# Reverse Personalization

## Supplementary Material

## 6. Balanced performance across privacy and utility

In Fig. 10, we show five privacy-utility trade-off plots on CelebA-HQ [38] (CHQ) and FFHQ [39] (FHQ). The x-axis shows the re-identification rate, computed using AdaFace [43]. The y-axis shows utility, with lower values indicating better performance for expression, gaze, pose, and FID [29], and higher values for Face IQA [13]. Each method (NullFace [50], FAMS [49], FALCO [6], RiDDLE [51], LDFA [45], DP2 [34]) appears as a point. A gradient background highlights the lower-left (or upper-left for Face IQA [13]) corner as the optimal balance between privacy and utility. This visualization shows that our method consistently lies closest to the "sweet spot," while others compromise either privacy or utility.

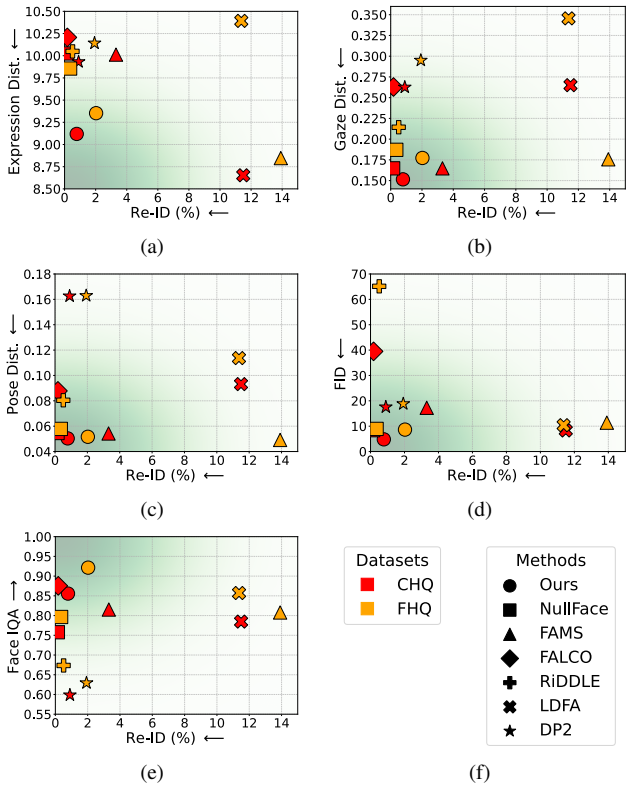


Figure 10. Privacy–utility trade-off plots. The gradient background highlights the optimal region (darker green) for balancing privacy and utility.

## 7. Privacy-utility trade-off across guidance scales

Figure 11 shows the effect of negative classifier-free guidance [30] scales. Re-identification rates, computed with AdaFace [43] and SwinFace [66], are compared to utility measures: expression distance, gaze distance, pose distance, FID [29], and Face IQA [13]. Points represent different guidance scales (-20, -15, -10, -5). More negative scales reduce re-identification rates but increase attribute distances (expression, gaze, pose) and degrade image quality (higher FID [29], lower Face IQA [13]). This trade-off occurs because excessively negative guidance values impose the synthetic identity too strongly, altering non-identity attributes that should remain unchanged. Consistent with prior findings [72], excessive classifier-free guidance [30] leads to overall image degradation.

