Fingerprinting Denoising Diffusion Probabilistic Models

Supplementary Material

6. More Details

6.1. More Experimental Details

In our **FingerInv**, we consistently used two default schedulers from the official code for PS-DDPMs and LDMs, respectively, to control variables and demonstrate the distinctiveness of our crossing route. Specifically, the official default scheduler settings for PS-DDPMs are identical, and for various LDMs, the default scheduler configuration from the official SD code is uniformly applied. To save time, all models used a total of $T = 20$ timesteps. To generate random strings for QR code creation, we used ChatGPT[7] and created QR codes through an online platform.[8] We used a Redmi K60 Pro smartphone to scan QR codes.

In Section 4.3, we explored the impact of different attacks on model performance in the main paper for robustness analysis. We fixed the same random seed to perform 100 samples on both the original model and the models subjected to attacks. This setting for FID calculation may not be conventional; due to our current inaccessibility of the LSUN server, we were unable to use samples from the training set. Moreover, in our context, we primarily aimed to assess distribution changes, and we believe that using a slightly reduced sample size—combined with sampling based on the original model as the target distribution—can still yield relevant insights.

6.2. Determination of Thresholds

In Section 4.2, we conduct a quantitative analysis of the uniqueness in the latent space. A threshold is determined for the squared l_2 distance between the owner fingerprint z and a suspicious fingerprint z', calculated following similar principles as in [24, 25]. We perform a hypothesis test based on the squared l_2 norm of their difference $\Delta z = z - z'$, assuming that z and z' are nearly standard Gaussian samples due to the nature of DDPM. Each element of Δz follows the distribution $\mathcal{N}(0, 2)$, so the variable $Z = \frac{\|\Delta z\|_2^2}{2}$ follows a scaled chi-squared distribution χ_ν^2. To safely reject the null hypothesis that z and z' are similar by applying the p-value approach with $p < 0.05$, we can find a threshold τ such that $P[Z \leq \tau] < 0.05$, or equivalently, find thresholγ such that $P[\|z - z'\|_2^2 \leq \gamma] < 0.05$.

In our experiments, different degrees of freedom ν are applicable for PS-DDPMs and LDMs ($21 \times 3 \times 256 \times 256$ and $21 \times 4 \times 64 \times 64$, respectively). As a result, we calculate two separate thresholds: $\gamma_{ps} = 8.27 \times 10^6$ for PS-DDPMs and $\gamma_{ldm} = 6.91 \times 10^5$ for LDMs.

7. Ablation Studies

7.1. For Fingerprint Extraction

Based on the fingerprint extraction process in Section 3.2, we conducted ablation experiments in three ways: first, using only our initialization noise samples; second, using DDPM inversion samples as initialization and optimizing it through our L_{critical}; and third by altering our loss function L_{critical} to $-L_{\text{critical}}$, similar to [25], which is denoted as MaxObj, acting as a nearly adversarial loss function. As shown in Figure 11, it is evident that using both our proposed initialization and L_{critical} achieve significantly better distinctiveness.

Notably, compared with baseline methods in Figure 6 in the main paper and thresholds $\gamma_{ps} = 8.27 \times 10^6$ and $\gamma_{ldm} = 6.91 \times 10^5$, utilizing either our noise initialization or optimization independently achieves considerable distinctiveness. For example, for DDPM inversion, applying our optimization increases the latent code distances by a factor of at least six. On the other hand, MaxObj does not enhance distinctiveness and performs worse than initialization alone. This aligns with the findings of [25], suggesting that MaxObj may not guide samples as close to or across the performance border-zone as our optimized methods do.

7.2. For Verification Images

In Section 3.2, we discussed the benefits of using QR code images for verification, and noted that they may serve as outliers for most generative models, potentially enhancing distinctiveness. However, we aim for our fingerprinting method's distinctiveness to rely on the definition and optimization of noise at each time step, forming a crossing route at the performance border-zone, rather than solely using specific outliers as x_0. This is because QR code images may not be outliers for all generative models. Therefore, we tested our method's distinctiveness using in-domain images as verification images.

In Figure 1 of the main paper, we use a natural verification image different from the QR code images in PS-DDPMs. Nevertheless, the training datasets of these four models are distributed differently, so we conducted experiments using three LDMs based on the LAION dataset to better illustrate the distinctiveness of our method. Specifically, we randomly sampled several images directly from the LAION-art dataset, and the results are shown in Figure 12. Even when using in-domain images as verification images and employing the three models within a similar generation domain, our method still exhibits strong distinctiveness.

