Supplementary Material for "RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-Key Identification"

Hai Ci*, Pei Yang*, Yiren Song*, and Mike Zheng Shou✉

Show Lab, National University of Singapore
cihai03@gmail.com yangpei@u.nus.edu
yiren@nus.edu.sg mike.zheng.shou@gmail.com

A Distribution Shift in Watermarking

The operation of discarding imaginary part during the watermarking process results in distribution shift in ℓ_1-to-reference distance for any watermarked noise. In the following, we first demonstrate this from a mathematical view. In order to present this conclusion more clearly and concisely, we consider a more general case where all watermark pixels are i.i.d. sampled from the same watermark distribution as *Tree-Ring* but without a specific pattern for math simplicity. Then we study the more complicated real scenario by empirical experiments.

A.1 Preliminaries

We following the notation convention of [2, 4], representing 2D spatial domain signals using lower case letters indexed by m, n (*e.g.* $x[m, n]$) and 2D frequency domain signals using upper case letters indexed by u, v (*e.g.* $X[u, v]$). We use $\mathcal{F}$ and $\mathcal{F}^{-1}$ to denote DFT and inverse DFT, respectively. The energy of a signal $X[u, v]$ is defined as

$$\mathcal{E}_X = \sum_{u,v} |X[u, v]|^2. \tag{1}$$

A signal can be represented as the sum of a real-valued signal and a complex-valued signal, $X[u, v] = X_{\mathtt{re}}[u, v] + jX_{\mathtt{im}}[u, v]$. It can also be represented as the sum of a conjugate symmetric signal and a conjugate asymmetric signal, $X[u, v] = X_{\mathrm{cs}}[u, v] + X_{\mathrm{ca}}[u, v]$. If a spatial domain signal is real-valued, then its frequency-domain counterpart is conjugate symmetric [4]:

$$\mathcal{F}\{x_{\mathtt{re}}[m, n]\} = X_{\mathrm{cs}}[u, v] = \frac{X[u, v] + X^*[u, v]}{2}, \tag{2}$$

where $X^*[u, v] = X_{\mathtt{re}}[u, v] - jX_{\mathtt{im}}[u, v]$ is the conjugate of $X[u, v]$.

* Equal Contribution. ✉Corresponding Author.

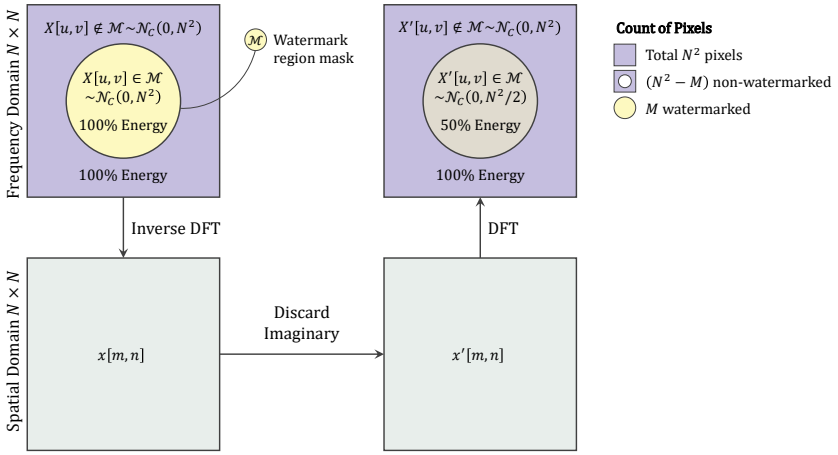

Fig. 1: Visualisation of the analysis. *Tree-Ring* discards the imaginary part of the spatial domain initial latent noise $x[m, n]$ for diffusion denoising, but this operation discards half of the energy from all watermarked frequency-domain pixels. These pixels also have their variance reduced by half.