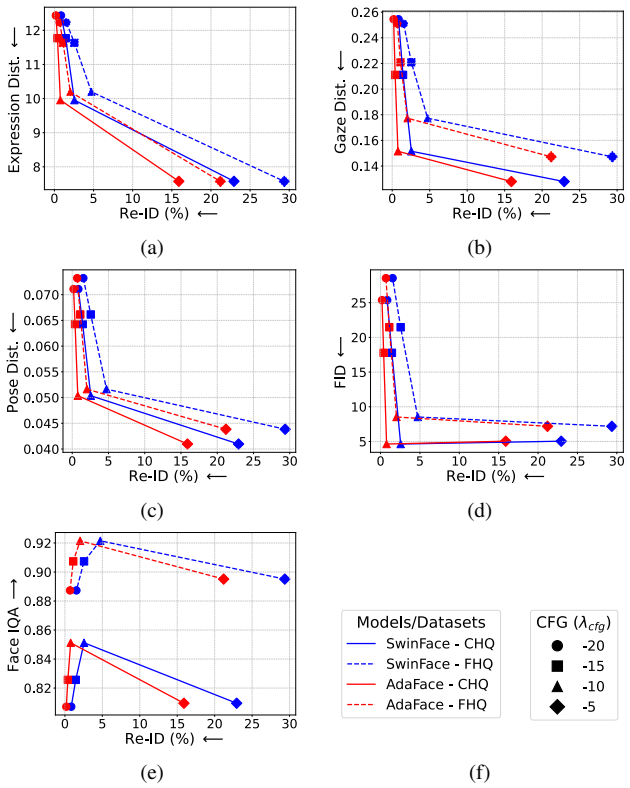


Figure 11. Privacy–utility trade-off plots showing that increasingly negative guidance scales lower re-identification rates but also degrade image quality and distort non-identity attributes (expression, gaze, pose).

## 8. Challenging cases and limitations

While prior methods struggle with extreme face angles or occlusions [36, 83], our approach performs reliably in these cases. The greater challenge occurs with extreme or uncommon expressions (see Fig. 12). In such cases, both our method and baselines struggle to preserve the original expression, likely due to limited training data for such expressions.

Image quality is also bounded by the underlying diffusion model (SDXL [65] in our experiments). Using more advanced diffusion models could further improve quality.

Our framework currently anonymizes single images. When applied to videos, it lacks temporal consistency. Extending the method to video anonymization using video diffusion models [7] is a promising direction for future work.

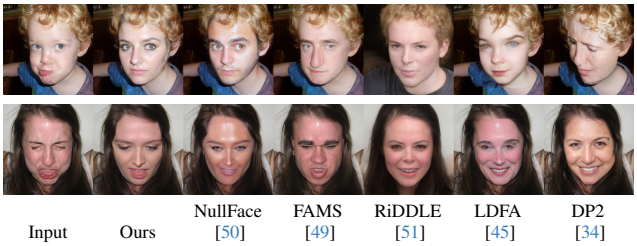


Figure 12. Examples where our method and baselines fail to preserve rare or extreme facial expressions.

## 9. Identity recovery test

A potential concern is whether the original identity can be recovered from an anonymized image by applying a negative classifier-free guidance [30] scale to reverse anonymization. We argue that this is not feasible, as the outcome of our method depends on the interaction of multiple components—including the model architecture, inversion process, and hyperparameter settings—which prevent reversibility.

We visualize recovery attempts in Fig. 13. The recovered images remain visibly different from the original inputs, aligning with the quantitative findings in Tab. 5, where re-identification rates remain low and comparable to anonymized images. These results demonstrate that the original identity cannot be restored through this approach.

## 10. Societal risks of AI-generated human faces

AI-generated human faces pose societal risks. Now nearly indistinguishable from real ones, such faces enable convincing fake identities on social media, fostering manipulation, fraud, and disinformation [20]. These technologies also risk amplifying cultural, racial, and gender biases [5]. When trained on imbalanced datasets, face-generation models often reinforce societal prejudices, producing skewed and exclusionary representations. AI-generated faces are also used for sexual exploitation, such as creating non-consensual deepfake pornography [76] that targets and violates victims without consent. While AI-generated faces have promising applications, their misuse highlights the urgent need for coordinated action by policymakers and technologists to address these harms.


Figure 13. Identity recovery attempts using negative classifier-free guidance [30] show that recovered images differ visually from the original inputs, demonstrating that the original identity cannot be restored.

| | Re-ID (%) ↓ | | | |
|---|---|---|---|---|
| | SwinFace | | AdaFace | |
| | CelebA-HQ | FFHQ | CelebA-HQ | FFHQ |
| Anonymized | 2.556 | 4.724 | 0.768 | 2.028 |
| Attempt to recover | 2.284 | 4.117 | 0.566 | 1.430 |

Table 5. Identity recovery test by applying negative classifier-free guidance [30] to anonymized images. The low re-identification rates indicate that the original identity cannot be recovered through this method.

