Anti-Tamper Protection for Unauthorized Individual Image Generation

Supplementary Material

In the appendix, we will include more experimental results and the detailed settings for anti-tamper perturbation (ATP).

1. Details of the Anti-tamper Perturbation

1.1. Hyper-parameter configuration

Experiment Environment. All experiments were conducted on a server equipped with 4 L40S GPUs (each with 48G) and an Intel(R) Xeon(R) Gold 6426Y CPU. The system had 251 GB of RAM. The software environment included Pytorch 2.4.1 running on Ubuntu 22.04.4 LTS, with CUDA 12.3 and cuDNN 9.1.0.70 for GPU acceleration. We didn't do distributed training, so the experiment can be conducted using one GPU.

Authorization Perturbation Hyper-Parameters. The authorization perturbation network is trained on FFHQ for 65,000 steps with a batch size of 8. For the weights of the loss function: $\lambda_{adv} = 1e-3$, $\lambda_{rec} = 0.7$, $\lambda_{reg} = 10$. The length of the authorization message m is 32, and the default mask ratio p is 0.5.

Protection Perturbation Hyper-Parameters. APT can adopt the existing protection design, and different baselines have varying choices for PGD radius and step size. Unlike the baselines, our method performs calculations in the frequency domain, so we did not select the same hyperparameters as the baseline.

Method	CelebA-HQ	VGGFace2
	Radius / Step Size	Radius / Step Size
Anti-DB+ours	5e-2 / 5e-3	250e-3 / 25e-3
AdvDM+ours	1e-1 / 2e-3	250e-3 / 25e-3
CAAT+ours	5e-2 / 5e-3	250e-3 / 25e-3
MetaCloak+ours	150e-3 / 5e-3	200e-3 / 5e-3

Table 1. PGD Radius and Step Size for different methods on CelebA-HQ and VGGFace2.

We observed the loss performance after adapting the baseline to our algorithm and selecting the appropriate PGD radius and step sizes. However, we did not perform detailed hyperparameter tuning experiments, as our main objective was to demonstrate that our method does not degrade the baseline's protection performance.

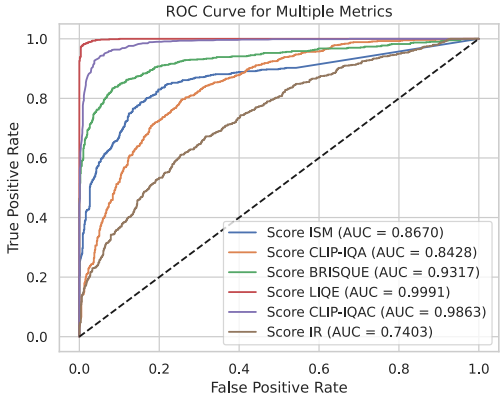


Figure 1. The ROC curve of different metrics.

1.2. Metric Selection

To select suitable metrics for evaluating the protection perturbation, we choose six metrics from the metrics adopted by existing works [2, 3, 7]: ISM [2], CLIP-IQA [5], BRISQUE [4], LIQE [8], CLIP-IQAC [3], IR [6]. When the model can't detect the face, the ISM value is set to -1 to guarantee that all generated images can get a corresponding ISM value. We first generate individual images using unprotected images from CelebA-HQ and those protected with Anti-DB. For each subject, we generate 16 images (50 subjects in total). We then calculate the value of the generated image's six metrics accordingly. We assume that the Anti-DB can often successfully protect the image when attackers do not make purification attempts. As a result, if the metric can classify images generated from protected images and those generated from unprotected images, it should be a reliable metric for evaluating protection performance. Images generated from protected images are categorized as negative samples, while those generated from unprotected images are categorized as positive samples. We then draw the ROC curves of the protection performance metrics, as shown in Figure 1. Among the metrics, CLIP-IQAC and LIQE show the highest AUC values, demonstrating the strongest discriminatory ability. As a result, we adopt them for the **Standard Protection Performance Comparison** in Section 4.1 (FDFR and ISM are also adopted, as ISM is the only metric among them that is directly related to facial identity. Furthermore, FDFR and ISM are typically computed together [2].). For the experiment of the **Protection Performance Under Attack Scenario**, we need to select one metric for calculating the Protection Success Rate