[7] https://chatgpt.com/

[8] https://cli.im/

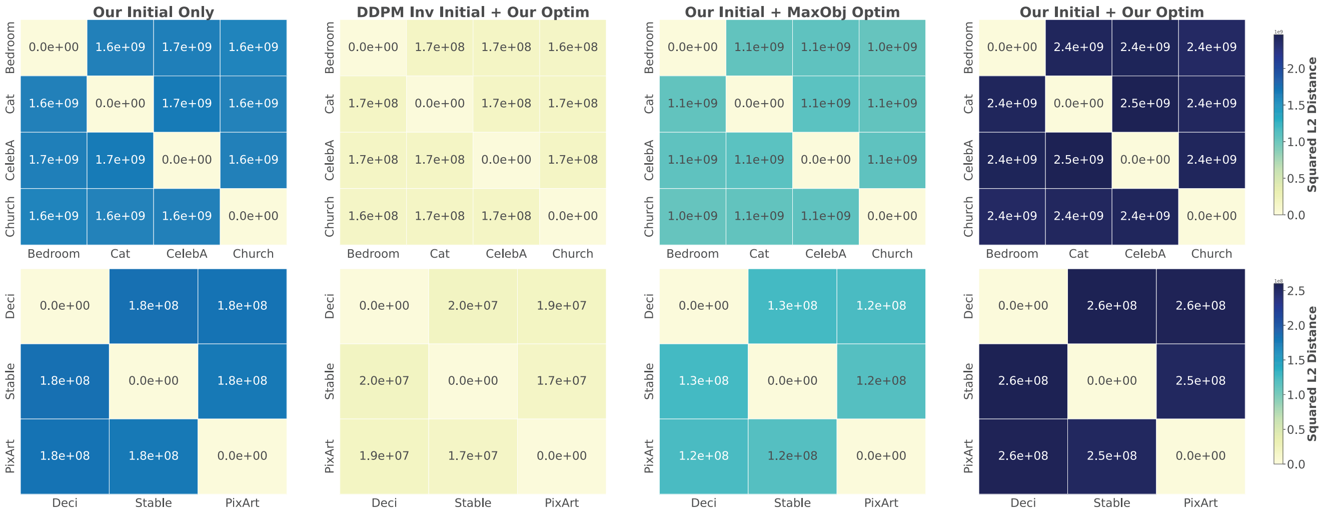

Figure 11. Results of our ablation studies. It is evident that employing both our proposed initialization and optimization methods results in significantly enhanced distinctiveness.

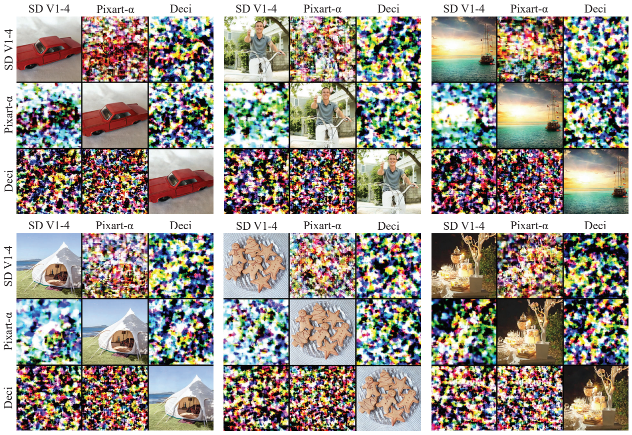

Figure 12. Our discriminative results for in-domain verification images.

8. Other Baseline Methods

In addition to the baseline methods discussed in the main paper, we experimented with additional methods, such as DDIM inversion and methods based on fingerprint restoration models (FRM) as described in [25]. However, their reconstruction and discriminative results indicate that these methods are not directly suited for DDPMs.

DDIM inversion. An intuitive baseline method involves the use of methods for deterministic diffusion models, like DDIM inversion, to estimate x_T as the latent code. However, this approach suffers from error accumulation due to linearization assumptions, making reconstruction challenging. As shown in Figure 13, using DDIM inversion with LDMs results in failed x_0 reconstructions. Additionally, the results also indicate that DDIM inversion may lead to misjudgment.

