0.5. As shown in Figure A2, these initial experiments demonstrated stable convergence across all metrics. Second, we expanded the parameter space to β values up to 1000 and learning rates up to 1e-4, revealing a crucial relationship between β and learning rate: $\alpha = c/\beta$ where $c \in$ [1e-3, 1e-2].

Third, we validated reproducibility using multiple random seeds (42, 123) and compared different learning rate schedulers. As demonstrated in Figure A3, training patterns remained consistent across different initialization conditions. Fourth, we investigated stability boundaries using extreme parameter values (β: 1500-2000, γ/β: 0.2-1.5), which revealed critical stability thresholds: gradient explosion occurred at β>2000 due to exponential scaling in the sigmoid function, and training destabilized when γ/β exceeded 1.5.

Based on these comprehensive experiments, we established our final configuration (β=200, learning rate=1e-5, γ/β=0.5). This configuration achieved stable optimization across key metrics, with raw model loss showing steady decrease while reward accuracies maintained stability between 0.6 and 0.8. The loss ratio and reward consistency stabilized after initial convergence, validating our parameter selection.

Configuration	Parameter	Value
Training Settings	Batch Size	8
	Gradient accumulation Steps	2
	Max Training Steps	8,000
	Validation Steps	100
	Mixed Precision	fp16
Optimizer Settings	Learning Rate	1e-5
	LR Scheduler	Constant with warmup
	Warmup Steps	800
	Optimizer8-bit	Adam
SimPO Parameters	β (Reward Scaling)	200
	γ/β Ratio (gamma ratio)	0.5
LoRA Settings	Rank	4
	Alpha	4
Model settings	Memory Required	24GB VRAM

Table A1: SimPO training hyperparameters.