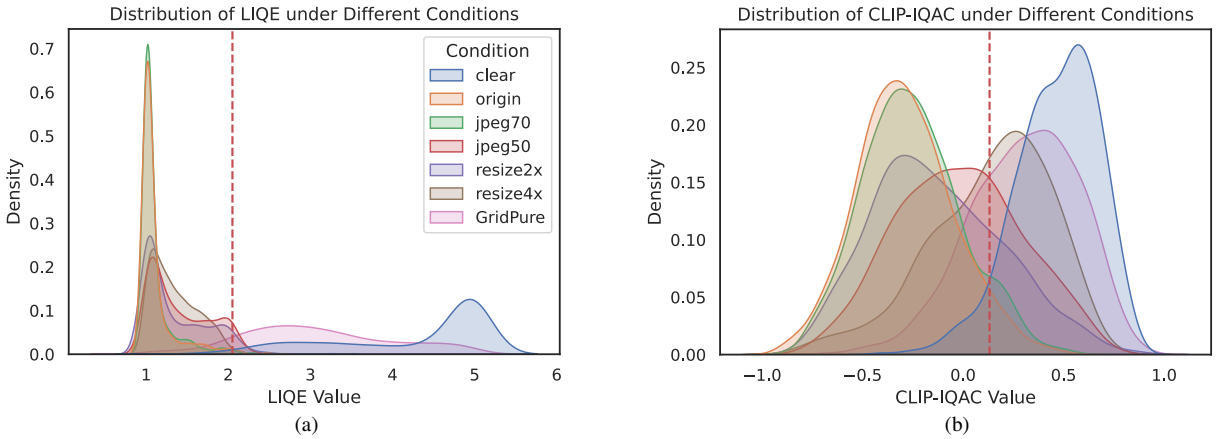


Figure 2. (a) Distributions of generated images evaluated by LIQE metric. (b) Distributions of generated images evaluated by CLIP-IQAC metric. The red dashed line illustrates the PSR threshold.

(PSR). We use the property of the ROC curve to decide the threshold of PSR. We select the threshold that can minimize $\sqrt{(1-TPR)^2+FPR^2}$, where TPR denotes true positive rate and FPR denotes false positive rate. **The threshold for CLIP-IQAC and LIQE are 0.1318359375, 2.05078125, respectively**.

Subsequently, we evaluate the performance of these metrics in capturing the impact of purification attempts on the protection mechanism. The distribution of the metrics for generated images is visualized through kernel density estimation. Specifically, "clear" and "origin" represent the generation results using unprotected and protected images, respectively. At the same time, the remaining categories correspond to the outcomes of applying the respective purification methods to protected images before generation.

As shown in Figure 2, the results demonstrate that CLIP-IQAC and LIQE effectively reflect the influence of purification attempts. Notably, following "resize 4x", "jpeg 50", and "GridPUre", the resulting distributions exhibit a convergence trend toward those of "clear." However, it can be observed that the PSR threshold of LIQE fails to capture the trend, as the majority of the samples fall to the left of the threshold. In contrast, CLIP-IQAC does not exhibit this issue, making it the preferred choice for calculating PSR.

1.3. Threshold Setting

We adopt Anti-DB for ATP to perturb the images in the CelebA-HQ test set. Then, we adopt purification techniques to purify the image. Figure 3 shows distinct differences in bit-error rate with and without purification. Since we aim to detect the occurrence of purification through the bit-error threshold, when the occurrence of purification significantly impacts the distribution of bit-errors, setting the threshold becomes a straightforward task. As a result, **we set the bit-error threshold of PSR to 3/32**. We adopt this value across different datasets and various protection perturbations, consistently finding that it can be effectively used to reject purification attempts.

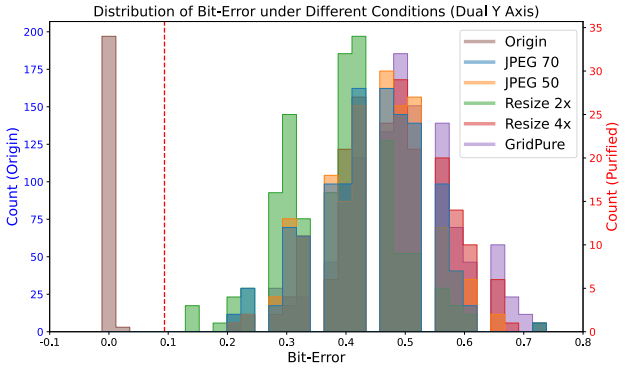


Figure 3. The distribution of bit-error under different purification settings. "Origin" denotes no purification applied. The red dashed line illustrates the PSR bit-error threshold.