Fingerprinting restoration models [25]. Another baseline involves adapting the FRM approach as outlined in [25]. Original [25] requiring white-box access to directly optimize the clean image x, we adapted it to optimize noise given a QR code image x in a black-box setting. However, the results underperformed. Figure 13 illustrates that, when applied in a black-box setting, FRM introduces considerable confusion and significant risk of misjudgment-all models produced the similar scannable QR code images.

9. More Results for Uniqueness Analysis

9.1. Scan Results for QR Images of Varying Lengths

In Section 4.2 of our main paper, we used QR code images with a string length of $l_{qr} = 32$ as a case study for scanning verification. Here, we present results for other l_{qr}. As shown in Table 3, the confusion matrices have ✓on the diagonal and ✗elsewhere. These results exhibit the same distinctiveness across different lengths, perfectly distinguishing between models, which are consistent with Figure 8 in the main paper.

9.2. Details for Discriminating Highly Similar Score-based Generative Denoisers

Algorithm. As mentioned in Section 4.2, according to [12], we obtained two highly similar score-based generative denoisers and implemented their **FingerInv** process based on their sampling methods. The purpose of training these denoisers is to remove blind Gaussian noise directly, unlike DDPMs which use a complex scheduler to iteratively map the distribution between pure noise images and target

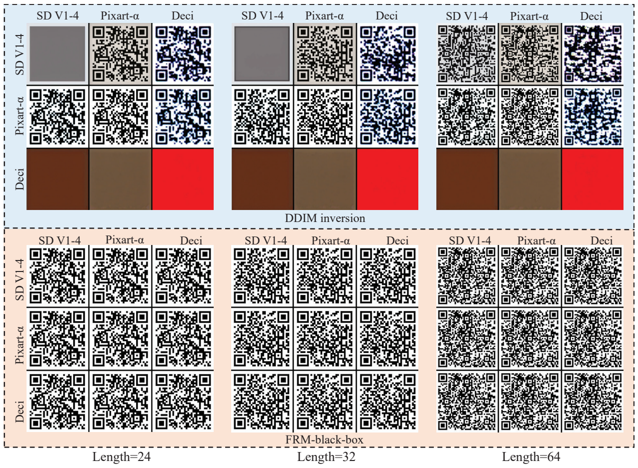

Figure 13. Results of other baseline methods. The upper row displays DDIM inversion outcomes, whereas the lower row shows the results for FRM. Ideally, effective confusion matrices should illustrate scannable QR codes only along the diagonal, thereby suggesting that these two baseline methods lack distinctiveness in differentiating between models.

Table 3. Our discriminative results for different string lengths of QR code images, which perfectly distinguish different models (only the diagonals are ✓).

PS-DDPMs					LDMs			
	Bedroom	Cat	CelebA	Church		SD	Pixart	Deci
String length = 24								
Bedroom	✓	✗	✗	✗	SD	✓	✗	✗
Cat	✗	✓	✗	✗	Pixart	✗	✓	✗
CelebA	✗	✗	✓	✗	Deci	✗	✗	✓
Church	✗	✗	✗	✓	-	-	-	-
String length = 32								
Bedroom	✓	✗	✗	✗	SD	✓	✗	✗
Cat	✗	✓	✗	✗	Pixart	✗	✓	✗
CelebA	✗	✗	✓	✗	Deci	✗	✗	✓
Church	✗	✗	✗	✓	-	-	-	-
String length = 64								
Bedroom	✓	✗	✗	✗	SD	✓	✗	✗
Cat	✗	✓	✗	✗	Pixart	✗	✓	✗
CelebA	✗	✗	✓	✗	Deci	✗	✗	✓
Church	✗	✗	✗	✓	-	-	-	-

images. So we simply iteratively add noises on the crossing route of the denoiser and adapt our fingerprinting method to these denoisers primarily based on their sampling process [12], which involves sampling via ascent of the log-likelihood gradient from a denoiser residual. Our algorithm is presented in Algorithm 2. Specifically, we set a maximum time step T. Given a target verification image x_0, we incrementally add noise to x_0 following our crossing route until reaching the noisy image x_T, then reverse to derive the latent components. We use a larger δ_1 during initialization. The process is simple yet effective, as confirmed by experimental results showing good distinctiveness. In the verification phase, we obtain the output image based on the original sampling algorithm from [12] with fixed latent noises, timesteps, and other parameters. Note that to maintain consistency with previous DDPMs, we label x_T as the noisy image and x_0 as the clean image, reversing the approach in [12]. Specifically, we follow the most of the original parameter settings as [12] and set $h = 0.01$, $\beta = 0.1$, $T = 50$, $N = 10$, $\delta_1 = 10^5$ and $\delta_2 = 1$,and $\lambda = 0.1$.