## 11. Additional qualitative results

Qualitative comparisons between our method and six state-of-the-art approaches—NullFace [50], FAMS [49], FALCO [6], RiDDLE [51], LDFA [45], and DP2 [34]—are shown for CelebA-HQ [38] and FFHQ [39]. CelebA-HQ [38] results are in Figs. 14 to 16, and FFHQ [39] results are in Figs. 17 to 19.


Figure 14. Qualitative comparison of anonymization results on CelebA-HQ [38].


Figure 15. Qualitative comparison of anonymization results on CelebA-HQ [38].

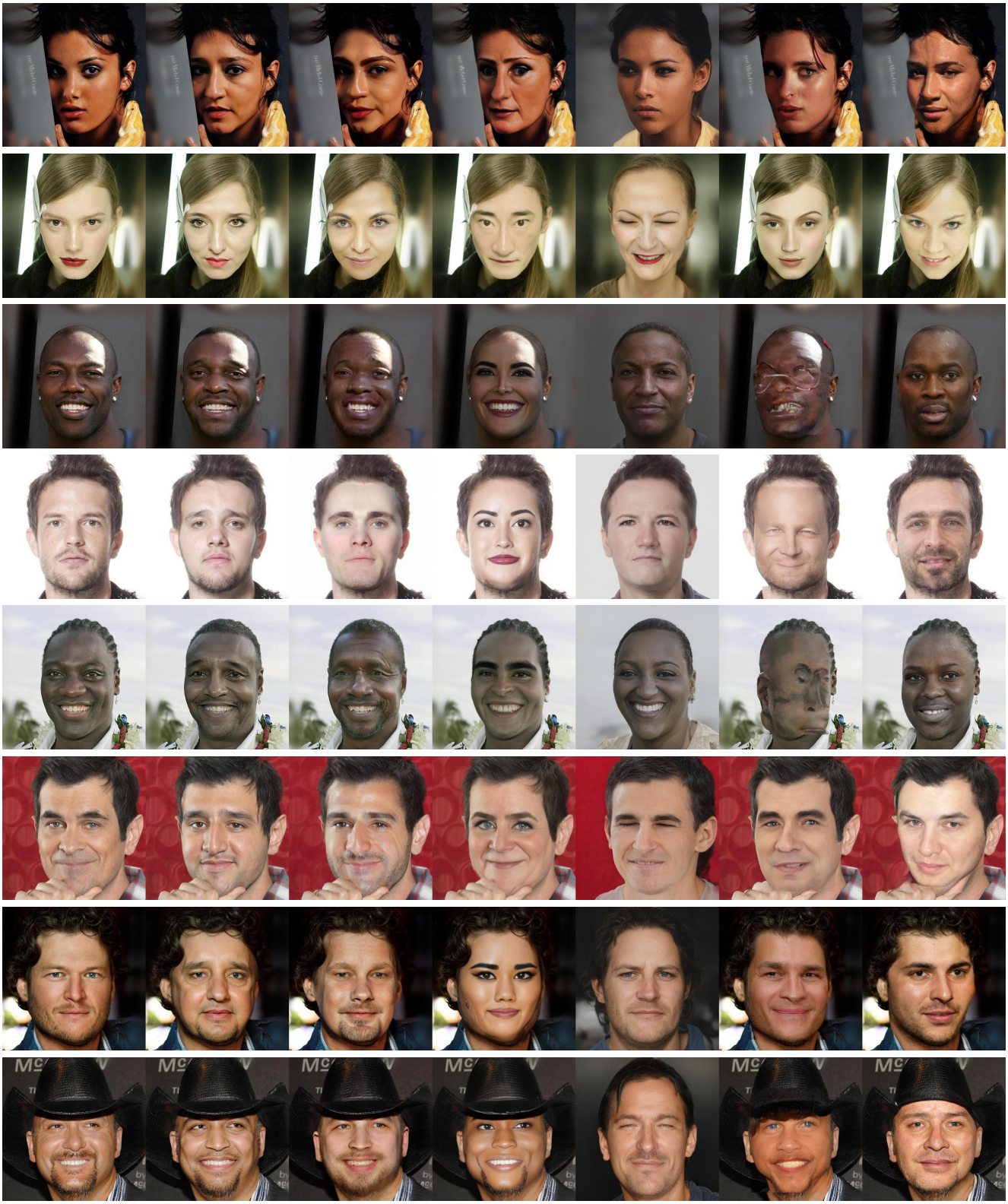


Figure 16. Qualitative comparison of anonymization results on CelebA-HQ [38].