1.4. Frequency-domain Sensitivity Analysis

In this section, we analyze why frequency-domain perturbation is inherently more sensitive (i.e., vulnerable) than the pixel-domain perturbation to the purifications. For pixel-domain purification (e.g., resizing), the vulnerability arises because the Block DCT computes each frequency coefficient as a weighted sum of all pixel values in a block. Thus, even a minor modification to a single pixel can affect all frequency coefficients. For frequency-domain purification (e.g., JPEG), the vulnerability stems from the fact that JPEG compression directly quantizes the frequency coefficients. These changes may be smoothed out in the pixel domain

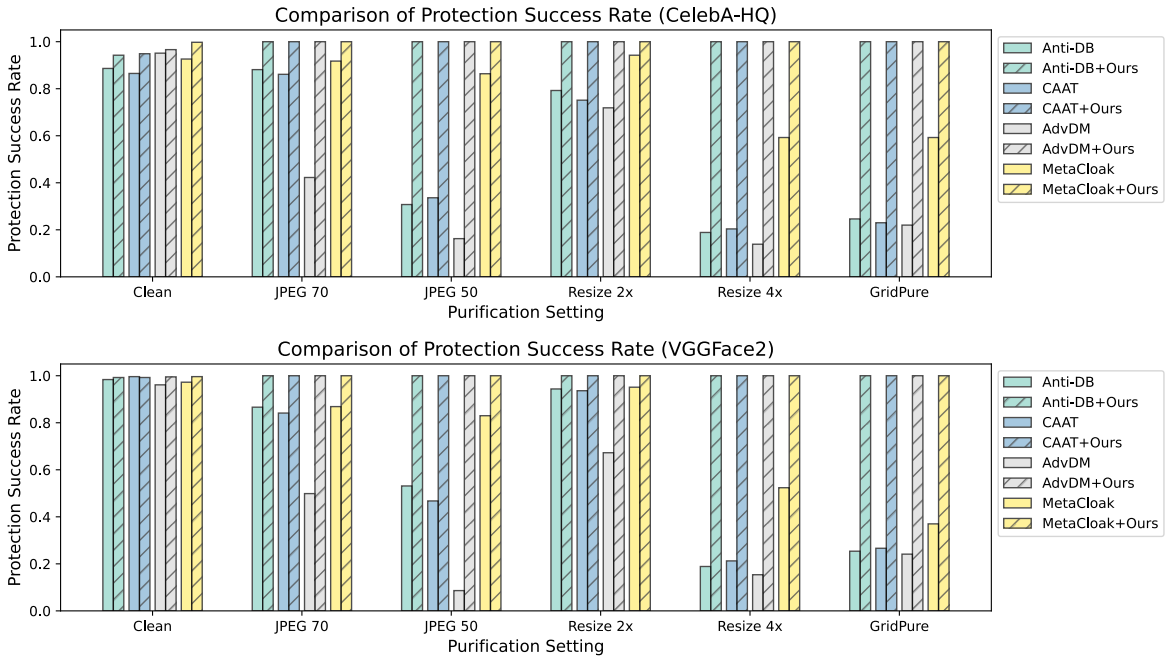


Figure 4. Comparison of Protection Success Rate for different methods across various purification settings. (Generated by prompt "a dslr portrait of sksperson")

Average Change Rate	Frequency Domain	Pixel Domain
4× Resizing	0.9714	0.7788
JPEG Compression (Q=50)	0.9594	0.8445

Table 2. The average change rate of coefficients and pixel values after performing different purifications.

due to the inverse DCT and pixel rounding. To support this explanation, we define the change rate as the proportion of frequency coefficients or pixel values that vary before and after purification. We evaluate it on CelebA-HQ. As shown in Table 2, the frequency-domain perturbations have a higher probability of being changed after the purification, resulting in the inherent vulnerability.

Comparison of high- vs. low-frequency resilience to purification. While it is commonly assumed that high-frequency components are more vulnerable to traditional purification methods (e.g., resizing), our findings show that advanced purification techniques such as GridPure challenge this assumption. We want to share that different purification techniques have different preferences for altering frequency bands. We computed the average normalized variance of the DCT coefficient differences (within 16×16 blocks) before and after purification. As shown in the Figure 5, resizing primarily affects higher frequency bands (green-box region), whereas GridPure significantly alters low-frequency bands (red-box region).

Consequently, we adopt a random and uniform perturbation design in this project to ensure sensitivity to different purifications.