Algorithm 2 Fingerprint inversion for a score-based generative denoiser

```
Require: denoiser $f$ that estimates the clean image, step size $h$, stochasticity from injected noise $\beta$, target verification image $x_0$, total timesteps T, hardness parameters $\delta_1$ and $\delta_2$, optimization steps $N$, learning rate $\lambda$
for $t = 1$ to $T$ do                                  ▷ Obtain $\{x_1, \ldots, x_T\}$
    $n_o \sim \mathcal{U}(-1, 1)$, $n_g \sim \mathcal{N}(0, 1)$
    $\tilde{\epsilon}_t = n_g + \delta_1 \frac{t-1}{T} n_o$
    for $i = 1$ to $N$ do
        $x_t \leftarrow x_0 + \tilde{\epsilon}_t$
        $L = \frac{T-t}{T} \|f(x_t) - x_0\|_2^2 - \delta_2 \frac{t-1}{T} \|\nabla x_t\|_1$
        $\tilde{\epsilon}_t = \tilde{\epsilon}_t - \lambda \nabla_{\tilde{\epsilon}_t} L$
    end for
    $x_t \leftarrow x_0 + \tilde{\epsilon}_t$
end for
for $t = T$ to $1$ do                                  ▷ Obtain $\{z_T, \ldots, z_1\}$
    $s_t \leftarrow f(x_t) - x_t$          ▷ Compute the score from the denoiser residual
    $\sigma_t^2 \leftarrow \|s_t\|^2 / d$     ▷ Compute the current noise level
    $\gamma_t^2 \leftarrow ((1 - \beta h)^2 - (1 - h)^2) \sigma_t^2$
    $z_t \leftarrow (x_{t-1} - x_t - h d_t) / \gamma_t$          ▷ Compute $z_t$
end for
return latent code $z = \{x_T, z_T, \ldots, z_1\}$
```

More results. In Section 4.2 of the main paper, considering that the resolution of these two models is only 80 × 80, we used QR code images with a string length of 16. Here, we extend the lengths to 24, 32, and 64, and present the results in Figure 14. The results indicate that even with QR code images that are complex relative to the 80 × 80 resolution, our method successfully distinguishes different models. This further demonstrates that our method is effective even for complex verification images, showcasing strong fingerprinting capability.

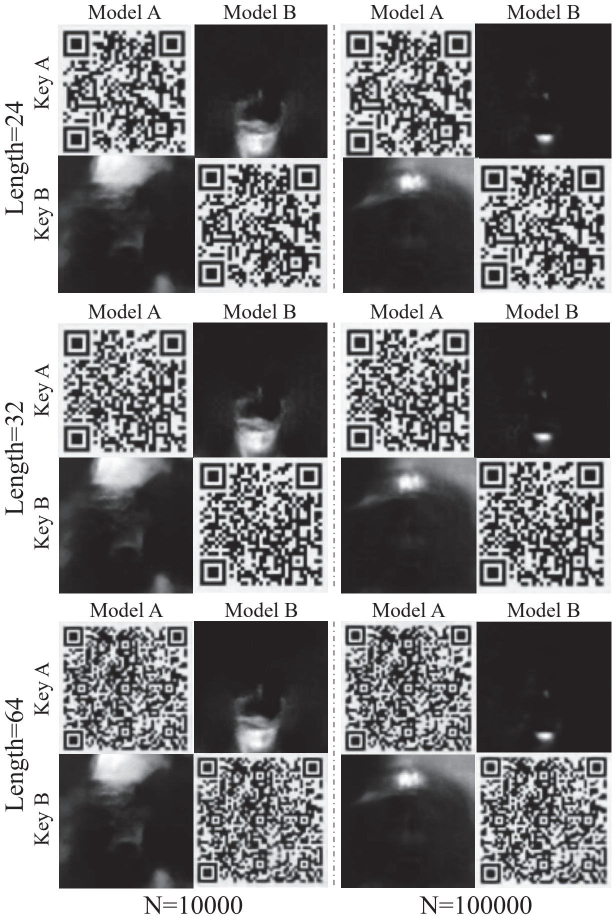

Figure 14. More results for two highly similar score based generative denoisers.