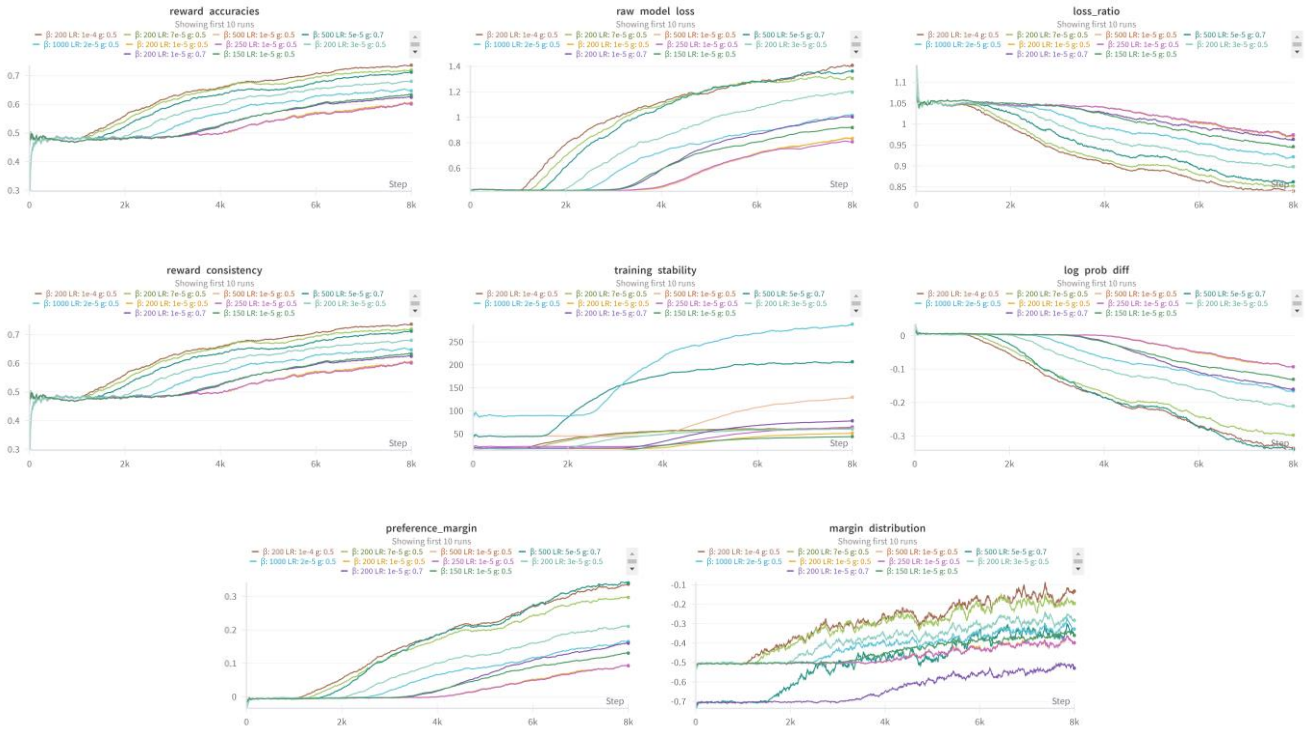


Figure A2: Training Dynamics Visualization: Parameter Sensitivity Analysis and Optimization. Results of initial parameter exploration demonstrate training dynamics under conservative ranges (β=200-250, learning rate=1e-5 to 3e-5). Evaluation metrics show the effectiveness of baseline configuration (β=200, learning