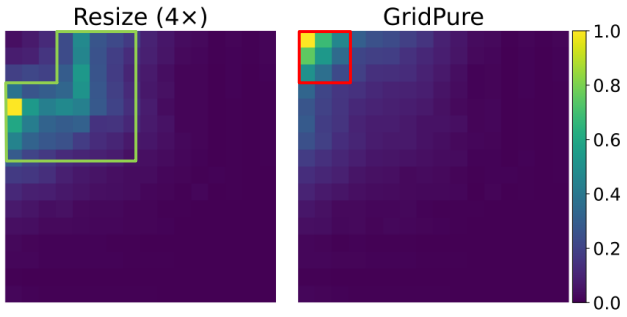
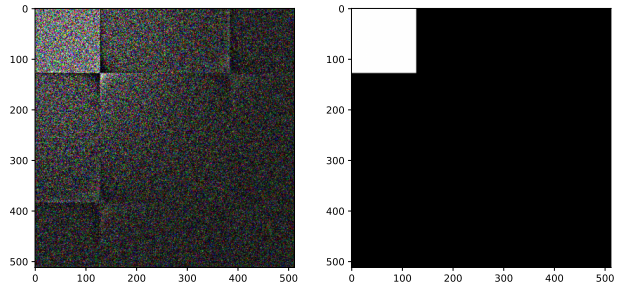


Figure 5. Visualization of the average normalized variance of the DCT coefficient differences (within 16×16 blocks) before and after purification.

2. More Experiment Results

2.1. Influence of Algorithm Design on Mask-guidance

In this experiment, we aim to demonstrate that Algorithm 1 validates the mask guidance. As outlined in the methodology section 3.2, the design of the projected gradient descent algorithm using Equation 8 is intended to invalidate the mask guidance. We verify this through a simulation experiment.



(a) By Equation 8 (b) By Algorithm 1

Figure 6. Visualization of change in the frequency domain after the gradient descent. The visualizations depict the changes in frequency domain coefficients after the updates, where black represents no change, and brighter values indicate greater changes.

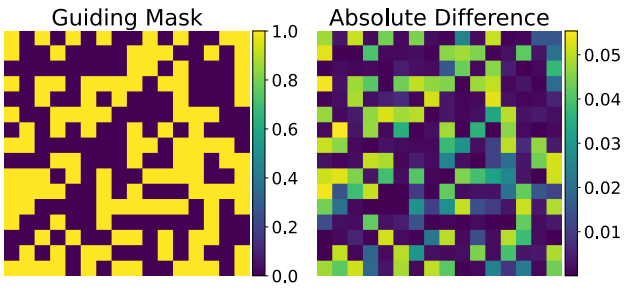


Figure 7. Visualization of the absolute difference in one 16×16 DCT coefficient map before and after applying the protection perturbation, along with the corresponding guiding mask.

Specifically, we generate random gradients in the frequency domain, ensuring they are concentrated in the top-left 128 × 128 region. A 512 × 512 image is transformed into the frequency domain via DCT, and a one-step gradient descent is conducted using Equation 8 and Algorithm 1 (the step size is 1, and the PGD radius is 1). Subsequently, we visualize the changes in frequency domain coefficients after the single gradient descent step. As illustrated in Figure 6, our algorithm successfully confines the coefficient updates to the designated region in the frequency domain, whereas the original algorithm fails to achieve such precise localization.

In addition to the simulation experiment, we also provide a visualization of the absolute difference in one 16×16 DCT coefficient map before and after applying the protection perturbation using Algorithm 1, along with the corresponding guiding mask used during optimization. As illustrated in Figure 7, the perturbation primarily affects regions where the guiding mask is activated (i.e., mask value = 1), confirming that the mask guidance effectively constrains the perturbation by Algorithm 1 (Improved Frequency Domain PGD).

2.2. Repeat Main Experiments with different prompt

Following the experimental setup described in [2, 3], we evaluate the protection performance of our method using an alternative prompt: "a dslr portrait of sks person", to generate individual images. We adopt the same experimental setup described in Section 4.1, with the sole distinction being the prompt utilized.

Figure 4 shows that, with the new prompt, ATP continues to safeguard individual image generation effectively. This is because the integrity-check mechanism prevents generation before the prompt is utilized, ensuring that the performance of this mechanism remains unaffected by variations in the prompt. Table 3 reveals that, under the new prompt, ATP still performs comparably to the original protection perturbation approaches when the purification techniques are not applied.

2.3. Generalizability Analysis

We report the protection performance of ATP (CAAT) trained on SD2.1 when applied to a different diffusion model (SD1.5) and personalization method (SVDiff [1]) using CelebA in Table 4. We compare the protection performance of ATP against that of the unprotected baseline (i.e., without any perturbation applied).

The results demonstrate that ATP is generalizable across diffusion models and personalization techniques.

2.4. Performance Trade-off on Mask Ratio

The authorization and protection perturbations in the frequency domain can be distinguished based on the random mask M. The mask ratio p controls the region in the frequency domain used for authorization versus protection. This experiment shows that adjusting the mask ratio achieves a performance balance between protection and authorization for the ATP.