A.2 Tree-Ring's Pipeline

As visualized in Fig. 1, *Tree-Ring* proposes to add a watermark to a frequency domain initial latent noise $X[u, v]$ (of size $N \times N$) by substituting the value of M pixels within a watermark region mask $\mathcal{M}$. For these watermarked pixels $X[u, v] \in \mathcal{M}$ (corresponds with w used in the paper main content), *Tree-Ring* samples their values from a circularly-symmetric complex normal distribution $\mathcal{N}_{\mathcal{C}}\left(0, N^2\right) = \mathcal{N}\left(0, \frac{N^2}{2}\right) + j\mathcal{N}\left(0, \frac{N^2}{2}\right)$. Other non-watermarked pixels $X[u, v] \notin \mathcal{M}$, individually, also follow this distribution, but the difference is that in *Tree-Ring*'s context, $X[u, v] \notin \mathcal{M}$ are spatially ensured to be conjugate symmetric $X[u, v] = X_{cs}[u, v]$, while $X[u, v] \in \mathcal{M}$ does not have such a guarantee. During the analysis, we view each pixel as a random variable.

After adding the watermark to $X[u, v]$, *Tree-Ring* transforms it back to the spatial domain, $x[m, n] = \mathcal{F}^{-1}\{X[u, v]\}$. The obtained $x[m, n]$ would typically have both the real and imaginary parts. However, since diffusion denoising would always start with a purely-real noise signal, *Tree-Ring* discards the imaginary part of $x[m, n]$, turning it into $x'[m, n] = x_{\mathfrak{re}}[m, n]$, whose frequency domain counterpart $X'[u, v] = \mathcal{F}\{x'[m, n]\}$ is the conjugate symmetric part of the original watermarked signal, $X'[u, v] = X_{\mathrm{cs}}[u, v]$.

A.3 Distribution Shift in Watermarked Region

Eq. (2) implies that the frequency domain signal $X'[u, v]$ satisfies:

$$X'[u,v] = \mathcal{F}\left\{x'[m,n]\right\} = \mathcal{F}\left\{x_{\mathfrak{re}}[m,n]\right\} = X_{\mathrm{cs}}[u,v] = \frac{X[u,v] + X^*[-u,-v]}{2}. \tag{3}$$

For pixels within the watermark region $X'[u,v] \in \mathcal{M}$, $X[u,v]$ and $X^*[-u,-v]$ are uncorrelated, and they both follow a circularly-symmetric complex normal distribution $\mathcal{N}_{\mathcal{C}}\left(0, N^2\right)$. Viewing $X'[u,v]$ as a random variable, its distribution is given by

$$X'[u,v] \sim \frac{1}{2}\left(\mathcal{N}_{\mathcal{C}}\left(0,N^2\right) + \mathcal{N}_{\mathcal{C}}\left(0,N^2\right)\right) = \mathcal{N}_{\mathcal{C}}\left(0, \left(\frac{1}{2}\right)^2 \cdot 2N^2\right) = \mathcal{N}_{\mathcal{C}}\left(0, \frac{N^2}{2}\right). \tag{4}$$

Oppositely, pixels outside the watermark region $X'[u,v] \notin \mathcal{M}$ are conjugate symmetric, which implies that $X'[u,v] = X_{\mathrm{cs}}[u,v] = X[u,v] \sim \mathcal{N}_{\mathcal{C}}\left(0,N^2\right)$. The distribution of non-watermarked pixels $X'[u,v] \notin \mathcal{M}$ are unchanged.

This shows that discarding the imaginary part from the spatial domain pixels $x[m,n]$ changes the distribution of the frequency domain pixels within the watermark region mask $\mathcal{M}$ from $X[u,v] \sim \mathcal{N}_{\mathcal{C}}\left(0,N^2\right)$ to $X'[u,v] \sim \mathcal{N}_{\mathcal{C}}\left(0, \frac{N^2}{2}\right)$.