9.3. Results for More Diffusion Models

The focus of this paper is on fingerprinting DDPMs for their wide adoption. We extend to other models, *e.g.* representative SGMs [12], with nearly identical score functions. Here, we again extend to more diffusion models, including flow-matching SOTA ones [14]. We added uniqueness results for these diffusion models, including SD2-1, Deci2, SD3-5 and Flux.1-dev, with training reportedly independent of earlier versions. We simply treat their models as denoisers and use the spirit of **FingerInv** to get their fingerprints, and then resized the resolution of their fingerprint latent codes for comparison. As demonstrated in Fig. 15, the cross-validation results still show distinguishability.

10. More Results for Robustness Analysis

10.1. Different String Lengths of QR images

In Section 4.3 of the main paper, due to space limitations, we conducted robustness analysis experiments using QR code images with a string length of 32. Here, we present results for string lengths of 24 and 64. As shown in Table 4 and Figures 16 and 17, our method demonstrates similar robustness and strong resistance to various attacks for different string lengths of QR images.

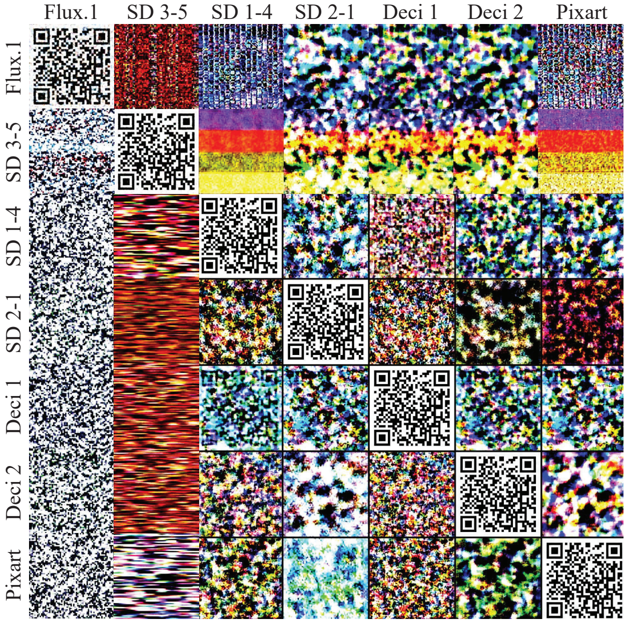

Figure 15. Unique results for more diffusion models.

10.2. Impacts of Attacks

In Section 4.3 of our experiments, we employed attack methods such as pruning, fine-tuning, and quantization, and detailed their significant impact on model performance in the main paper. Here, we provide additional results illustrating the effects of these attacks on the generation outcomes. Figures 18 and 19 show that these attacks clearly influence both unconditional and conditional generation.

It is evident that pruning attacks have a significant impact. Quantization with bfloat16 precision affects PS-DDPMs more significantly than float16 quantization, which has a smaller impact. For fine-tuning attacks, the LAION-art dataset was used to fine-tune PS-DDPMs, and pre-trained fine-tuned models for SD were sourced online. Although the quality of some images generated through fine-tuning attacks may not degrade, their distribution changes. Our fingerprinting method effectively counters these attacks, demonstrating significant robustness.