rate =1e-5), where reward accuracies and loss curves exhibit stable convergence. Further parameter analysis identifies a crucial relationship between β and learning rate ($\alpha = c/\beta$, $c \in$ [1e-3, 1e-2]), determining optimal ranges for stable training.

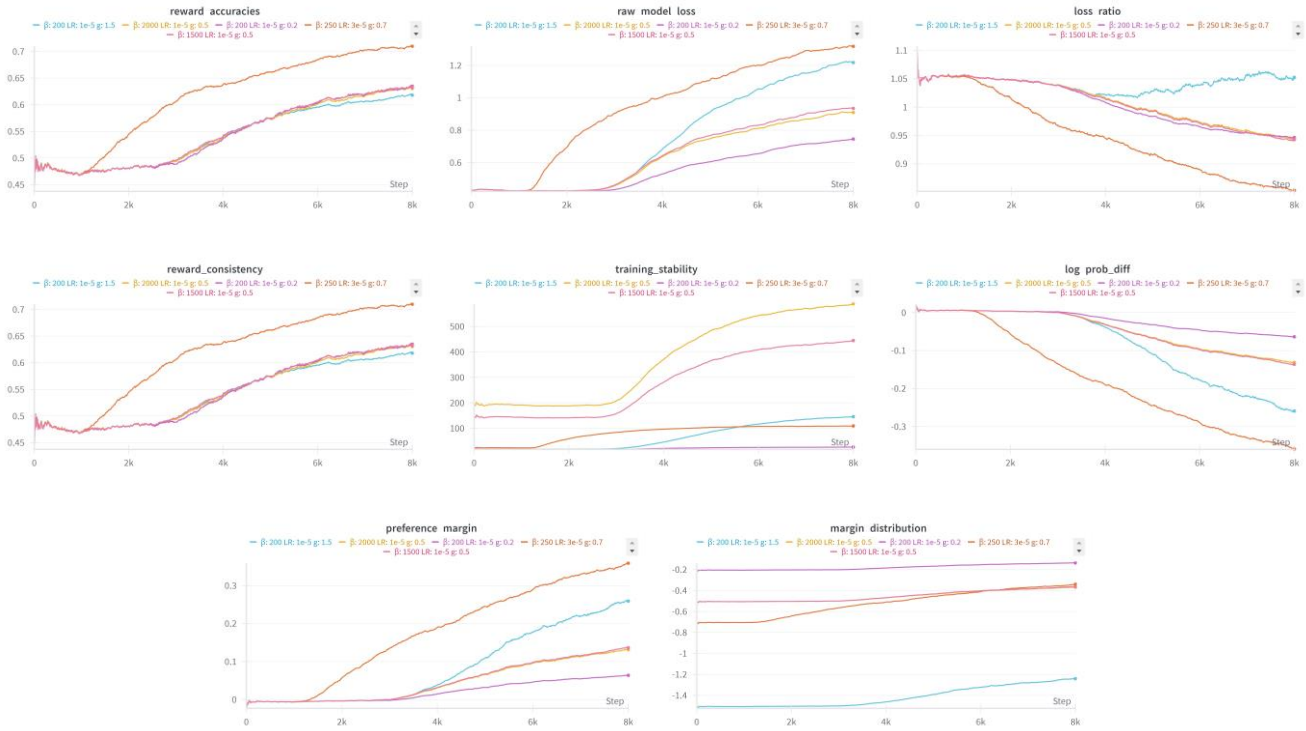
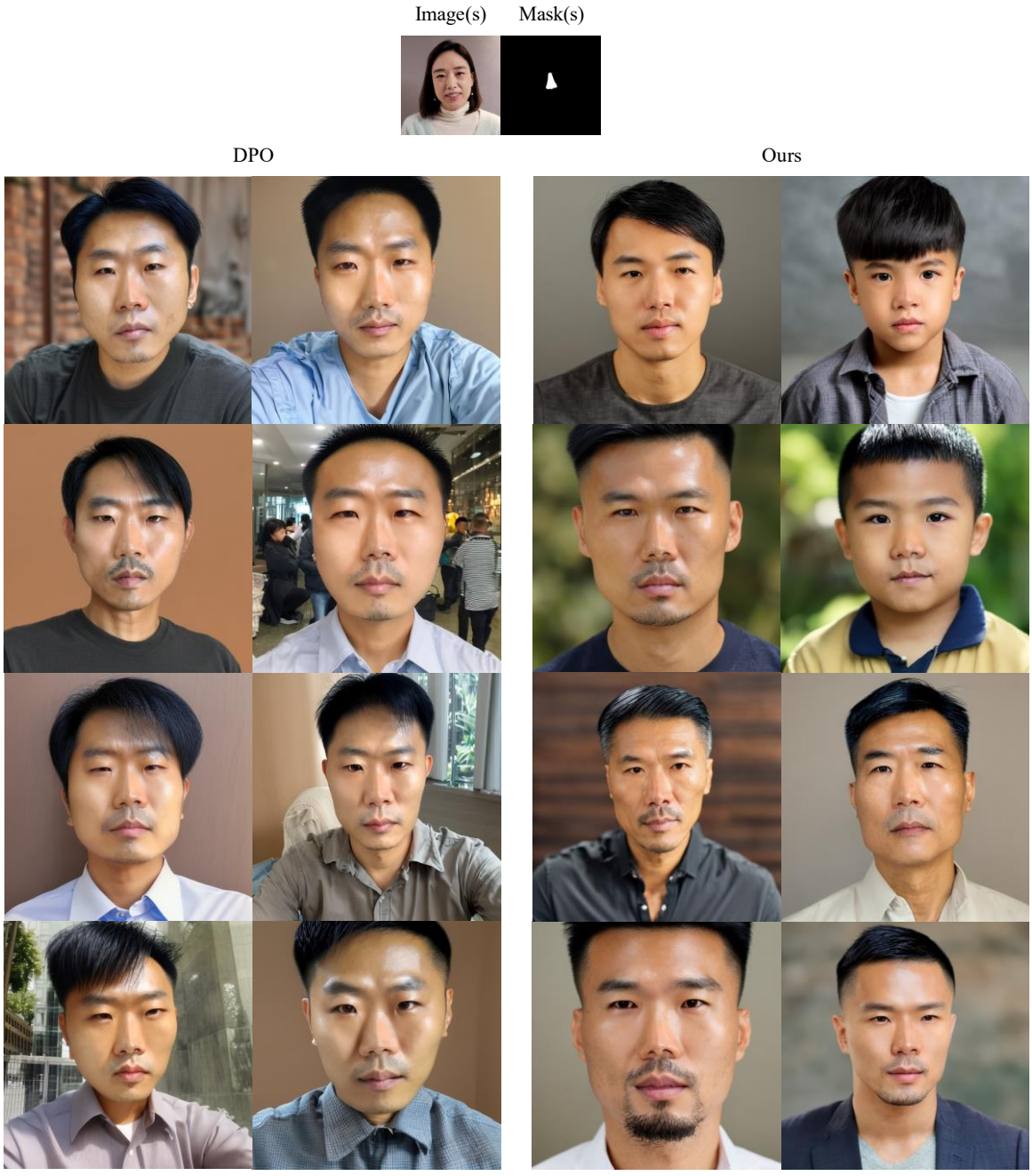