A.4 From Energy's Viewpoint

Discarding the imaginary part of $x[m,n]$ also causes watermarked region $X'[u,v] \in \mathcal{M}$ to lose half of its energy. Since $X[u,v] \sim \mathcal{N}_{\mathcal{C}}\left(0,N^2\right)$, $|X[u,v]|^2 \sim N^2\chi^2(1)$. Therefore, the expected energy of $X[u,v] \in \mathcal{M}$ is given by

$$\mathbb{E}\left[\sum_{u,v\in\mathcal{M}} |X[u,v]|^2\right] = (N^2 - M)\mathbb{E}_{u,v\in\mathcal{M}}\left[|X[u,v]|^2\right] = (N^2 - M)N^2. \tag{5}$$

Similarly, $|X'[u,v]|^2 \sim \frac{N^2}{2}\chi^2(1)$ for $X'[u,v] \in \mathcal{M}$. Therefore, the expected energy of $X'[u,v] \in \mathcal{M}$ is given by

$$\mathbb{E}\left[\sum_{u,v\in\mathcal{M}} |X'[u,v]|^2\right] = (N^2 - M)\mathbb{E}_{u,v\in\mathcal{M}}\left[|X'[u,v]|^2\right] = (N^2 - M)\frac{N^2}{2}. \tag{6}$$

The fraction of energy of $X'[u,v] \in \mathcal{M}$, compared to $X[u,v] \in \mathcal{M}$, is given by

$$\eta = \frac{\mathbb{E}\left[\sum_{u,v\in\mathcal{M}} |X'[u,v]|^2\right]}{\mathbb{E}\left[\sum_{u,v\in\mathcal{M}} |X[u,v]|^2\right]} = \frac{(N^2 - M)\frac{N^2}{2}}{(N^2 - M)N^2} = \frac{1}{2}, \tag{7}$$

So the watermarked region loses half of its energy.

A.5 Distribution Shift in ℓ_1 Distance

In this section, we clarify the systematic ℓ_1-to-reference shift introduced by discarding the imaginary part during the watermark injection process. To simplify the math, we assume that the recovered watermark comes from the same distribution as the injected one. We consider four different watermarks $\in \mathcal{M}$ in frequency domain:

- $\hat{X}[u,v] \sim \mathcal{N}_{\mathcal{C}}(0, N^2)$: Recovered watermark that never experience imaginary part discarding.
- $\hat{X}'[u,v] \sim \mathcal{N}_{\mathcal{C}}(0, \frac{N^2}{2})$: Recovered watermark that experienced imaginary part discarding.
- $\hat{Y}[u,v] \sim \mathcal{N}_{\mathcal{C}}(0, N^2)$: Null watermark recovered from unwatermarked images.
- $Z[u,v] \sim \mathcal{N}_{\mathcal{C}}(0, N^2)$: The reference watermark to imprint.

Since $\hat{X}'_{\mathfrak{re}}[u,v] \sim \mathcal{N}\left(0, \left(\frac{N}{2}\right)^2\right)$ and $Z_{\mathfrak{re}}[u,v] \sim \mathcal{N}\left(0, \left(\frac{N}{\sqrt{2}}\right)^2\right)$ are both Gaussian, their combination is also Gaussian with summed variance:

$$\left(\hat{X}'_{\mathfrak{re}}[u,v] \pm Z_{\mathfrak{re}}[u,v]\right) \sim \mathcal{N}\left(0, \left(\frac{N}{2}\right)^2 + \left(\frac{N}{\sqrt{2}}\right)^2\right) = \mathcal{N}\left(0, \left(\frac{\sqrt{3}}{2}N\right)^2\right). \tag{8}$$