For example, as the mask ratio increases, a larger portion of the frequency domain will be allocated to authorization. As shown in Table 6 and Table 7, the increase in mask ratio leads to a decrease in bit-error, reflecting an improvement in message embedding accuracy. It also decreases protection performance, as LIQE, CLIP-IQAC, ISM, and FDFR scores indicate. Thus, we adopt a mask ratio of 0.5 as the default setting to achieve a balanced trade-off between authorization and protection performance.

2.5. Performance Trade-off on Block Size

The frequency domain transformation is achieved by BDCT. One of the hyperparameters for it is the size of the Block. In this section, we report the influence of this hyperparameter on the information hiding of authorization perturbation. We train the authorization model using different

	CelebA-HQ				VGGFace2			
	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑	CLIP-IQAC ↓	LIQE ↓	ISM ↓	FDFR ↑
Anti-DB	-0.2047	1.3403	0.3944	0.3775	-0.4274	1.0228	0.3233	0.7950
Anti-DB+Ours	-0.3085	1.1027	0.3509	0.4513	-0.4635	1.0250	0.3073	0.6850
AdvDM	-0.2979	1.0450	0.3193	0.6325	-0.3763	1.0305	0.3650	0.6213
AdvDM+Ours	-0.3367	1.0459	0.3634	0.4638	-0.4703	1.0126	0.3103	0.6538
CAAT	-0.1927	1.3018	0.4139	0.3025	-0.4890	1.0080	0.2819	0.7888
CAAT+Ours	-0.3257	1.0999	0.3725	0.4075	-0.4902	1.0192	0.2914	0.6963
MetaCloak	-0.2573	1.4254	0.3892	0.5000	-0.4485	1.1075	0.3513	0.8613
MetaCloak+Ours	-0.4049	1.0891	0.3488	0.7975	-0.4694	1.0447	0.3500	0.8875

Table 3. Quantitative results for CelebA-HQ and VGGFace2 datasets across various metrics. (Generated by prompt "a dslr portrait of sks person")

SD1.5 + DreamBooth				
	CLIP-IQAC↓	LIQE↓	ISM↓	FDFR↑
Origin	0.5007	4.5427	0.6824	0.0125
ATP	-0.2893	1.1929	0.4329	0.2988
SD2.1 + SVDiff				
	CLIP-IQAC↓	LIQE↓	ISM↓	FDFR↑
Origin	0.3837	4.3704	0.6679	0.1338
ATP	-0.3307	1.0484	0.3861	0.5575

Table 4. Protection Performance of ATP when generation model and algorithm are changed.

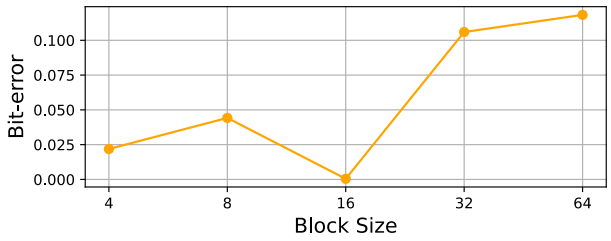


Figure 8. The Bit-error variation under different block size.

block sizes and evaluate it on CelebA-HQ. Figure 8 visualizes the variation in Bit-error under different block sizes. Since the block is square-shaped, we use its side length to represent the block size. It can be observed that a size of 16×16 yields the lowest Bit-error, supporting our design choice adopted in the project.

2.6. Protection Performance Achieved Using Only Authorization Perturbation

In this section, we discuss the protection performance when we don't include protection perturbation in the ATP design. We compare the protection performance of images with no perturbation, images with authoirzation perturbation and images with ATP (taking CAAT as protection perturbaiton) in CelebA-HQ. As shown in the Table 5, authorization perturbation alone fails to provide strong protection when purification is not applied.

	CLIP-IQAC↓	**LIQE↓**	**ISM↓**	**FDFR↑**
Origin (No Perturb)	0.4659	4.2340	0.7053	0.0975
Authorization Alone	0.3258	3.6935	0.6414	0.1075
ATP (CAAT)	-0.3568	1.0768	0.4315	0.6338

Table 5. The protection performance using only the authorization perturbation is significantly worse than that of ATP.

As a result, the combination of protection perturbation and authorization perturbation (ATP) is crucial for achieving reliable protection.

2.7. Repeat Experiments on VGGFace2

We repeat the experiment on VGGFace2 to further validate the credibility of our conclusions in Section 4. We adopt the same experimental setup described in Section 4, with the sole distinction being the dataset utilized. The experiment results are shown in Table 8 and Figure 9.