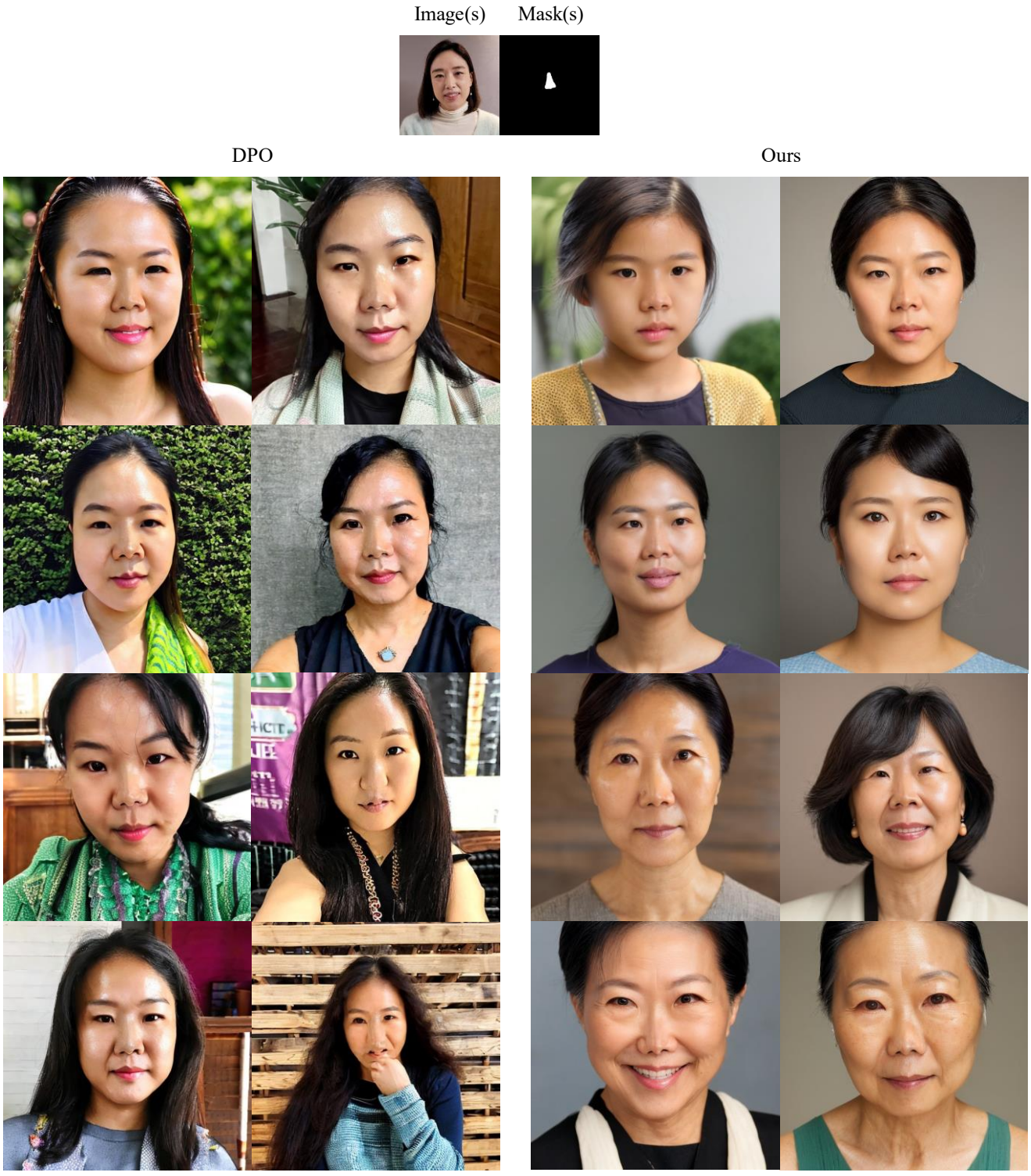
Figure A3: Training Dynamics Visualization: Reproducibility Validation and Stability Boundaries. Stability boundary analysis and reproducibility tests. Training patterns remain consistent across different random seeds (42, 123), confirming method reliability. Performance metrics reveal clear stability thresholds: gradient explosion occurs at $\beta>2000$, and training destabilizes when γ/β exceeds 1.5. These findings establish operational boundaries for reliable model training, supporting our final configuration selection ($\beta=200$, learning rate=1e-5, $\gamma/\beta=0.5$).

