No Way To Steal My Face: Proactive Defense Against Identity-Preserving Personalized Generation

Supplementary Material

A. Methodological and Evaluation Details

While the main paper introduces the two-stage defense principle of **IDGuardian**, this section provides the complete algorithmic details of its optimization process. Specifically, IDGuardian jointly optimizes adversarial perturbations through two complementary strategies—**Cross-Encoder Identity Field Confounding** and **Guidance-Flow Identity Deflection**—which correspond to the disruption of identity extraction and injection, respectively. For clarity and reproducibility, we detail the optimization objectives, loss composition, and implementation settings that were omitted for brevity in the main text.

A.1. Optimization Process of IDGuardian

Our defense operates in the latent denoising space of the pretrained SDXL backbone within the IP-Adapter SDXL Plus Face pipeline, and optimizes an adversarial perturbation δ on the input image x, such that the final protected image $x_{\text{adv}} = x + \delta$ preserves perceptual quality while significantly reducing identity consistency in downstream personalized generation.

The optimization is carried out iteratively over N_{iter} steps. At each step:

- The input clean image is encoded into the latent space using the VAE encoder.
- A random diffusion timestep t is sampled, and corresponding noise is added to simulate a forward noisy latent x_t.
- Identity embeddings are computed from the clean and perturbed versions of the input, and serve as conditioning inputs for the UNet during denoising.
- Two complementary gradients are computed: i) the identity loss gradient $\nabla_\delta \mathcal{L}_{\text{ID}}$ based on FaceNet and CLIP similarity; ii) a distributional score gradient S^*, which measures the semantic divergence between adversarial and clean identity trajectories in diffusion space.
- To match the spatial resolution and RGB channels of the perturbation, S^* is upsampled via bilinear interpolation (PyTorch: *F.interpolate(..., mode='bilinear', align_corners=False)*) and only the first three channels (RGB) are retained to align with the perturbation tensor.
- These gradients are normalized and combined to update δ via a projected gradient descent step, bounded by the perturbation constraint $\|\delta\|_\infty < \epsilon$.

Algorithm 1 outlines the full optimization pipeline. This instantiation reflects a synergy between semantic feature-space objectives and internal generative guidance dynamics, ensuring that the optimization is both perceptually aware and diffusion-aligned.

Algorithm 1 IDGuardian

Input: original image x, perturbation budget ϵ, learning rate α, text prompt y, max steps N_{iter}
Output: optimized adversarial image x_{adv}
Load pretrained **IP-Adapter SDXL Plus Face** pipeline components: VAE Encoder, UNet, and CLIP image encoder
Encode image into latent space: $x_0 \leftarrow \text{VAE\_Encoder}(x)$
Initialize perturbation $\delta \leftarrow 0$
for $i = 1$ to N_{iter} **do**
  Sample random timestep t, corresponding noise ϵ_t
  Compute noisy latent: $x_t \leftarrow \text{AddNoise}(x_0, \epsilon_t, t)$
  $x_{\text{adv}} \leftarrow \text{clip}(x + \delta, 0, 1)$
  Compute embeddings: $\text{CLIP}(x), \text{CLIP}(x_{\text{adv}})$
  Predict denoising outputs:
    $\hat{\epsilon}_{\text{clean}} \leftarrow \text{UNet}(x_t, t, \text{CLIP}(x), y)$
    $\hat{\epsilon}_{\text{adv}} \leftarrow \text{UNet}(x_t, t, \text{CLIP}(x_{\text{adv}}), y)$
  Compute score gradient:
    $S^* \leftarrow -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \cdot (\hat{\epsilon}_{\text{adv}} - \hat{\epsilon}_{\text{clean}})$
  Upsample S^* to image resolution using bilinear interpolation and align channels with RGB:
    $S^*_{\text{up}} \leftarrow \text{BilinearUp}(S^*, \text{size} = x.\text{shape})[:, :3, :, :]$
  Compute identity loss gradient: $\nabla_\delta \mathcal{L}_{\text{ID}}$
  Combine gradients:
    $\text{total\_grad} \leftarrow \text{Normalize}(\nabla_\delta \mathcal{L}_{\text{ID}}) - \text{Normalize}(S^*_{\text{up}})$
  Update perturbation: $\delta \leftarrow \delta - \alpha \cdot \text{sign}(\text{total\_grad})$
  Clip: $\delta \leftarrow \text{clip}(\delta, -\epsilon, \epsilon)$
end for
Return $x_{\text{adv}} = x + \delta$ =0

A.2. DreamBooth Fine-Tuning and Similarity Evaluation

For DreamBooth and other training-based methods, which typically require multiple images per identity for fine-tuning, we provide four distinct images per identity. For each generated image of that identity, we compute its similarity to each of the four original images, yielding four similarity scores. We then average these four scores to obtain a single similarity metric for that generated image. This procedure is repeated for all generated images of the identity, and the final reported similarity is obtained by averaging across all generated images. This ensures that the evaluation reflects the model's performance using all available reference images rather than relying on a single image.

A.3. Reproducibility Resources

For reproducibility, we list the main public resources used in this work:

- **FaceNet implementation:** We adopt the official FaceNet implementation from the GitHub repository `davidsandberg / facenet`, using the *Inception-ResNet-v1* checkpoint with identifier `20180402-114759` in our loss computation.
- **Evaluation toolkit:** We use the public DeepFace toolkit from the GitHub repository `serengil/deepface` for identity similarity evaluation.
- **Datasets:** The VGGFace2 dataset is accessed from the HuggingFace repository `datasets / ProgramComputer / VGGFace2`. The CelebA-HQ dataset is obtained from the GitHub repository `IIGROUP/MM-CelebA-HQ-Dataset`.

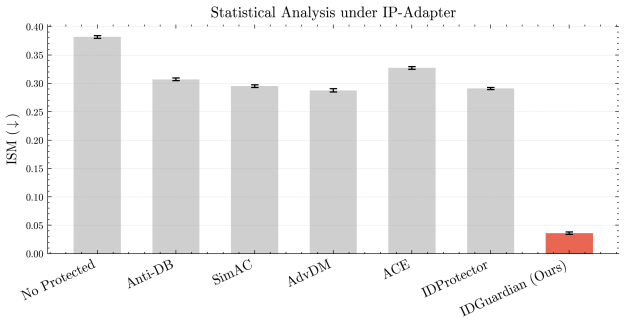

Figure 1. Statistical analysis of identity similarity (ISM) under IP-Adapter. Error bars denote the variability across repeated runs.

B. Detailed Quantitative Results

B.1. Comparative Evaluation with Baselines

In the main paper, we summarized the performance of IDGuardian against baseline methods using average identity similarity scores and a representative perceptual quality metric. Here, we provide comprehensive quantitative results for all individual identity similarity metrics (ArcFace, FaceNet, VGGFace) and perceptual quality metrics (FaceQNet, MagFace) across the tested methods. Table 3 and Table 4 present detailed comparisons highlighting the consistent superiority of IDGuardian in identity suppression and its effectiveness in preserving perceptual quality.

We further include statistical analysis results, demonstrating that the observed performance gains remain stable and consistent across repeated runs. Taking IP-Adapter as an example, as shown in Figure 1, IDGuardian exhibits consistently lower identity similarity with small variance compared to baseline methods.