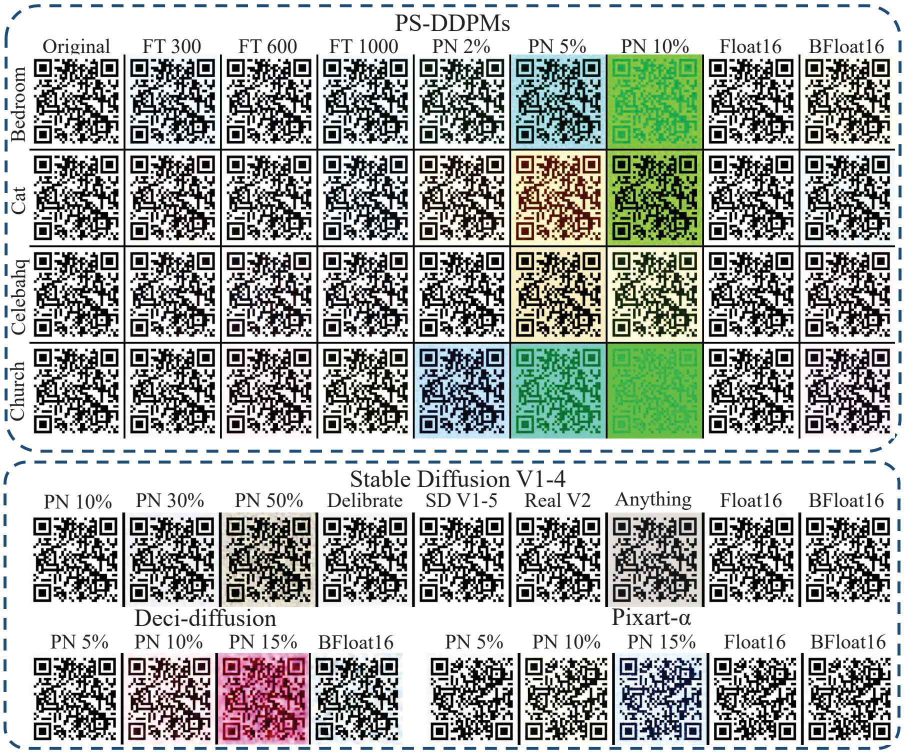

Figure 16. Visual results of robustness analysis using the QR code verification image with string length 24.

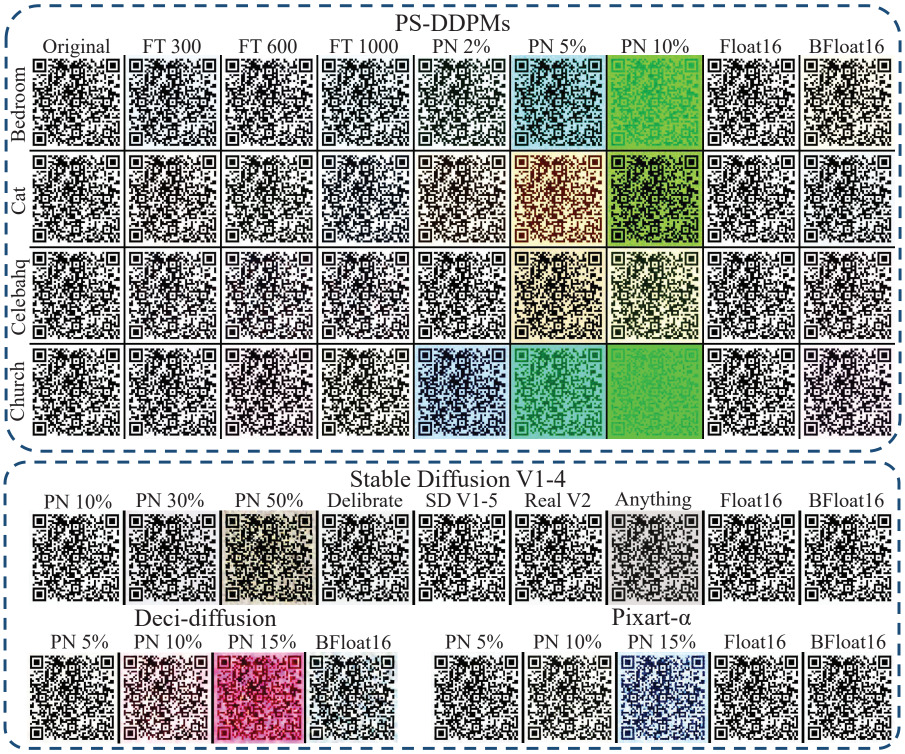

Figure 17. Visual results of robustness analysis using the QR code verification image with string length 64.