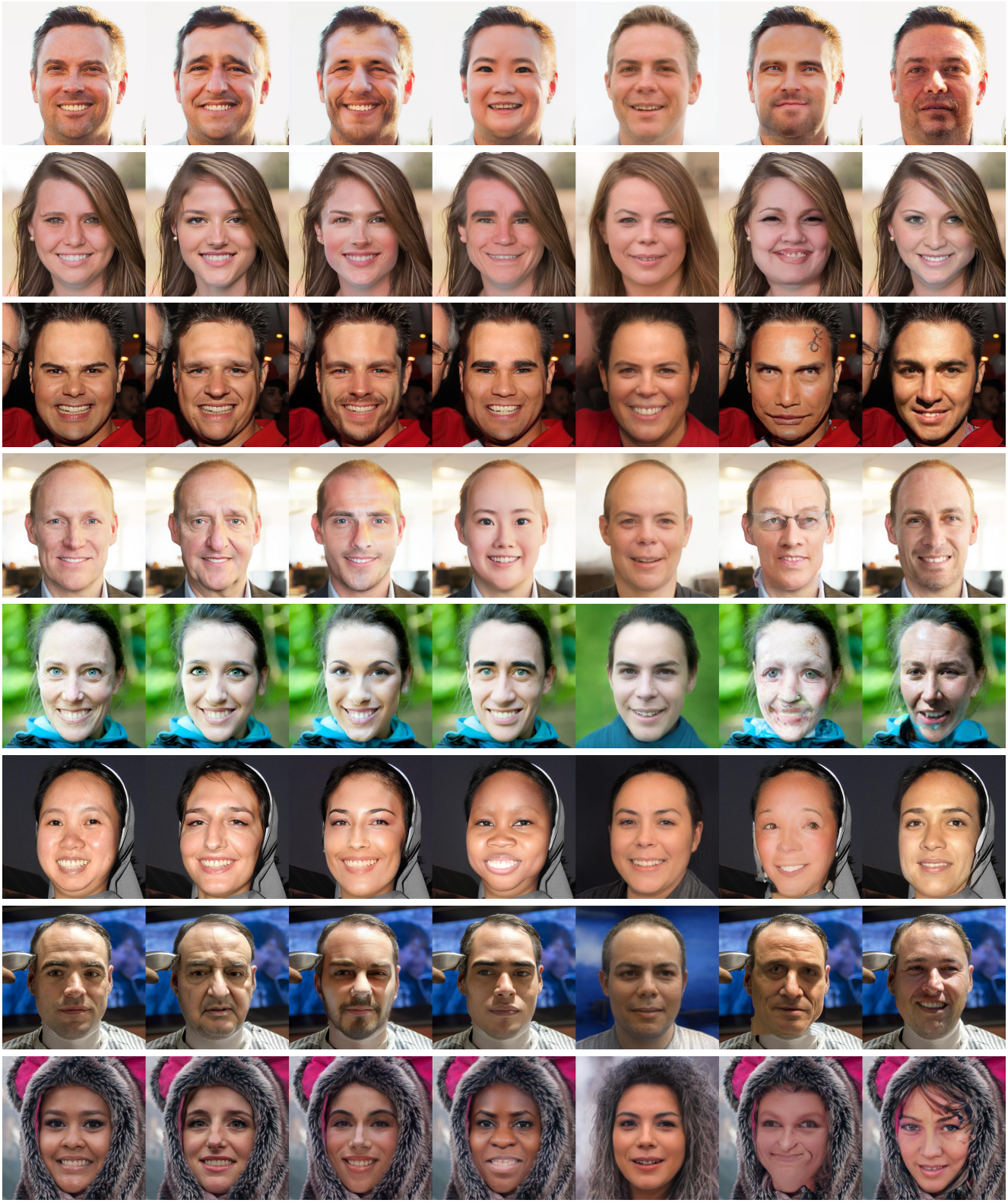


Figure 17. Qualitative comparison of anonymization results on FFHQ [39].