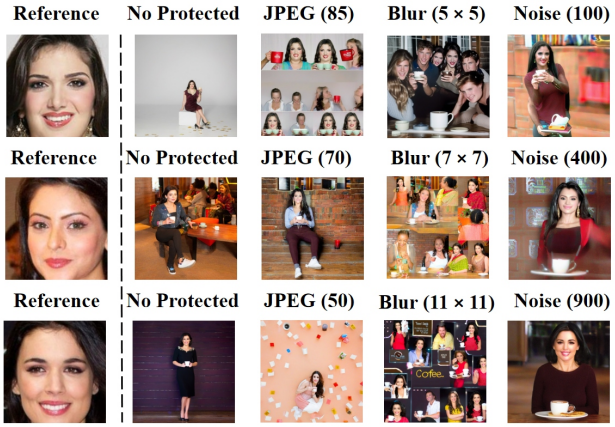

Figure 2. Visual results under three levels of JPEG compression, Gaussian blur, and Gaussian noise, showing that our protection mechanism remains effective even under stronger distortions.

B.2. Ablation Study on Identity Similarity and Perceptual Quality Metrics

We further expand the ablation study presented in the main paper by showing the full set of identity similarity metrics and perceptual quality metrics across various ablation configurations. Tables 5 and 6 provide detailed insights into the contributions of different identity encoder combinations and adversarial conceptual bridge. These results validate the complementary roles of identity supervision and conceptual bridge in enhancing identity suppression, where visual degradation is part of the intended defense effect.

B.3. Robustness Evaluation Settings.

To evaluate the robustness of our protection mechanism, we applied three types of distortions to the protected images: JPEG compression, Gaussian blur, and Gaussian noise. Specifically, we used JPEG compression with a quality factor of 85, Gaussian blur with a kernel size of 5×5 and $\sigma = 1.5$, and noise perturbation with a variance of 100. These parameters were selected to simulate common post-processing or transmission degradations while maintaining perceptual plausibility. As illustrated in Figure 5 of the main text, the protection performance under these distortions remains comparable to that on undistorted inputs, demonstrating that our method exhibits strong robustness against typical image degradations. To further evaluate the robustness, we conducted additional experiments with stronger distortions. Specifically, we tested two additional levels of perturbations: JPEG compression with quality factors of 70 and 50, Gaussian blur with kernel sizes of 7×7 and 11×11 (both with $\sigma = 1.5$), and Gaussian noise with variances of 400 and 900. Table 1 summarizes the results, demonstrating that our method maintains protection performance even under these more severe image degradations.

Table 1. Robustness evaluation under different distortion levels. The results are reported on Ip-Adapter Plus XL as an example.

No Protected			**JPEG (85)**			**Blur (5 × 5, 1.5)**			**Noise (100)**		
Arc ↓	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**
0.4985	0.6899	0.5573	0.3209	0.3983	0.3006	0.3317	0.4443	0.3520	0.3078	0.3970	0.3113
No Protected			**JPEG (70)**			**Blur (7 × 7, 1.5)**			**Noise (400)**		
Arc ↓	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**
0.4985	0.6899	0.5573	0.3520	0.4863	0.3730	0.3303	0.4473	0.3338	0.3955	0.4778	0.3746
No Protected			**JPEG (50)**			**Blur (11 × 11, 1.5)**			**Noise (900)**		
Arc ↓	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**
0.4985	0.6899	0.5573	0.3426	0.4606	0.3382	0.3416	0.4728	0.3632	0.3094	0.4495	0.3604

As illustrated in Figure 2, we also provide visual comparisons under three levels of post-processing strength, showing that the perceptual quality of the protected images remains largely preserved despite the increased distortions.

B.4. Robustness Evaluation on Identity Similarity and Perceptual Quality Metrics

Robustness experiments are conducted under both common image degradations (JPEG compression, Gaussian blur, and additive noise) and adaptive defense-aware attacks such as Impress, which is explicitly designed to circumvent identity protection mechanisms. As shown in Tables 7 and 8, IDGuardian maintains low identity similarity scores across all perturbation types, while perceptual quality metrics naturally decline under heavy input distortion. These results demonstrate that our method retains strong protection capability even under adverse conditions. Qualitative examples in Figure 8 further support these findings by illustrating that identity-relevant features remain suppressed even under various post-processing transformations. These detailed robustness results complement the summary metrics presented in the main paper and underscore IDGuardian's practical effectiveness in real-world usage scenarios.

C. Additional Ablation Study

C.1. Visualization of the Main Ablation Results

To further validate the effectiveness of our proposed identity guidance strategy and loss function design, we provide an additional visualization of the ablation results in Figure 9. The figure clearly demonstrates the influence of different identity loss combinations and guidance directions on the overall performance. Consistent with the quantitative results in Table 4 of the main text, IDGuardian achieves superior identity protection while maintaining high visual quality.

C.2. Perturbation Budget and Learning Rate

We perform a grid search over the perturbation magnitude ϵ and learning rate α to determine an optimal trade-off between identity consistency suppression and perceptual quality. Specifically, we evaluate the following five configurations:

- $\alpha = 0.005$, $\epsilon = 4/255$
- $\alpha = 0.005$, $\epsilon = 8/255$
- $\alpha = 0.005$, $\epsilon = 16/255$
- $\alpha = 0.001$, $\epsilon = 8/255$
- $\alpha = 0.01$, $\epsilon = 8/255$

As depicted in Figure 3, the configuration $\alpha = 0.005$, $\epsilon = 8/255$ achieves the best balance:

- It consistently reduces identity consistency below 0.15 across ArcFace, FaceNet, and VGGFace.
- It maintains high perceptual quality, with SSIM exceeding 0.8419 and PSNR surpassing 32.186 dB.

Smaller perturbations (e.g., $\epsilon = 4/255$) yield insufficient identity suppression, while larger ones (e.g., $\epsilon = 16/255$) compromise visual fidelity. Similarly, extremely small learning rates ($\alpha = 0.001$) lead to under-optimization, whereas large values ($\alpha = 0.01$) introduce instability and visual artifacts. These observations underscore the importance of jointly tuning the perturbation magnitude and step size, as a delicate balance is required to ensure both identity confusion and natural-looking outputs.

C.3. Number of Optimization Steps

We investigate the convergence behavior of identity loss over a range of optimization steps, up to 1000 iterations. As shown in Figure 4, the identity loss declines rapidly during the initial 100 steps, indicating early-stage effectiveness in suppressing identity-related features. To illustrate optimization dynamics more clearly, we apply a shaded band using

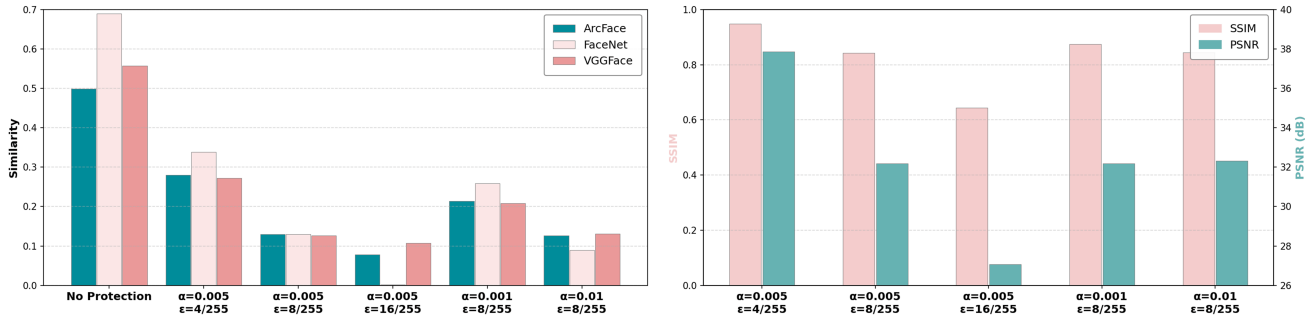

Figure 3. Grid search results for perturbation budget ϵ and learning rate α. The configuration $\alpha = 0.005$, $\epsilon = 8/255$ achieves the best trade-off between identity suppression (measured by ArcFace, FaceNet, and VGGFace similarity) and perceptual quality (SSIM and PSNR).