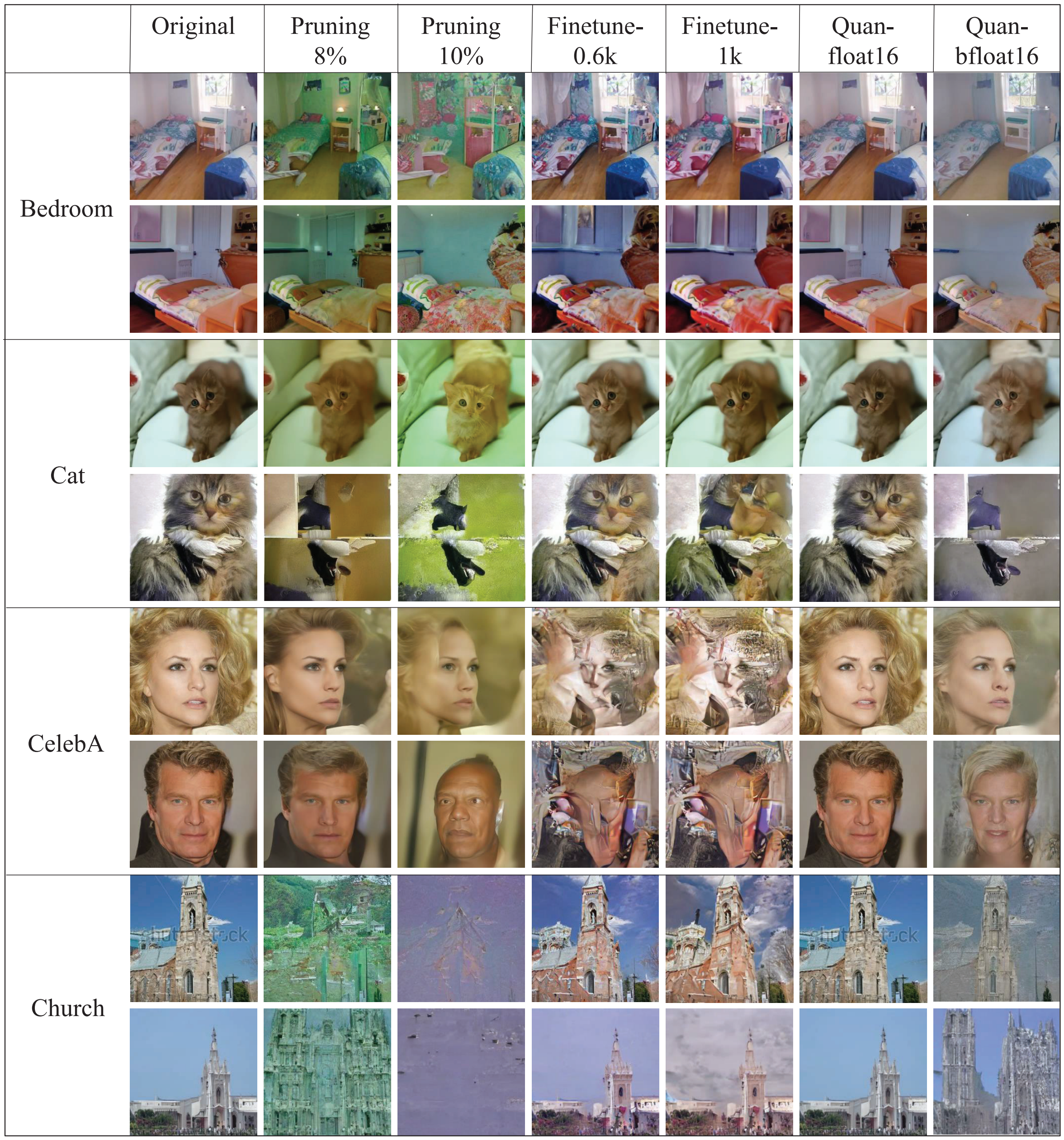

Figure 18. Visual effects of various attacks on unconditional generation.