2.8. Visualization of Perturbed Images

In Figure 10, we present perturbed images generated using different methods from the CelebA-HQ and VGGFace2 datasets. We observe that while perturbations are difficult

Ratio	**Bit-error (e^{-3}) ↓**	**CLIP-IQAC ↓**	**LIQE ↓**	**ISM ↓**	**FDFR ↑**
0.25	0.7813	-0.3561	1.0471	0.3805	0.6400
0.50	0.4688	-0.3139	1.0741	0.4647	0.5213
0.75	0.3125	-0.1480	1.3582	0.5765	0.2225

Table 6. Performance comparison for different mask ratios on CelebA-HQ.

Ratio	**Bit-error (e^{-3}) ↓**	**CLIP-IQAC ↓**	**LIQE ↓**	**ISM ↓**	**FDFR ↑**
0.25	3.1250	-0.422682	1.135657	0.205465	0.91375
0.50	0.7813	-0.386397	1.098367	0.254911	0.81875
0.75	1.2500	-0.371436	1.03989	0.340458	0.70625

Table 7. Performance comparison for different mask ratios on VGGFace2.

	BDCT	Improved-PGD	Mask	Bit-error (e^{-3})↓
(a)				366.09
(b)			✓	43.594
(c)	✓			503.75
(d)	✓		✓	79.844
Ours	✓	✓	✓	0.7813

Table 8. Comparison of different fusion designs with Bit-error values on VGGFace2.

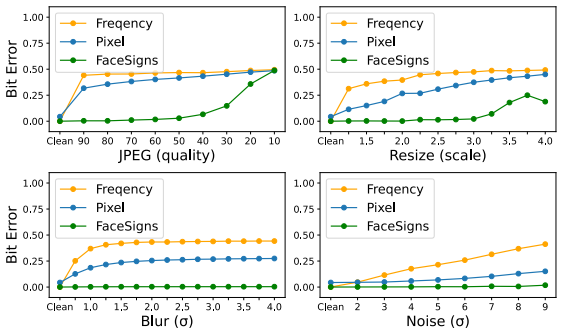


Figure 9. Sensitivity of ATP to different types of purification. The x-axis indicates the hyperparameter of different purifications, while the y-axis indicates the Bit-error. The images are from VGGFace2.

to detect at normal scales, they become noticeable when viewed at an enlarged scale. This remains an unresolved challenge in the field and a focus of our future research efforts.

2.9. Visualization of Generation Results Applied Purification Techniques

We prepared visual cases to illustrate how purification techniques bypass existing protection mechanisms. Specifically, we present the results of individual image generation for images from CelebA-HQ and VGGFace2 after applying different protection perturbation algorithms. As demonstrated in Figure 11 and Figure 12, purification can bypass the protection provided by protection perturbation, compromising the safeguarding of individual image generation.

2.10. More Qualitative Results of Main Experiments

We present additional qualitative comparison results across various methods under two datasets (i.e., CelebA-HQ, VGGFace2) and two different prompts (i.e., a photo of sks person, a dslr portrait of sks person) in Figure 13, Figure 14, Figure 15, and Figure 16.

2.11. Scalability and Computational Efficiency Analysis

Scalability. A safety checker is deployed by the widely used diffusion model library "diffusers", which takes up 1159.60 MB. The authorization model only takes up 1.58 MB, which should be affordable by the service providers.

Computational Efficiency. The ATP requires extra time in authorization message hiding and verification. With batch size = 4, the averaged inference time costs are: Autoencoder encoding/decoding: 0.0201s/0.0274s; BDCT + IBDCT: 0.0016s. In the protection phase, ATP using CAAT as protection perturbation performs autoencoder encoding once and applies mask-guided PGD, which requires two additional BDCT+IBDCT operations per PGD step. This results in **0.38% increase** of the total protection time compared to the original CAAT protection (77.33s). In the generation phase, autoencoder decoding is performed once. When considering a generation method like DreamBooth (341.9s), the added decoding introduces **0.008% increase**.

References

[1] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7323–7334, 2023. 4

[2] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pages 2116–2127. IEEE, 2023. 1, 4

[3] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 24219–24228. IEEE, 2024. 1, 4

[4] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process., 21(12):4695–4708, 2012. 1

[5] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 2555–2563. AAAI Press, 2023. 1

[6] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. 1

Figure 10. Perturbed images of different methods from two datasets.