Table 2. Identity Suppression Performance Across Surrogate Generative Models. The metrics Arc, Face, and VGG correspond to identity similarity scores computed by the face recognition models ArcFace, FaceNet, and VGGFace, respectively.

Method	Ip-Adapter			Ip-Adapter Plus XL			Blip-Diffusion			InfiniteID		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3369	0.4774	0.3308	0.4985	0.6899	0.5573	0.3305	0.3918	0.2848	0.6345	0.7627	0.6286
IDGuardian-SD1.5	**0.0425**	**0.0119**	**0.0650**	**0.1589**	**0.1413**	**0.1403**	**0.2670**	**0.2550**	**0.2204**	**0.4338**	**0.4989**	**0.4113**
Method	**Photomaker**			**Dreambooth**			**InfiniteYou**			**Mean**		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3488	0.3513	0.2806	0.4681	0.6580	0.4477	0.6788	0.7554	0.5847	0.4708	0.5838	0.4450
IDGuardian-SD1.5	**0.1353**	**0.1556**	**0.2233**	**0.2907**	**0.3940**	**0.3455**	**0.3328**	**0.3573**	**0.2809**	**0.2373**	**0.2591**	**0.2409**

`fill_between`[1] to visualize artificial fluctuation bounds around the loss curve. This is not a confidence interval but a heuristic indicator of stability and oscillation patterns.

The loss saturates around 200 iterations, beyond which further gains become marginal. Based on this trend, we adopt $N_{\text{iter}} = 200$ as the default number of steps in our experiments, ensuring adequate convergence without excessive computation.

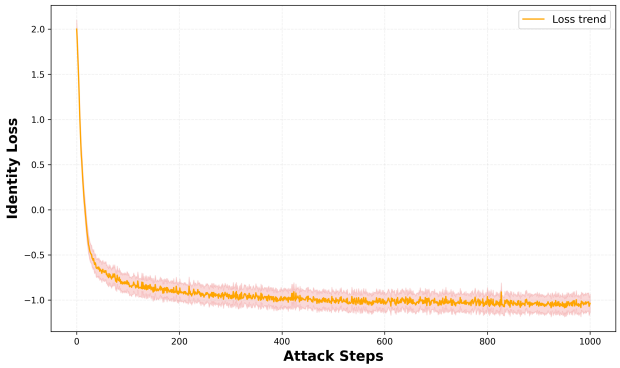

Figure 4. Identity loss convergence over optimization steps with shaded band indicating fluctuation bounds (± 0.1) around the mean curve. The identity loss stabilizes after approximately 200 steps, guiding the selection of the default iteration count.

D. Generalization Analysis

D.1. Evaluation on Commercial Face Personalization APIs

To further validate the generalization ability of our protection method, we conducted additional evaluations on several commercial face personalization APIs, including ***Seedream 4.0***, ***Owen-Image***, ***Kling-AI***, and ***Ihuiwa***. As shown in Figure 5, our method still provides a certain degree of defense under these black-box conditions, demonstrating its robustness and transferability beyond the training environ-

[1] `fill_between` is a Matplotlib function that fills the area between two curves, allowing visual emphasis on regions such as fluctuation bounds. In this context, it is used to highlight the range of ±0.1 around the loss curve to help visualize the stability and oscillation of the optimization process.

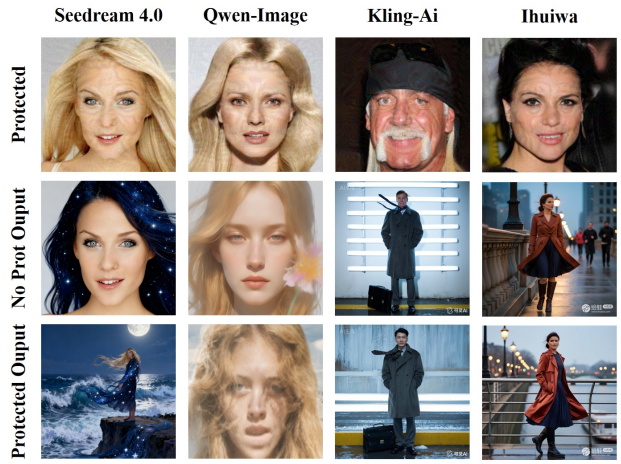

Figure 5. Visualization of generalization results on commercial face personalization APIs.

ment.

D.2. Flexibility in Surrogate Selection

To evaluate the generalizability of our approach at the model level, we replace the surrogate model used for adversarial optimization with a significantly different generative backbone. Our main experiments utilize the IP-Adapter SDXL Plus Face model based on Stable Diffusion XL (SDXL). Additionally, we assess our method on the earlier Stable Diffusion 1.5 (SD1.5) backbone along with its corresponding IP-Adapter variant. All attack configurations, including text prompts and optimization parameters, are held constant to ensure comparability.

The results demonstrate that the proposed method consistently achieves strong identity suppression across distinct surrogate models, confirming its robustness and applicability beyond specific model architectures, as summarized in Table 2. Representative visual results further illustrate the consistency and effectiveness of identity suppression across models, as shown in Figure 6.

D.3. Dataset Generalization

In addition to the VGGFace2 dataset results presented in the main paper, we further evaluate the method on the high-resolution CelebA-HQ dataset. We employ multiple identity similarity metrics, namely ArcFace, FaceNet, and VGGFace, to measure the extent of identity removal across diverse personalized generation models. Perceptual quality is concurrently assessed through FaceQnet and MagFace to verify preservation of visual fidelity.

The quantitative results summarized in Table 9 and Table 10 indicate that the method maintains effective identity suppression and reduced perceptual quality on CelebA-HQ, thus demonstrating robustness across datasets varying in resolution and demographic composition. As illustrated in Figure 10, the visual results confirm that the method successfully removes identity-specific features while preserving photorealism, supporting the effectiveness of the proposed defense. We further provide results on the LFW dataset using the identity similarity metric in Table 11. The results remain consistent with those on VGGFace2 and CelebA-HQ, further demonstrating the robustness and cross-dataset generalization ability of IDGuardian. Collectively, these experiments verify that IDGuardian consistently preserves identity suppression capability and visual quality across different surrogate model architectures and datasets, highlighting its strong generalization capability in diffusion-based personalized face generation.

E. Human and GPT-Based Evaluation of Identity Protection Effectiveness

As illustrated in Figure 7, the upper part of the figure presents the results of the human perceptual study, while the lower part corresponds to the GPT-based evaluation. This joint visualization highlights the consistency between subjective human judgments and automated assessments, with our method favored in 92% of human responses and 94% of GPT comparisons across over 50 test cases, further reinforcing the effectiveness of our approach.

E.1. Human Study

To further assess the effectiveness of identity protection from a human perceptual perspective, we conducted a comprehensive subjective evaluation involving human participants. These participants were tasked with rating the perceived identity similarity between the original identity images and the personalized images generated by our method as well as several state-of-the-art baseline methods. The results consistently demonstrate that images produced by our approach exhibit significantly lower perceived identity similarity compared to those generated by other methods, thereby indicating superior identity obfuscation. This extensive subjective evaluation complements the quantitative metrics by capturing nuanced perceptual differences and providing insights into human visual cognition that automated measures alone may not fully capture.