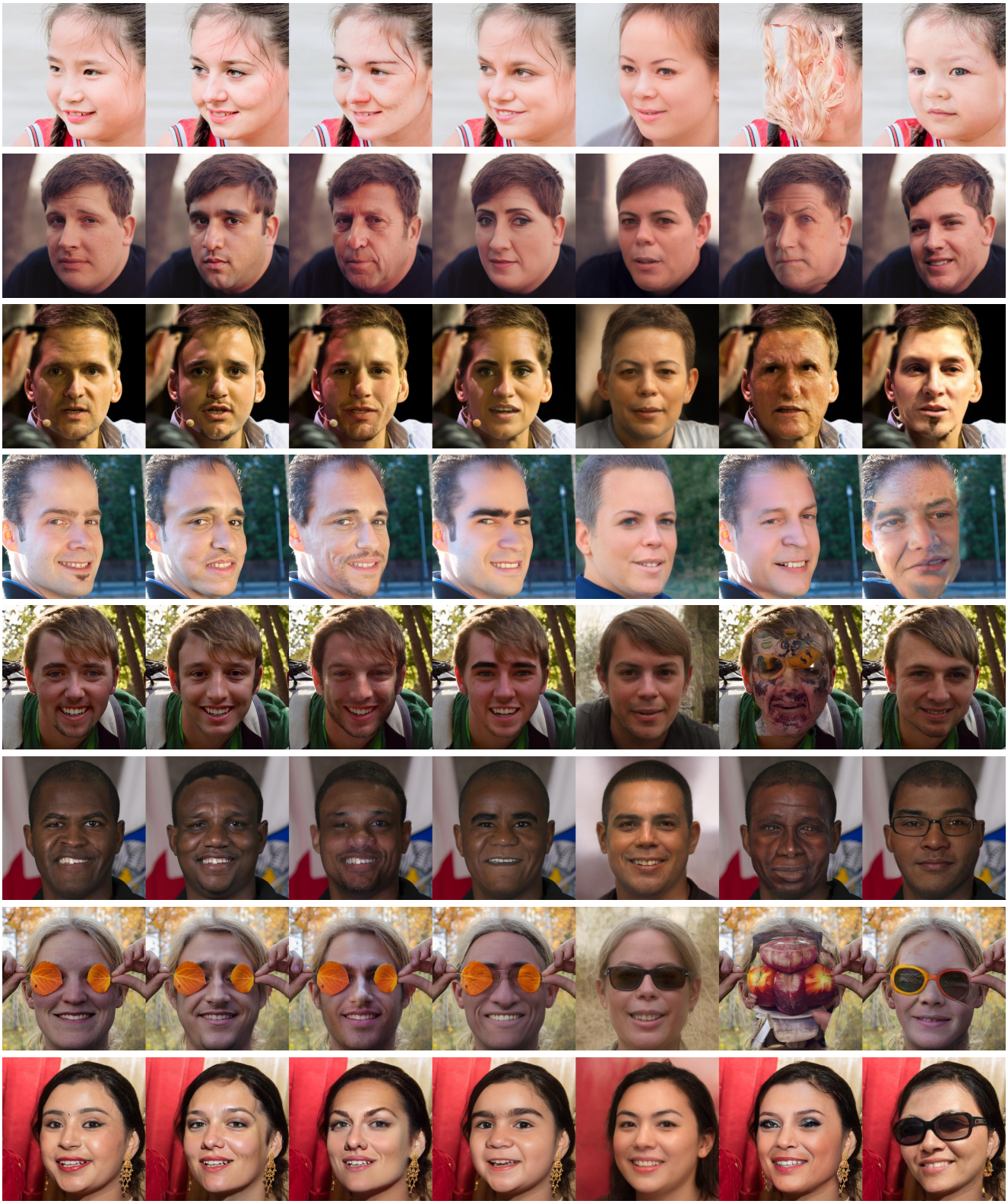


Figure 18. Qualitative comparison of anonymization results on FFHQ [39].