[7] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 24534–24543. IEEE, 2024. 1

[8] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 14071–14081. IEEE, 2023. 1

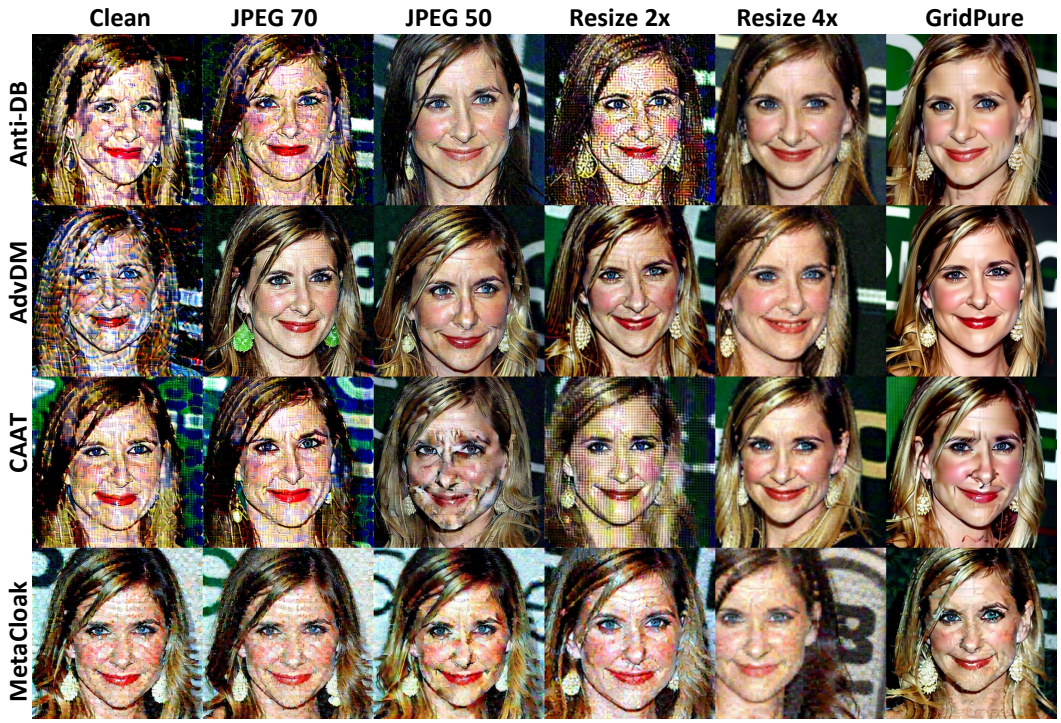


Figure 11. Visual cases showing the purification results bypassing the protection mechanisms on images from the CelebA-HQ dataset. "Clean" indicates no purification applied.

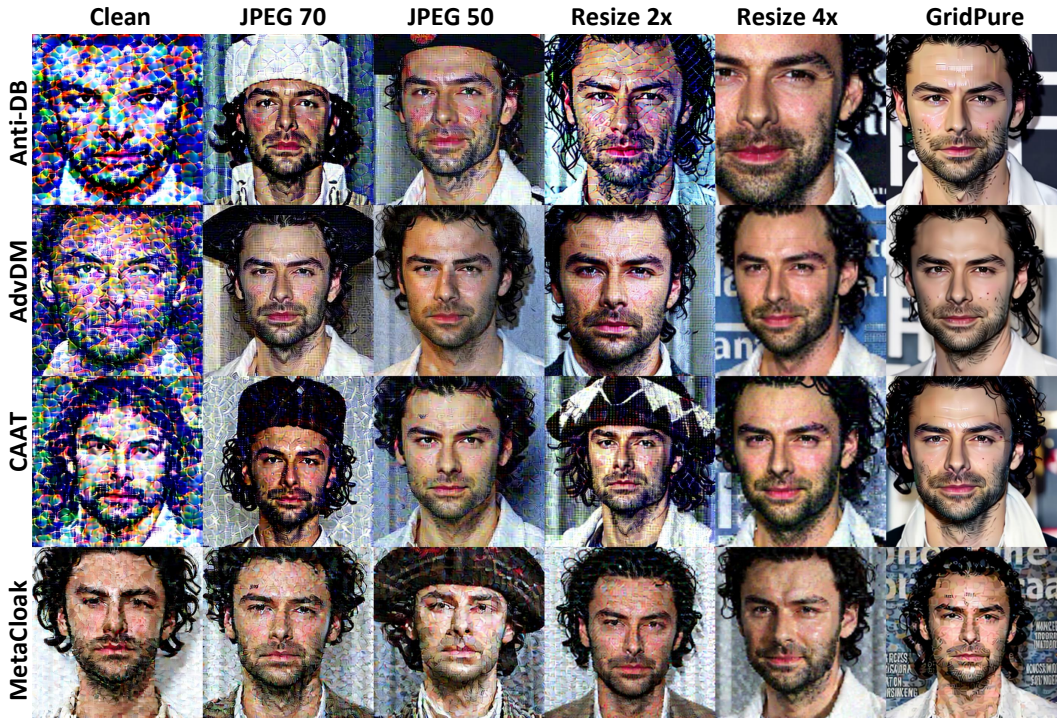


Figure 12. Visual cases showing the purification results bypassing the protection mechanisms on images from the VGGFace2 dataset. "Clean" indicates no purification applied.