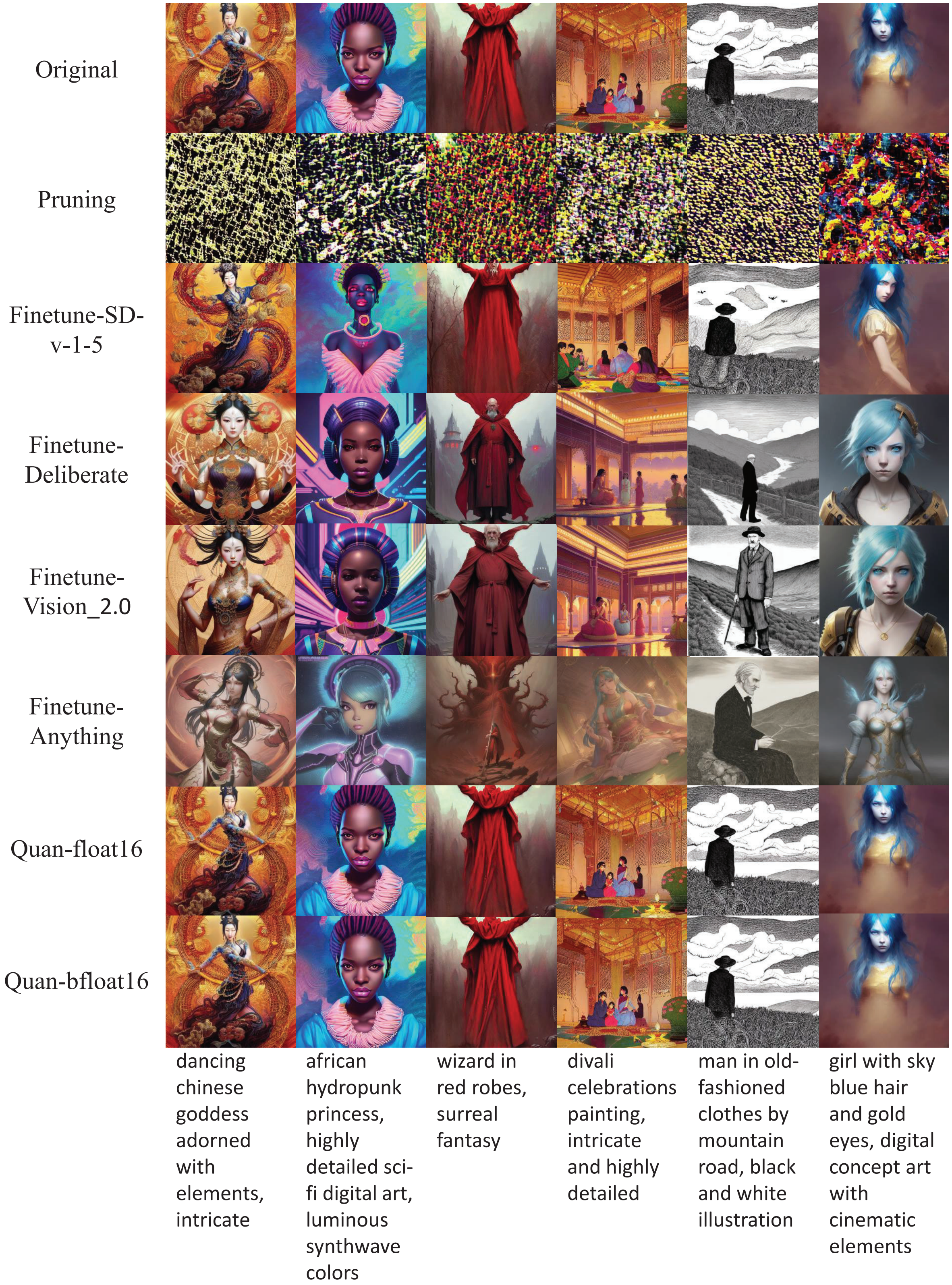
Original
Pruning
Finetune-SD-v-1-5
Finetune-Deliberate
Finetune-Vision_2.0
Finetune-Anything
Quan-float16
Quan-bfloat16
dancing chinese goddess adorned with elements, intricate
african hydropunk princess, highly detailed sci-fi digital art, luminous synthwave colors
wizard in red robes, surreal fantasy
divali celebrations painting, intricate and highly detailed
man in old-fashioned clothes by mountain road, black and white illustration
girl with sky blue hair and gold eyes, digital concept art with cinematic elements

Figure 19. Visual effects of various attacks on conditional generation. The bottom row shows the prompts used for generating the images.

Table 4. Robustness results for verifying QR images with different string lengths l_{qr} of our method. We successfully detected fingerprint information in all attack scenarios.

	Eval	Length=24	Length=32	Length=64
Pruning	Bedroom	✓	✓	✓
	Cat	✓	✓	✓
	CelebA	✓	✓	✓
	Church	✓	✓	✓
	SD V1-4	✓	✓	✓
	Deci	✓	✓	✓
	Pixart	✓	✓	✓
Finetuning	Bedroom	✓	✓	✓
	Cat	✓	✓	✓
	CelebA	✓	✓	✓
	Church	✓	✓	✓
	SD V1-5	✓	✓	✓
	Delibrate	✓	✓	✓
	Realistic	✓	✓	✓
	Anything	✓	✓	✓
Quantization	Bedroom	✓	✓	✓
	Cat	✓	✓	✓
	CelebA	✓	✓	✓
	Church	✓	✓	✓
	SD V1-4	✓	✓	✓
	Deci	✓	✓	✓
	Pixart	✓	✓	✓
Success Rate		100.00%	100.00%	100.00%