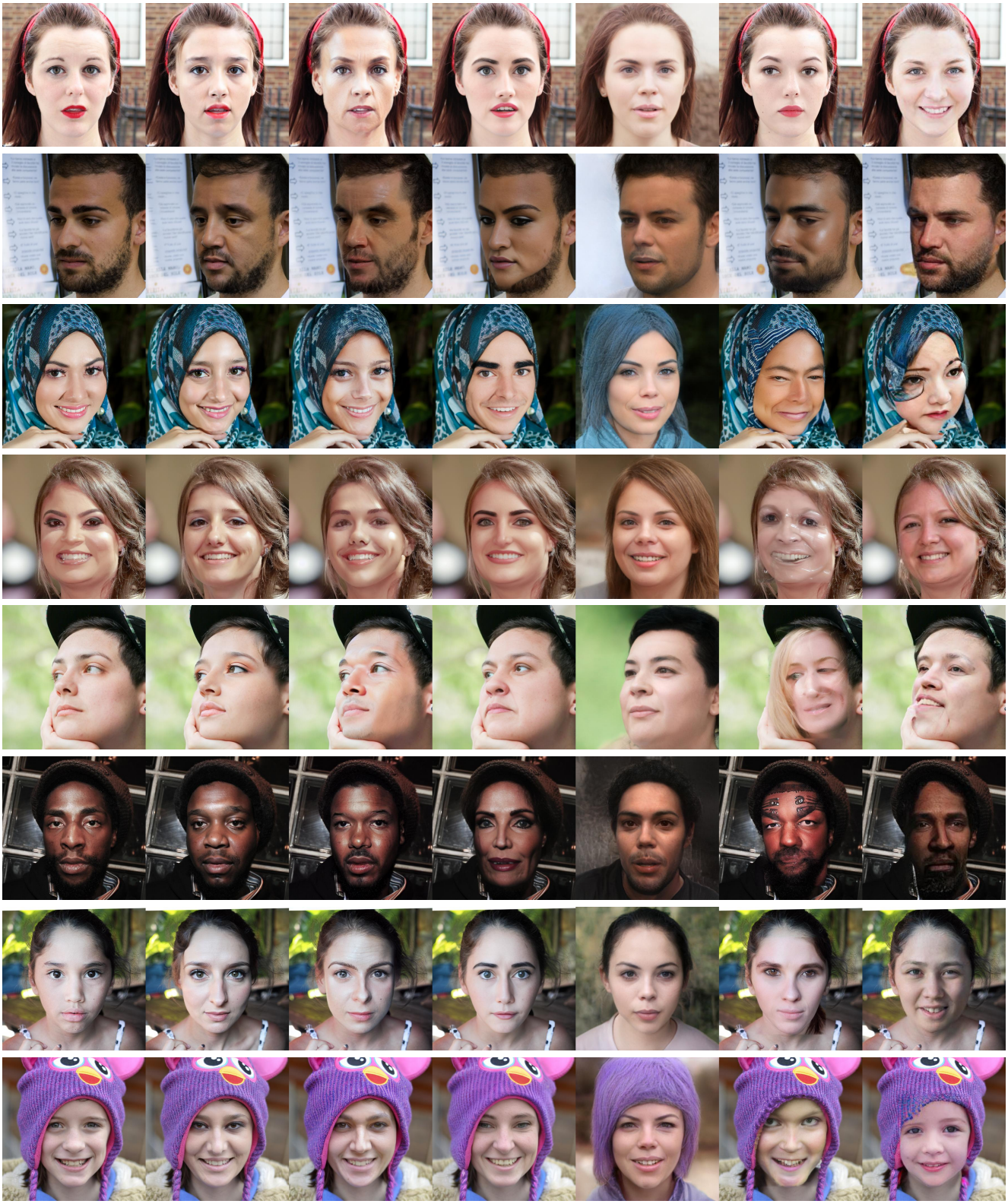


Figure 19. Qualitative comparison of anonymization results on FFHQ [39].