E.2. GPT-Based Evaluation

In addition to the human evaluation, we also employed a state-of-the-art large multimodal language model, GPT, to automatically assess the identity protection performance. Given pairs of original and generated images, GPT was prompted to analyze and identify which generated image exhibits the greatest discrepancy in identity features relative to the original input. Leveraging GPT's advanced visual-language reasoning capabilities, this automated evaluation

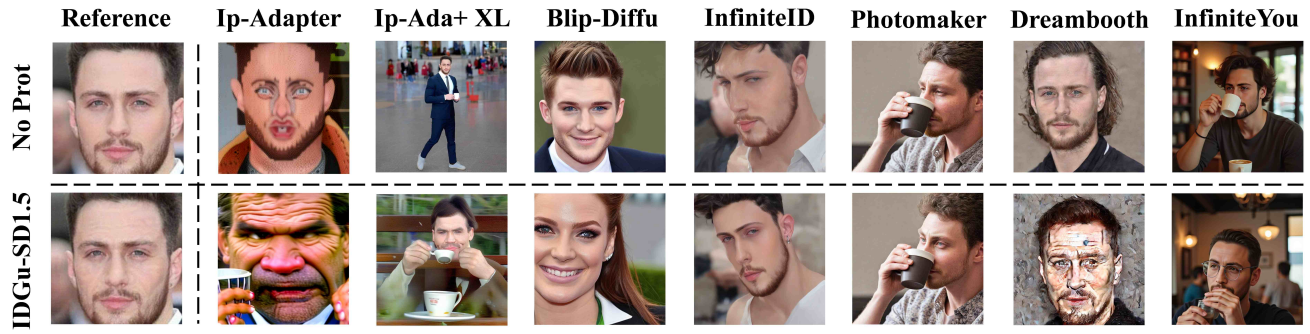

Figure 6. Visual comparison of identity suppression results across different surrogate generative models.

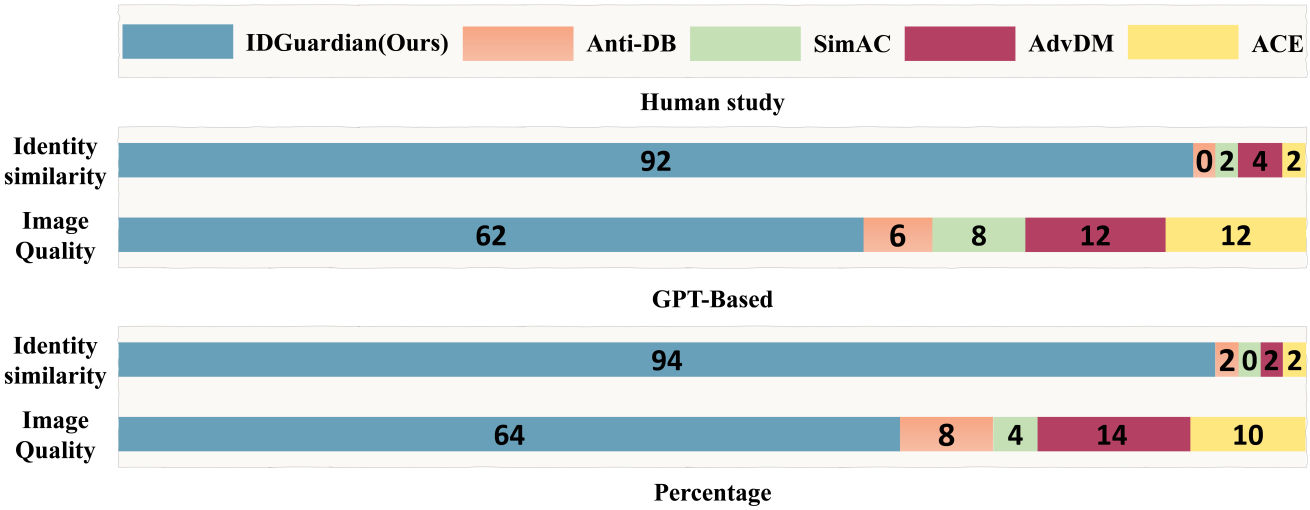

Figure 7. Comparison of Identity Protection Effectiveness Based on Human Perception and GPT-Based Evaluation.

approach offers a scalable and efficient complement to human judgment. The findings from the GPT-based evaluation are consistent with the results obtained from the user study, thereby further validating the robustness and effectiveness of our method in substantially reducing identity similarity across different evaluation paradigms.

E.3. Ethics Statement

Our user study involved voluntary participation of adult subjects recruited within our institution. Participants were asked to provide subjective ratings of visual similarity without disclosing or providing any personally identifiable or sensitive information. The study used publicly available face datasets and collected no new identifiable personal data. All user evaluations were performed on synthetic or anonymized data, and no sensitive or personally identifiable information was recorded. The study posed no potential harm or risk to participants. While formal Institutional Review Board (IRB) approval was not required under our institution's policy for this low-risk, anonymized perceptual evaluation, all participants were informed about the study's purpose and provided consent prior to participation. The study was conducted in accordance with recognized ethical research principles.

Table 3. Comparison of identity protection performance against baseline methods. Evaluation is conducted using three different identity similarity metrics: ArcFace, FaceNet, and VGGFace. "No Protected" denotes generation without defense.

Method	Ip-Adapter			Ip-Adapter Plus XL			Blip-Diffusion			InfiniteID		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3369	0.4774	0.3308	0.4985	0.6899	0.5573	0.3305	0.3918	0.2848	0.6345	0.7627	0.6286
Anti-DB	0.2525	0.3964	0.2715	0.4815	0.6419	0.5219	0.2685	0.3153	0.2286	0.5897	0.6859	0.5848
SimAC	0.2048	0.4005	0.2792	0.3454	0.4634	0.3956	0.3238	0.3344	0.2563	0.5472	0.6670	0.5599
AdvDM	0.2409	0.3624	0.2535	0.3427	0.4810	0.3700	0.2461	0.2989	0.2363	0.5069	0.6107	0.5087
ACE	0.2573	0.4249	0.2996	0.4615	0.6377	0.5079	0.2738	0.3160	0.2336	0.5419	0.6489	0.5383
IDProtector	0.2059	0.4167	0.2657	0.3324	0.5297	0.4120	0.2565	0.2967	0.2346	0.4972	0.7009	0.6112
IDGuar (Ours)	**0.0673**	**-0.0226**	**0.0640**	**0.1296**	**0.1231**	**0.1260**	**0.2255**	**0.2376**	**0.1964**	**0.4121**	**0.5215**	**0.3990**

Method	Photomaker			Dreambooth			InfiniteYou			Mean		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3488	0.3513	0.2806	0.4681	0.6580	0.4477	0.6788	0.7554	0.5847	0.4709	0.5838	0.4449
Anti-DB	0.2331	0.3104	0.2716	0.3590	0.5420	0.3976	0.5867	0.6569	0.5343	0.3959	0.5070	0.4014
SimAC	0.2267	0.2614	0.2564	0.2987	0.4760	0.3942	0.5604	0.6472	0.5039	0.3581	0.4643	0.3780
AdvDM	0.2083	0.2189	0.2477	**0.2191**	0.4507	**0.3327**	0.5546	0.6309	0.4905	0.3312	0.4362	0.3485
ACE	0.2246	0.2545	0.2756	0.3252	0.5328	0.4033	0.6450	0.7207	0.5504	0.3900	0.5051	0.4012
IDProtector	0.2170	0.2212	0.2807	0.3372	0.4768	0.3932	0.5047	0.6067	0.5105	0.3358	0.4641	0.3868
IDGuar (Ours)	**0.0747**	**0.1284**	**0.2239**	0.2792	**0.3893**	0.3404	**0.3230**	**0.3265**	**0.2522**	**0.2159**	**0.2434**	**0.2288**

Table 4. Comparison of identity protection performance against baseline methods. Evaluation is conducted using FaceQNet and MagFace quality scores. "No Protected" denotes generation without defense.