Figure 13. Qualitative comparison of original perturbation algorithms and their ATP modified versions in CelebA-HQ.

Figure 14. Qualitative comparison of original perturbation algorithms and their ATP modified versions in CelebA-HQ.

Figure 15. Qualitative comparison of original perturbation algorithms and their ATP modified versions in VGGFace2

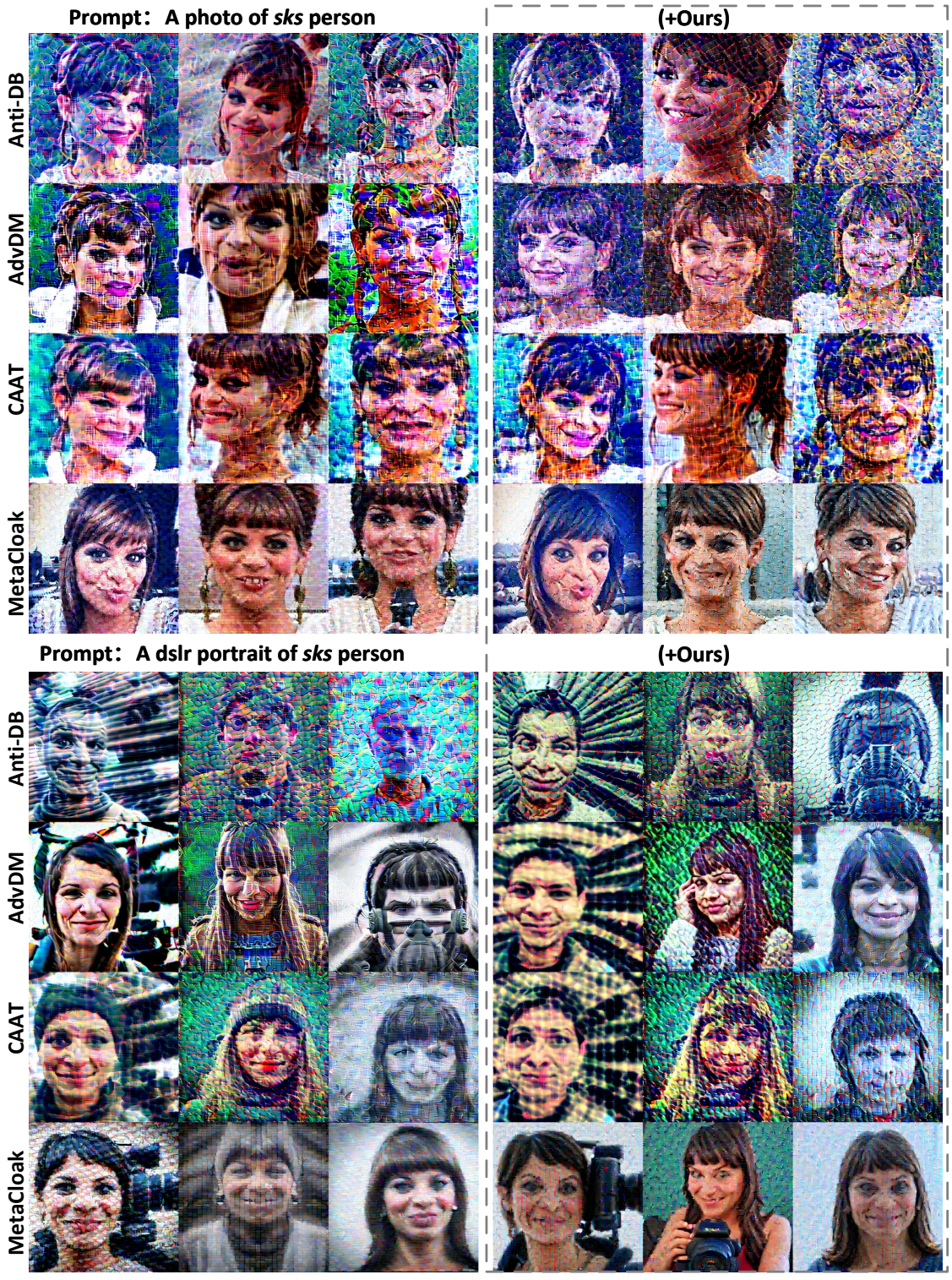


Figure 16. Qualitative comparison of original perturbation algorithms and their ATP modified versions in VGGFace2