More Qualitative Results. We provide more results of ***Ano-Face-Fair*** through various experiments and comparisons. First, Figure A4-A5 demonstrates the effectiveness of our method compared to DPO for Asian face anonymization, with Figure A4 showing results for Asian men and Figure A5 for Asian women. Figure A6-A7 shows the qualitative results of face anonymization for Asian race. Figure A8-A9 presents the comparison of face anonymization results across different races (Female and Male). Figure A10-A11 illustrates the qualitative results of face anonymization using mask-based, mask and text-based prompting.



"He is an Asian man."

Figure A4: Comparison of face male face anonymization results (DPO vs Ours)



"She is an Asian woman."

Figure A5: Comparison of face female face anonymization results (DPO vs Ours)

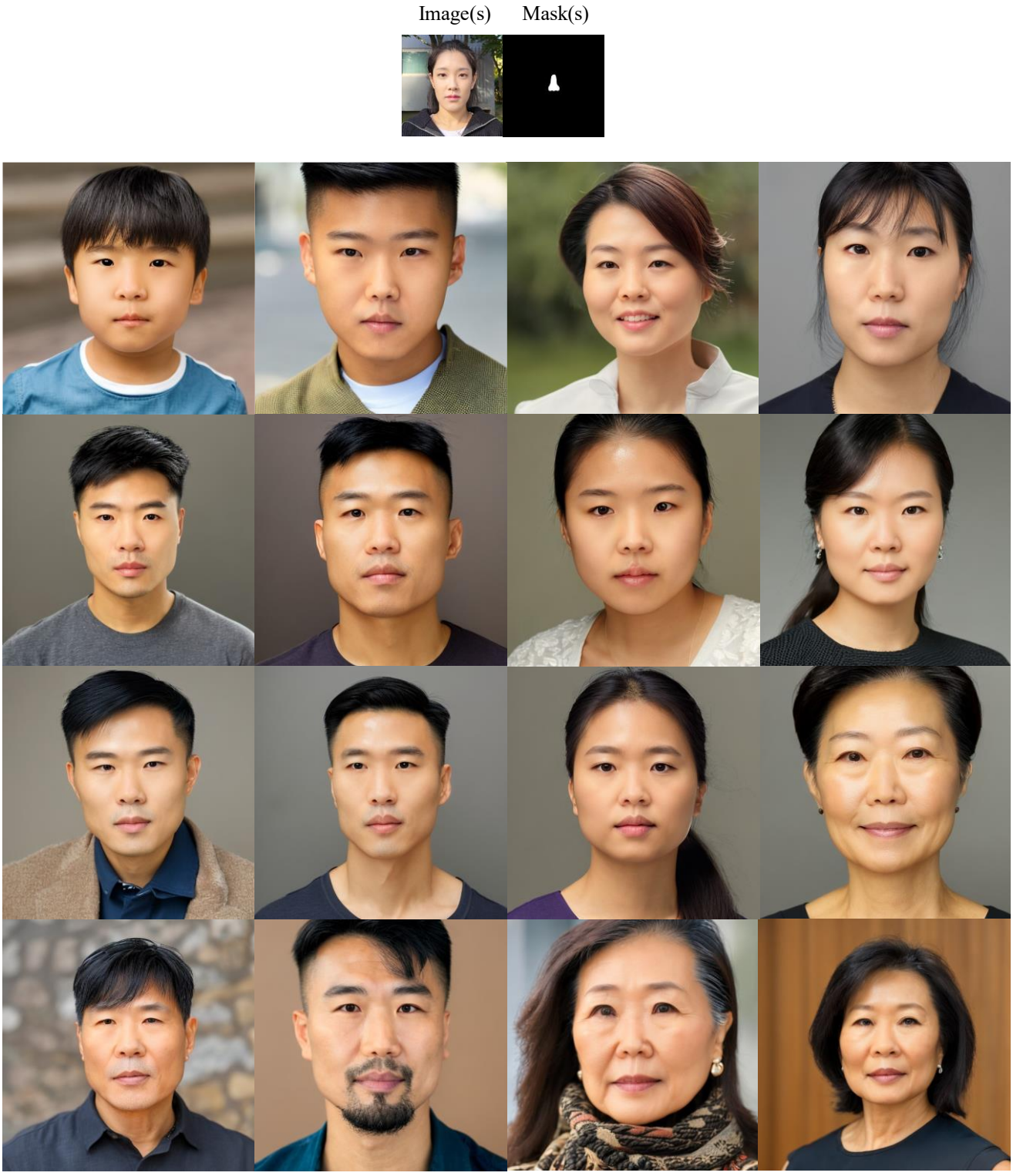


Figure A6: Qualitative results of face anonymization for Asian race

Figure A7: Qualitative results of face anonymization for Asian race

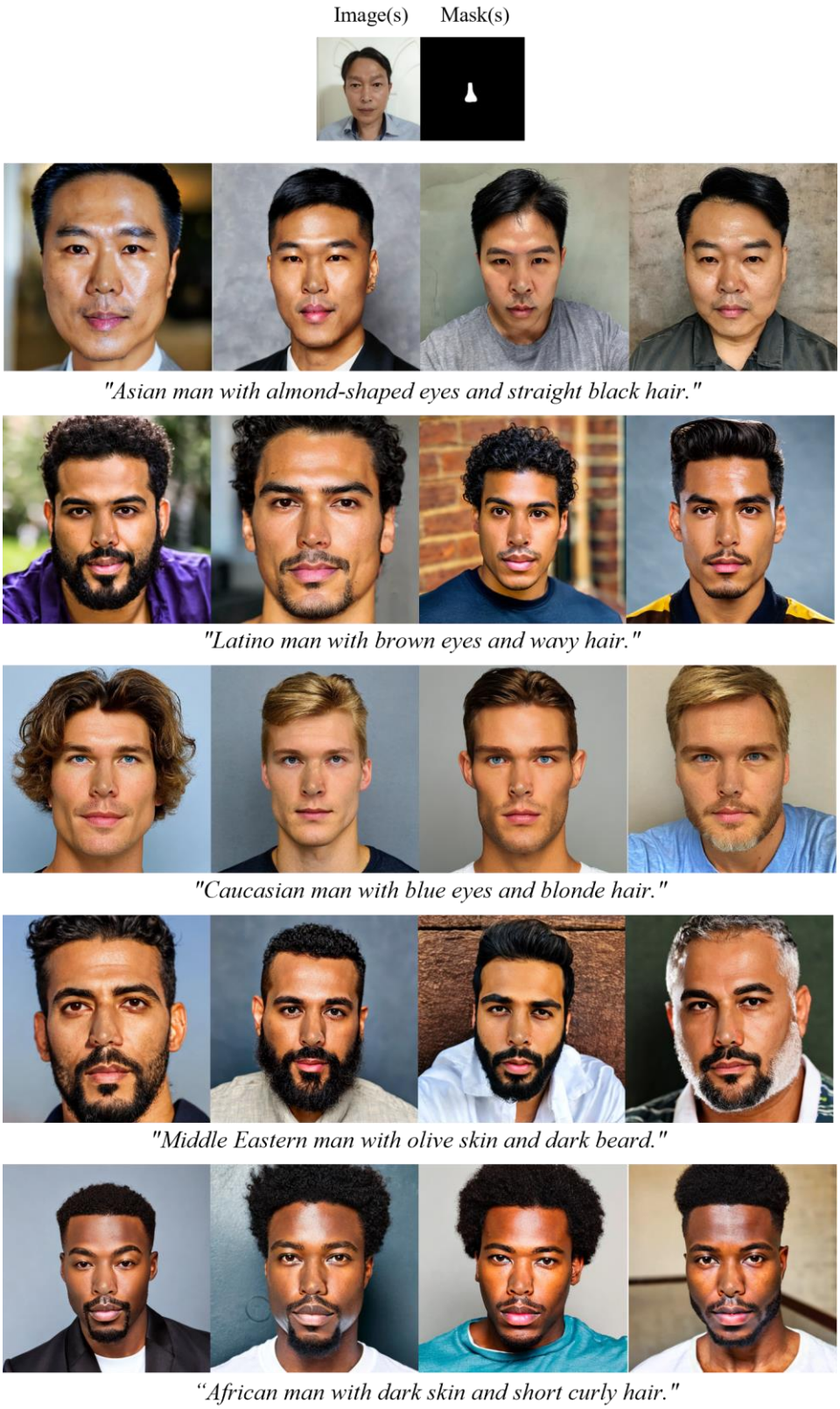


Figure A8: Comparison of face anonymization results across different races (Male)