Method	Ip-Adapter		Ip-Adapter Plus XL		Blip-Diffusion		InfiniteID	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.3339	21.3847	0.4564	19.3085	0.3305	21.8112	0.4946	19.0747
Anti-DB	0.3045	21.1209	0.4553	**18.4376**	0.3565	20.8991	0.4957	18.9697
SimAC	0.3124	20.8217	0.4492	18.9702	0.3628	21.4539	**0.4844**	19.0755
AdvDM	0.3486	20.6235	**0.4379**	18.8501	0.3547	21.6632	0.4896	18.9385
ACE	0.3627	20.6235	0.4528	18.5293	0.3630	21.7078	0.4994	19.0961
IDProtector	0.3367	**19.9330**	0.4318	18.8869	0.3680	21.3936	0.4885	18.9833
IDGuar (Ours)	**0.2820**	20.8712	0.4503	18.6099	**0.3511**	**20.8498**	0.4938	**18.9031**

Method	Photomaker		Dreambooth		InfiniteYou		Mean	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.4564	20.0541	0.3316	20.2766	0.4927	20.7298	0.4137	20.3771
Anti-DB	0.2883	20.0427	0.3186	20.2498	0.4666	19.9501	0.3836	19.9529
SimAC	0.2959	19.9726	0.3427	20.1656	0.4590	19.9077	0.3866	20.0525
AdvDM	0.2934	20.1307	**0.1921**	**19.7116**	0.4586	19.9321	0.3678	19.9785
ACE	0.2866	20.1201	0.2147	20.1827	0.4610	19.8904	0.3772	19.8786
IDProtector	0.3119	28.5475	0.2248	20.6773	0.4535	19.5957	0.3736	21.1453
IDGuar (Ours)	**0.2733**	**19.8202**	0.2203	20.2199	**0.4517**	**19.8111**	**0.3604**	**19.8693**

Table 5. Ablation study of identity loss combinations and guidance directions, where **G adv** denotes guidance solely toward the adversarial identity distribution and **G clean** denotes guidance solely away from the clean identity distribution.

Method	Ip-Adapter			Ip-Adapter Plus XL			Blip-Diffusion			InfiniteID		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3369	0.4774	0.3308	0.4985	0.6899	0.5573	0.3305	0.3918	0.2848	0.6345	0.7627	0.6286
CLIP	0.2289	0.3058	0.1949	0.2032	0.2328	0.2147	0.2864	0.3078	0.2664	0.5048	0.6775	0.5478
Facenet	0.2412	0.2897	0.2058	0.2145	0.2281	0.2223	0.2948	0.2989	0.2716	0.5126	0.6843	0.5534
CLIP + Facenet	0.1578	0.2347	0.1326	0.1665	0.1707	0.1638	0.2534	0.2764	0.2355	0.4932	0.6047	0.5089
IDGuar (Ours)	**0.0673**	**-0.0226**	**0.0640**	**0.1296**	**0.1231**	**0.1260**	**0.2255**	**0.2376**	**0.1964**	**0.4121**	**0.5215**	**0.3990**
G adv	0.0688	-0.0006	0.0650	0.1318	0.1476	0.1536	0.2971	0.3072	0.2509	0.4501	0.5369	0.4181
G clean	0.0869	-0.0061	0.0668	0.1442	0.1274	0.1432	0.2400	0.2751	0.2029	0.4265	0.5261	0.4073
IDGuar (Ours)	**0.0673**	**-0.0226**	**0.0640**	**0.1296**	**0.1231**	**0.1260**	**0.2255**	**0.2376**	**0.1964**	**0.4121**	**0.5215**	**0.3990**

Method	Photomaker			Dreambooth			InfiniteYou			Mean		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3488	0.3513	0.2806	0.4681	0.6580	0.4477	0.6788	0.7554	0.5847	0.4709	0.5838	0.4449
CLIP	0.2465	0.2696	0.2757	0.4634	0.5890	0.4584	0.5273	0.6043	0.4599	0.3458	0.4210	0.3397
Facenet	0.2593	0.2634	0.2810	0.4721	0.5743	0.4628	0.5345	0.5926	0.4658	0.3612	0.4188	0.3518
CLIP + Facenet	0.2056	0.2142	0.2202	0.4577	0.5606	0.4726	0.5032	0.5357	0.3932	0.3063	0.3577	0.2905
IDGuar (Ours)	**0.0747**	**0.1284**	**0.2239**	**0.2792**	**0.3893**	**0.3404**	**0.3230**	**0.3265**	**0.2522**	**0.2159**	**0.2434**	**0.2288**
G adv	0.1437	0.1477	0.2324	0.2796	0.4187	0.3432	0.3279	0.3405	0.2526	0.2427	0.2711	0.2451
G clean	0.1562	0.1974	0.2333	0.3046	0.4813	0.3410	0.3379	0.3472	0.2565	0.2423	0.2783	0.2358
IDGuar (Ours)	**0.0747**	**0.1284**	**0.2239**	**0.2792**	**0.3893**	**0.3404**	**0.3230**	**0.3265**	**0.2522**	**0.2159**	**0.2434**	**0.2288**

Table 6. Ablation study of different identity loss combinations and guidance directions. Evaluation is conducted using FaceQNet and MagFace to measure the perceptual quality of generated images.

Method	Ip-Adapter		Ip-Adapter Plus XL		Blip-Diffusion		InfiniteID	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.3339	21.3847	0.4564	19.3085	0.3305	21.8112	0.4946	19.0747
CLIP	0.3129	21.7832	0.4726	20.4175	0.3683	20.9487	0.5142	19.6254
Facenet	0.3185	21.6910	0.4791	20.5083	0.3724	**20.0284**	0.5076	19.4939
CLIP + Facenet	**0.2714**	21.8426	0.4689	19.5268	0.3776	20.4123	0.5128	20.1884
IDGuar (Ours)	0.2820	**20.8712**	**0.4503**	**18.6099**	**0.3511**	20.8498	**0.4938**	**18.8031**
G adv	0.2900	**20.8653**	0.4603	19.2300	0.3569	20.7490	**0.4899**	19.0814
G clean	**0.2692**	21.6448	0.4526	19.2488	0.3557	**19.3413**	0.4968	18.8645
IDGuar (Ours)	0.2820	20.8712	**0.4503**	**18.6099**	**0.3511**	20.8498	0.4938	**18.8031**

Method	Photomaker		Dreambooth		InfiniteYou		Mean	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.4564	20.0541	0.3316	20.2766	0.4527	19.7298	0.4083	20.2342
CLIP	0.3012	20.2741	0.3310	**19.5135**	0.4729	**19.1276**	0.3962	20.2414
Facenet	0.3075	20.3417	0.3298	19.8824	0.4660	19.4059	0.3972	20.1930
CLIP + Facenet	0.2916	20.1042	0.2983	19.8631	0.4679	20.0893	0.3768	20.2895
IDGuar (Ours)	**0.2733**	**19.8202**	**0.2203**	20.2199	**0.4517**	19.8111	**0.3604**	**19.8693**
G adv	0.2813	**19.7368**	0.2754	**19.9528**	0.4536	20.2709	0.3725	19.9837
G clean	0.2789	19.8856	0.2270	20.5887	0.4519	20.3764	0.3617	19.9929
IDGuar (Ours)	**0.2733**	19.8202	**0.2203**	20.2199	**0.4517**	**19.8111**	**0.3604**	**19.8693**

Table 7. Robustness evaluation against common image post-processing operations, including JPEG compression, Gaussian blur, and Gaussian noise, and attacks targeting active defenses such as Impress.

Method	Ip-Adapter			Ip-Adapter Plus XL			Blip-Diffusion			InfiniteID		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3369	0.4774	0.3308	0.4985	0.6899	0.5573	0.3305	0.3918	0.2848	0.6345	0.7627	0.6286
JPEG	0.1068	0.1900	0.1263	0.3209	0.3983	0.3006	0.2236	0.2196	0.2026	0.4286	0.5158	0.4062
Blur	0.1678	0.2948	0.1967	0.3317	0.4443	0.3520	0.2663	0.2663	0.2077	**0.3854**	**0.4780**	**0.3734**
Noise	0.1365	0.1794	0.1295	0.3078	0.3970	0.3113	**0.1916**	**0.1910**	**0.1716**	0.4482	0.5581	0.4391
Impress	0.1018	0.1696	0.1230	0.3184	0.3735	0.2956	0.2261	0.2653	0.2277	0.4478	0.5309	0.4074
IDGuar (Ours)	**0.0673**	**-0.0226**	**0.0640**	**0.1296**	**0.1231**	**0.1260**	0.2255	0.2376	0.1964	0.4121	0.5215	0.3990

Method	Photomaker			Dreambooth			InfiniteYou			Mean		
	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓	Arc ↓	Face ↓	VGG ↓
No Protected	0.3488	0.3513	0.2806	0.4681	0.6580	0.4477	0.6788	0.7554	0.5847	0.4709	0.5838	0.4449
JPEG	0.1199	**0.0754**	0.2082	0.2723	0.3963	**0.3267**	0.3344	0.3728	0.2590	0.2581	0.3097	0.2614
Blur	**0.0709**	0.1591	0.2102	0.2899	0.5136	0.4110	0.3371	0.3510	0.2587	0.2642	0.3582	0.2871
Noise	0.1785	0.1618	**0.1971**	0.3418	0.5177	0.3844	0.3410	0.3625	0.2623	0.2780	0.3382	0.2708
Impress	0.1910	0.1494	0.2097	**0.2516**	0.4479	0.3386	**0.3166**	**0.3233**	**0.2387**	0.2648	0.3229	0.2629
IDGuar (Ours)	0.0747	0.1284	0.2239	0.2792	**0.3893**	0.3404	0.3230	0.3265	0.2522	**0.2159**	**0.2434**	**0.2288**

Table 8. Robustness evaluation against common image post-processing operations. Evaluation is conducted using FaceQNet and MagFace to measure the perceptual quality of generated images.

Method	Ip-Adapter		Ip-Adapter Plus XL		Blip-Diffusion		InfiniteID	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.3339	21.3847	0.4564	19.3085	0.3305	21.8112	0.4946	19.0747
JPEG	**0.2211**	21.3679	0.4539	18.5163	0.3233	21.7881	0.4918	**18.8227**
Blur	0.3211	21.3901	0.4667	19.7750	**0.2895**	**19.8607**	0.5031	18.8229
Noise	0.2499	**20.6822**	0.4497	**18.4623**	0.3460	20.6286	0.5016	19.0230
Impress	0.2501	20.7604	**0.4270**	19.6015	0.3353	20.9642	0.4935	19.1813
IDGuar (Ours)	0.2820	20.8712	0.4503	18.5099	0.3511	20.8498	**0.4908**	18.9031

Method	Photomaker		Dreambooth		InfiniteYou		Mean	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.4564	20.0541	0.3316	20.2766	0.4527	19.7298	0.4080	20.2342
JPEG	0.2743	**19.6998**	0.2535	20.5918	0.4544	20.0803	**0.3532**	19.9809
Blur	0.2754	19.8549	0.2814	19.6256	**0.4483**	20.0016	0.3694	19.9044
Noise	0.2809	19.8303	0.3583	**19.2192**	0.4528	19.8165	0.3770	**19.6660**
Impress	0.2762	20.0743	0.2475	21.0550	0.4536	**19.6948**	0.3547	20.1902
IDGuar (Ours)	**0.2733**	19.8202	**0.2203**	20.2199	0.4517	19.8111	0.3604	19.8693

Table 9. Comparison of identity protection performance against baseline methods on the CelebA-HQ dataset. Evaluation is conducted using three widely adopted identity similarity metrics: ArcFace, FaceNet, and VGGFace.

Method	Ip-Adapter			Ip-Adapter Plus XL			Blip-Diffusion			InfiniteID		
	Arc ↓	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**
No Protected	0.4104	0.4714	0.3764	0.5706	0.7152	0.6062	0.2862	0.3495	0.3061	0.6352	0.6886	0.5642
Anti-DB	0.3061	0.3647	0.3193	0.4940	0.6321	0.5260	0.2257	0.3369	0.2894	0.5344	0.5946	0.5031
SimAC	0.2398	0.3390	0.3021	0.4049	0.5131	0.4339	0.2499	0.3240	0.2849	0.4483	0.5262	0.4455
AdvDM	0.2827	0.3907	0.3125	0.4158	0.5007	0.4464	0.2222	0.3484	0.2680	0.5056	0.5435	0.4669
ACE	0.3462	0.4354	0.3505	0.5412	0.6673	0.5834	0.2418	0.3422	0.2880	0.4813	0.5480	0.4839
IDProtector	0.2543	0.3071	0.2922	0.4363	0.5421	0.4486	0.1881	**0.3003**	**0.2256**	0.5065	0.6161	0.5611
IDGuar (Ours)	**0.1003**	**0.0515**	**0.0940**	**0.2686**	**0.2905**	**0.2385**	**0.1794**	0.3061	0.2431	**0.3469**	**0.4198**	**0.3489**

Method	Photomaker			Dreambooth			InfiniteYou			Mean		
	Arc ↓	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**	**Arc ↓**	**Face ↓**	**VGG ↓**
No Protected	0.6934	0.4219	0.2612	0.5421	0.6555	0.4335	0.6614	0.7485	0.5640	0.5428	0.5787	0.4445
Anti-DB	0.6504	0.4065	0.2560	0.1031	0.0666	0.0761	0.6035	0.6869	0.5312	0.4167	0.4412	0.3573
SimAC	0.6251	0.3564	0.2321	**0.0831**	0.0852	0.0665	0.5854	0.6846	0.4654	0.3766	0.4041	0.3186
AdvDM	0.6313	0.3718	0.2672	0.0868	0.0365	0.1100	0.5207	0.6293	0.4654	0.3807	0.4030	0.3338
ACE	0.6422	0.3880	0.2384	0.1401	0.2271	0.0944	0.6358	0.7283	0.5560	0.4327	0.4766	0.3707
IDProtector	0.6151	0.3013	0.2022	0.3298	0.3727	0.2578	0.5617	0.6609	0.4943	0.4131	0.4429	0.3545
IDGuar (Ours)	**0.3320**	**0.2654**	**0.1996**	0.0838	**0.0320**	**0.0610**	**0.3331**	**0.3871**	**0.2428**	**0.2349**	**0.2503**	**0.2040**

Table 10. Comparison of identity protection performance against baseline methods on the CelebA-HQ dataset. Evaluation is conducted using FaceQNet and MagFace to measure the perceptual quality of generated images.