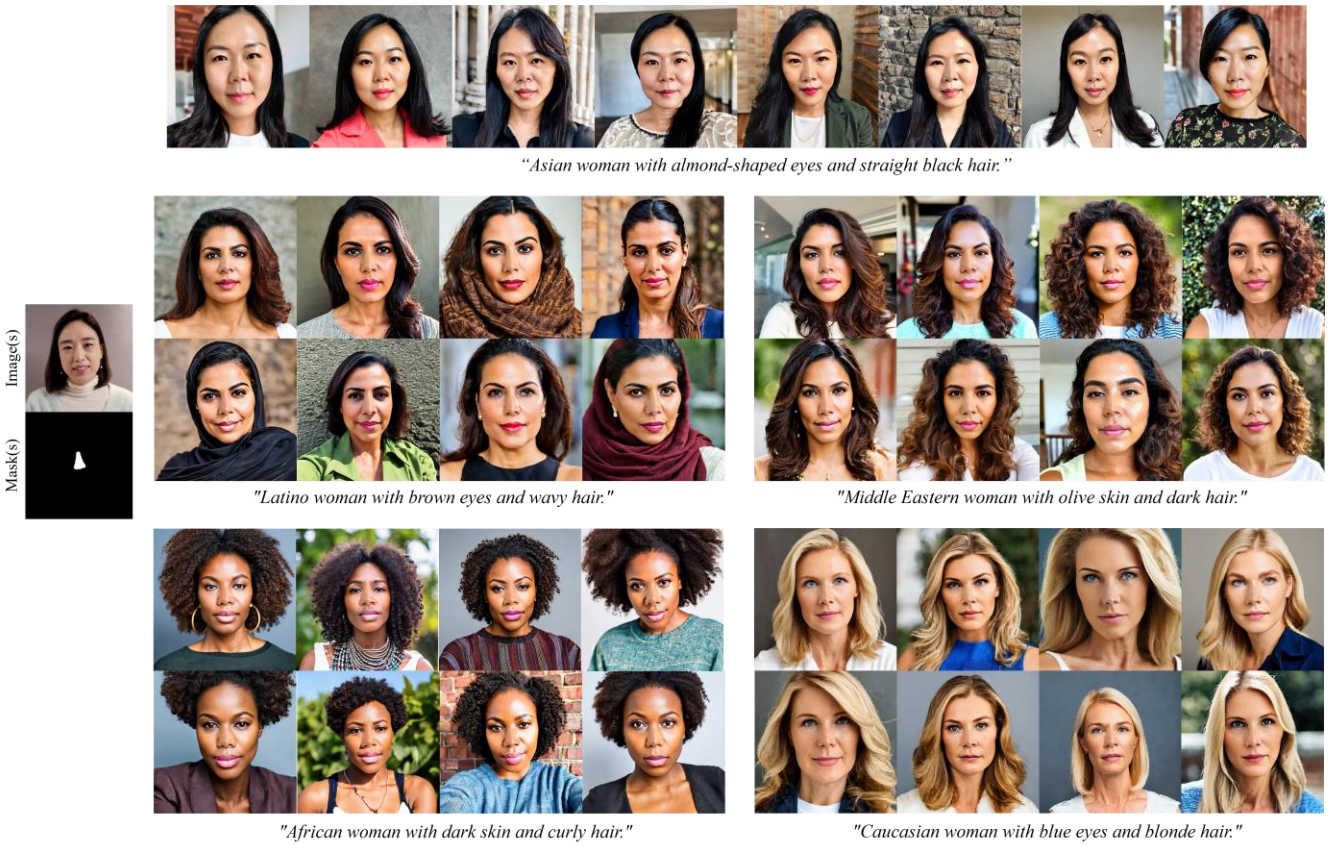


Figure A9: Comparison of face anonymization results across different races (Female)