Method	Ip-Adapter		Ip-Adapter Plus XL		Blip-Diffusion		InfiniteID	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.3031	20.6785	0.4490	18.8023	0.3344	20.9884	0.4527	19.6607
Anti-DB	0.2599	**19.5696**	0.4515	**18.5281**	0.3170	21.4418	0.4731	18.9686
SimAC	0.2670	19.9800	0.4493	19.2084	0.3433	**19.4777**	0.4695	19.0426
AdvDM	0.2842	19.9800	0.4505	19.2084	**0.2941**	20.0386	0.4645	19.2935
ACE	0.2928	20.0025	0.4540	19.3925	0.3363	20.7060	0.4539	19.1724
IDProtector	0.2945	20.0540	0.4544	19.0985	0.3039	20.4320	**0.4204**	19.7844
IDGuar (Ours)	**0.2556**	19.9783	**0.4490**	18.7536	0.3098	20.6180	0.4574	**18.7834**
Method	**Photomaker**		**Dreambooth**		**InfiniteYou**		**Mean**	
	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓	FaceQNet ↓	MagFace ↓
No Protected	0.2674	20.5900	0.3222	21.1998	0.4463	19.4956	0.3679	20.2022
Anti-DB	0.2693	**20.4309**	0.2802	21.8453	0.4451	19.6502	0.3566	20.0621
SimAC	0.2693	20.5074	0.2650	21.6122	0.4489	19.7678	0.3589	19.9423
AdvDM	0.2672	20.5074	0.3072	21.3755	0.4418	19.7678	0.3585	20.0245
ACE	0.2677	20.7290	**0.2501**	21.5659	0.4463	19.4555	0.3573	20.1463
IDProtector	0.2907	28.8511	0.3740	20.0474	0.4565	21.7575	0.3706	21.4321
IDGuar (Ours)	**0.2653**	20.5549	0.3073	**21.2664**	**0.4352**	**19.3972**	**0.3542**	**19.9074**

Table 11. ArcFace-only identity similarity on LFW. Method names are abbreviated for brevity.

Method	IP-Adapter↓	IP-Ada. + XL↓	BLIP-D.↓	InfiniteID↓	PhotoM.↓	DreamB.↓	InfiniteYou↓	Mean↓
No Protected	0.3580	0.5210	0.3010	0.6150	0.3420	0.4550	0.6670	0.4656
Anti-DB	0.2550	0.4890	0.2490	0.5720	0.2300	0.3380	0.5980	0.3901
SimAC	0.2070	0.3620	0.2760	0.5490	0.2200	0.2870	0.5570	0.3511
AdvDM	0.2320	0.3480	0.2410	0.5030	0.2050	0.2540	0.5410	0.3320
ACE	0.2480	0.4620	0.2610	0.5330	0.2270	0.3210	0.6400	0.3846
IDProtector	0.2130	0.3310	0.2380	0.4920	0.2180	0.3160	0.5090	0.3310
IDGuar (Ours)	**0.0790**	**0.1360**	**0.1980**	**0.4050**	**0.0910**	**0.2520**	**0.3180**	**0.2113**

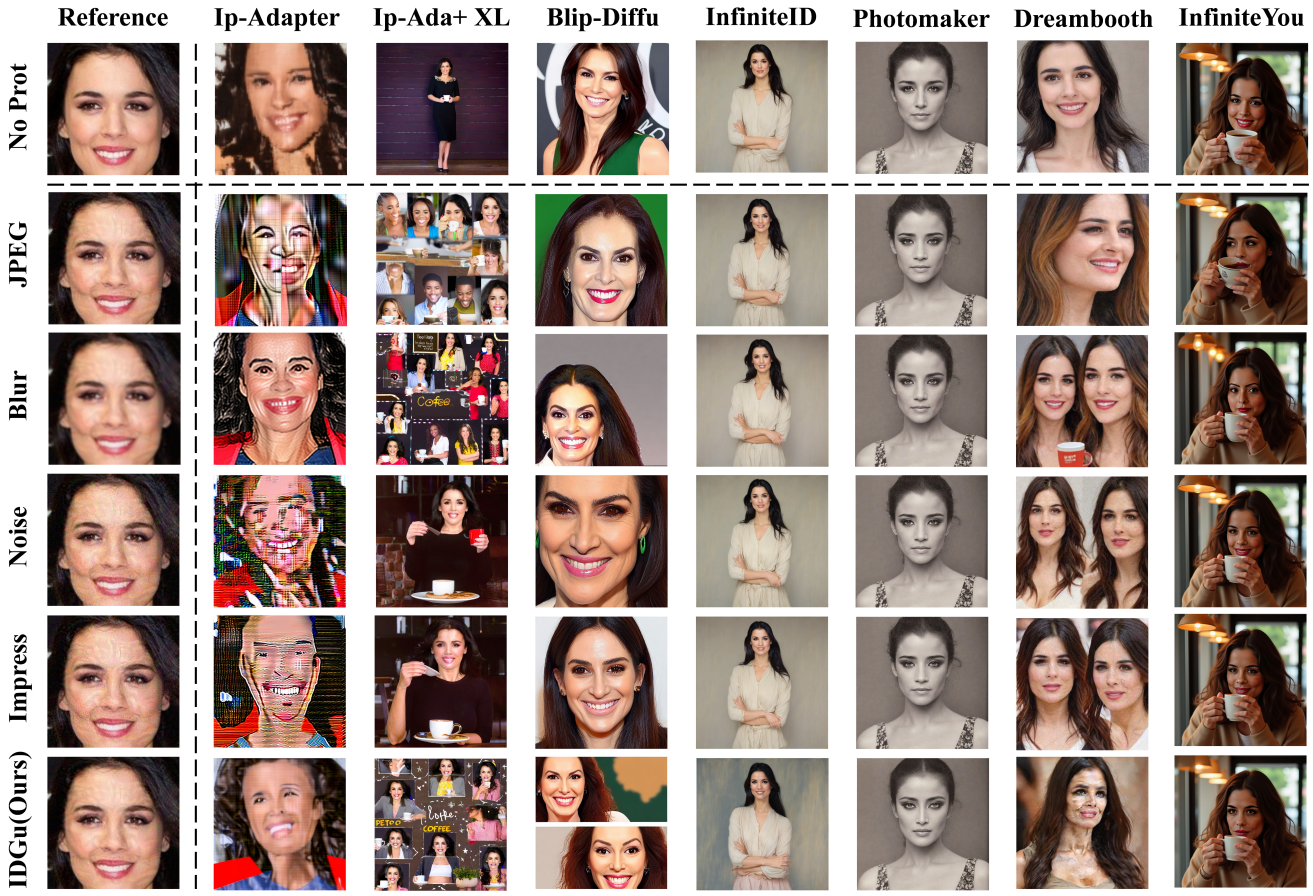

Figure 8. Robustness Evaluation Against Image Post-Processing and Defense Attacks.

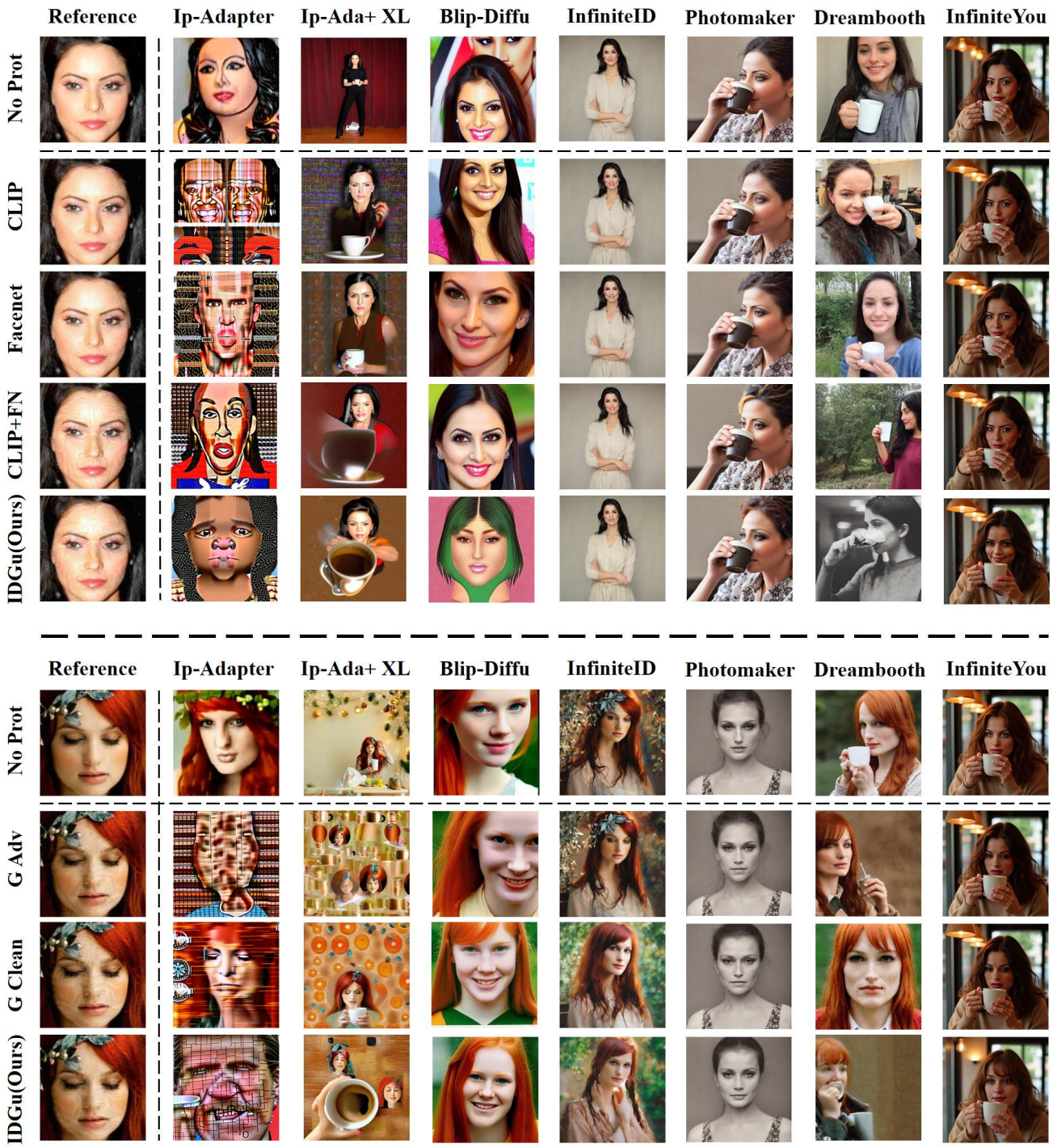

Figure 9. Visualization of the ablation study on identity loss combinations and guidance directions.

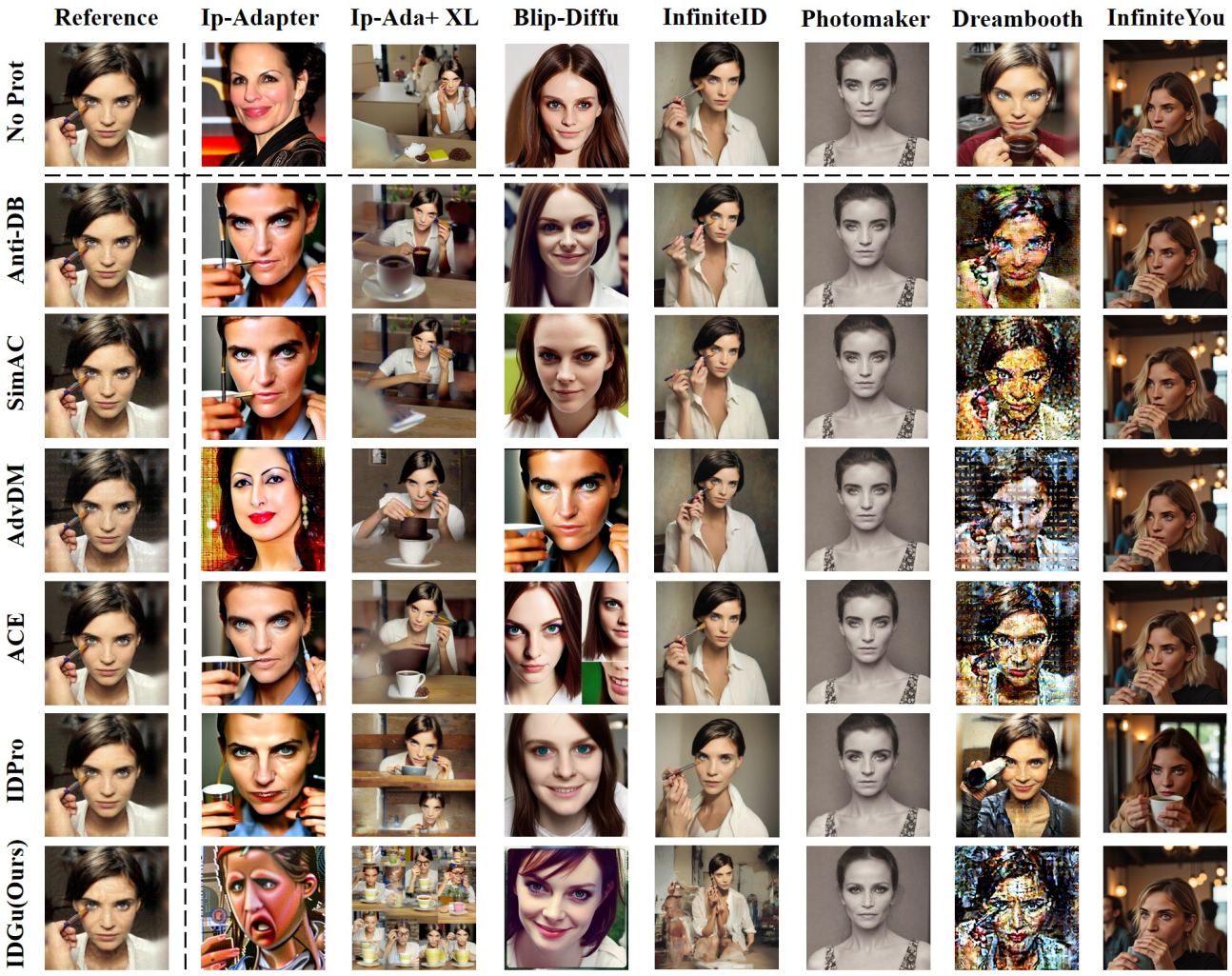

Figure 10. Visualization of Identity Suppression Effectiveness on CelebA-HQ Dataset for IDGuardian and Comparison Methods.