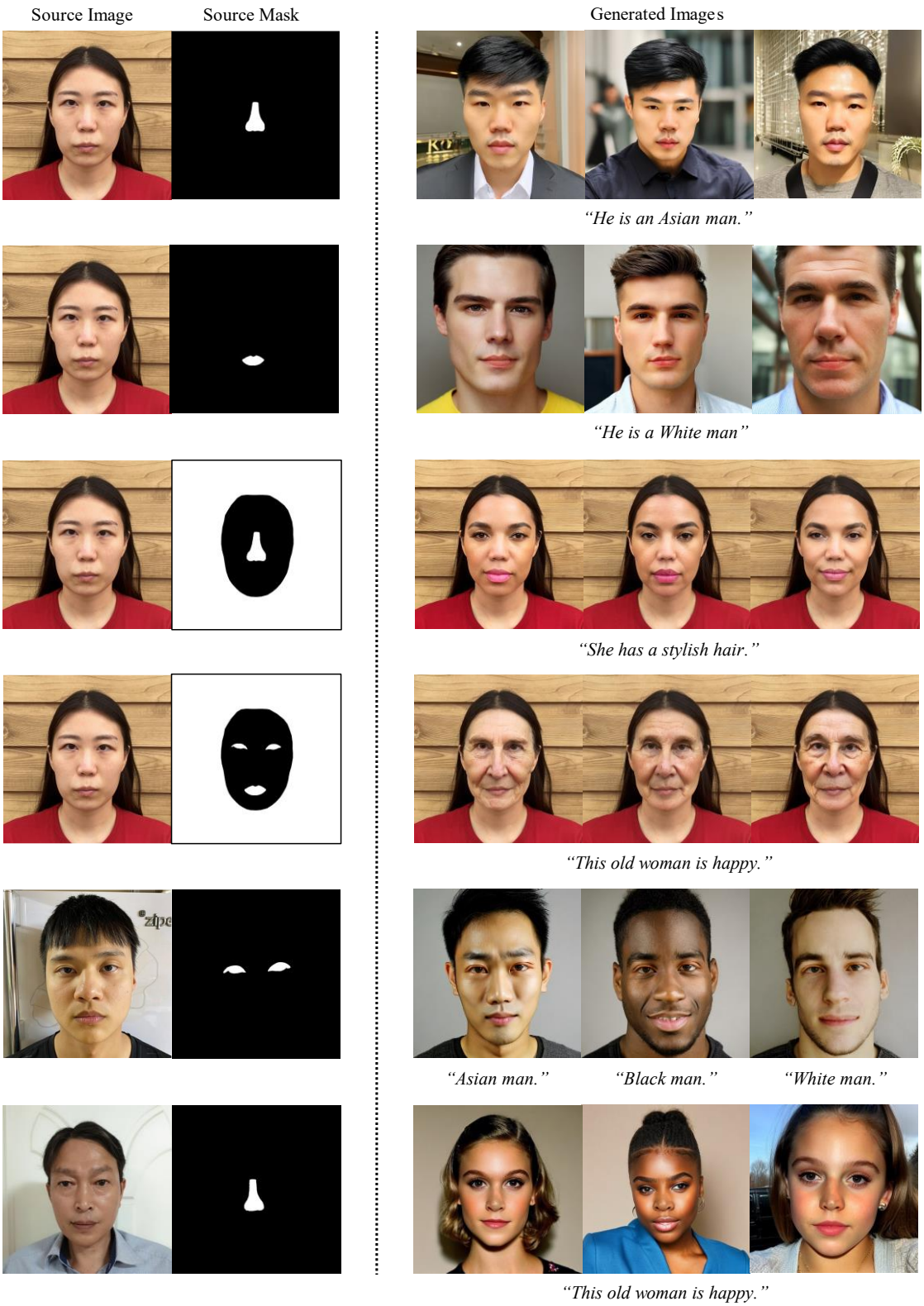


Figure A10: Qualitative results of face anonymization using mask-based prompting

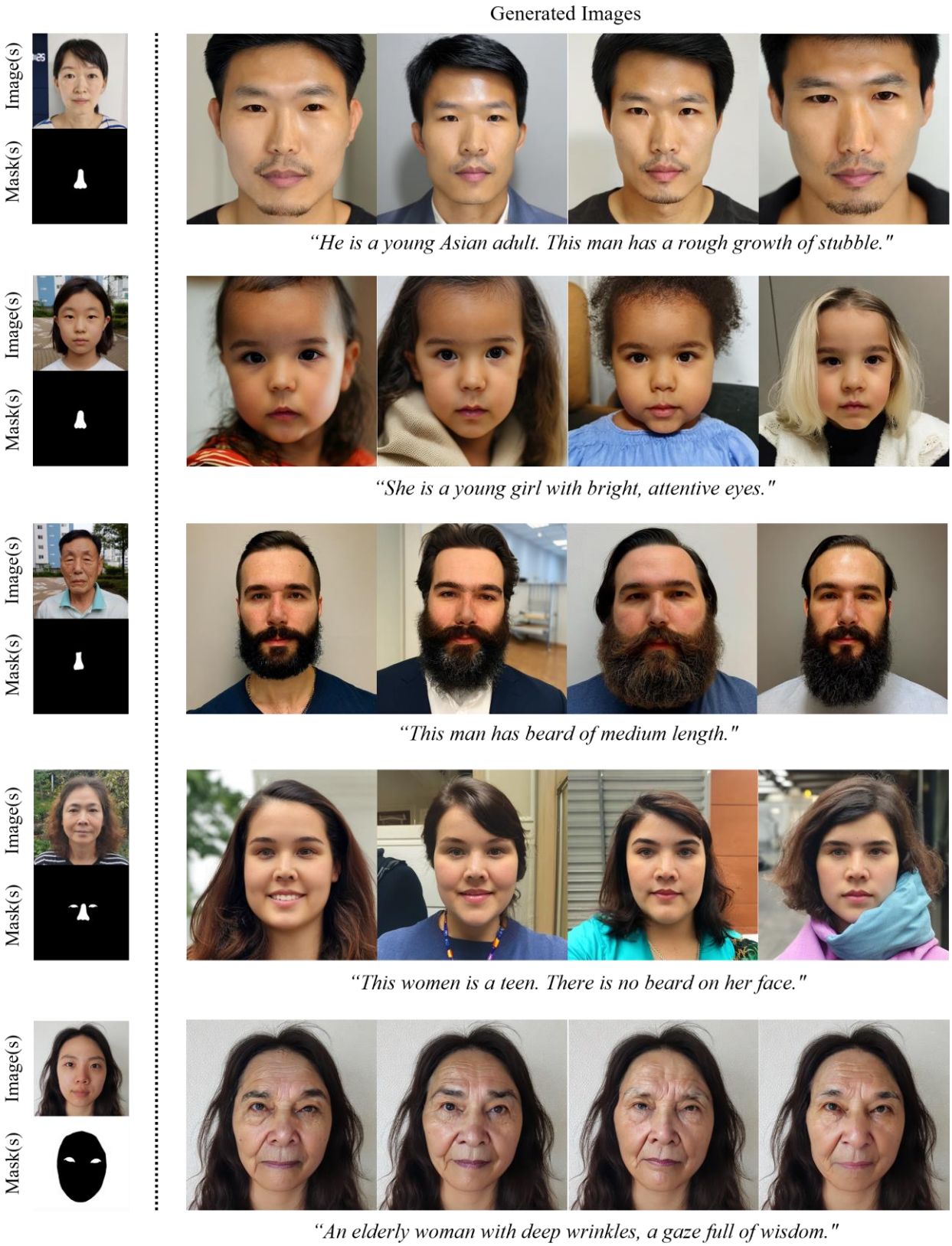


Figure A11: Qualitative results of face anonymization results using mask-based and text-based prompting