And similarly for the imaginary parts. Therefore,

$$\left(\hat{X}'_{\mathfrak{re}}[u,v] \pm Z_{\mathfrak{re}}[u,v]\right)^2 + \left(\hat{X}'_{\mathfrak{im}}[u,v] \pm Z_{\mathfrak{im}}[u,v]\right)^2 \sim \left(\frac{\sqrt{3}}{2}N\right)^2 \chi^2(2), \tag{9}$$

where the pdf of a $\chi^2\,(2)$-distributed random variable is given by:

$$f_{\chi^2(2)}(x) = \frac{e^{-\frac{x}{2}}}{2\Gamma(1)} \tag{10}$$

Hence, the expected ℓ_1 distance between $\hat{X}'[u,v] \in \mathcal{M}$ and $Z[u,v] \in \mathcal{M}$, normalised by the count of watermarked pixels M, is given by:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{M}\|\hat{X}'-Z\|_1\right] &= \mathbb{E}\left[\frac{1}{M}\sum_{u,v\in\mathcal{M}}|\hat{X}'[u,v]-Z[u,v]|\right] \\
&= \mathbb{E}\left[\frac{1}{M}\sum_{u,v\in\mathcal{M}}|(\hat{X}'_{\mathfrak{re}}[u,v]+j\hat{X}'_{\mathfrak{im}}[u,v])-(Z_{\mathfrak{re}}[u,v]+jZ_{\mathfrak{im}}[u,v])|\right] \\
&= \mathbb{E}\left[\frac{1}{M}\sum_{u,v\in\mathcal{M}}\sqrt{\left(\hat{X}'_{\mathfrak{re}}[u,v]-Z_{\mathfrak{re}}[u,v]\right)^2+\left(\hat{X}'_{\mathfrak{im}}[u,v]-Z_{\mathfrak{im}}[u,v]\right)^2}\right] \\
&= \mathbb{E}_{u,v\in\mathcal{M}}\left[\sqrt{\left(\hat{X}'_{\mathfrak{re}}[u,v]-Z_{\mathfrak{re}}[u,v]\right)^2+\left(\hat{X}'_{\mathfrak{im}}[u,v]-Z_{\mathfrak{im}}[u,v]\right)^2}\right] \\
&= \frac{\sqrt{3}}{2}N\int_{\mathbb{R}}\sqrt{x}f_{\chi^2(2)}(x)dx = \frac{\sqrt{3}}{2}N\sqrt{\frac{\pi}{2}}.
\end{aligned} \tag{11}$$

Similarly, making use of $\hat{X}_{\mathfrak{re}}[u,v], \hat{X}_{\mathfrak{im}}[u,v], \hat{Y}_{\mathfrak{re}}[u,v], \hat{Y}_{\mathfrak{im}}[u,v], Z_{\mathfrak{re}}[u,v], Z_{\mathfrak{im}}[u,v] \sim \mathcal{N}\left(0, \left(\frac{N}{\sqrt{2}}\right)^2\right)$, the pixel-number-normalized ℓ_1 distance is given by

$$\mathbb{E}\left[\frac{1}{N^2}\|\hat{X}[u,v]-Z[u,v]\|_1\right] = N\sqrt{\frac{\pi}{2}}. \tag{12}$$

$$\mathbb{E}\left[\frac{1}{N^2}\|\hat{Y}[u,v]-Z[u,v]\|_1\right] = N\sqrt{\frac{\pi}{2}}. \tag{13}$$

These expectations satisfy:

$$\mathbb{E}\left[\|\hat{X}'-Z\|_1\right] = \frac{\sqrt{3}}{2}\mathbb{E}\left[\|\hat{X}-Z\|_1\right] = \frac{\sqrt{3}}{2}\mathbb{E}\left[\|\hat{Y}-Z\|_1\right]. \tag{14}$$

Generally speaking, discarding the imaginary part causes the expectation of ℓ_1-to-reference distance to shift by $\frac{\sqrt{3}}{2}$ statistically. This shift factor is derived under several assumptions. In practice, situations are more complicated. *Tree-Ring* injects a fixed watermark pattern. Thus $\hat{X}$, $\hat{X}'$ and Z all carries information of a specific pattern and correlates with each other. They are not i.i.d samples. Both pattern matching and the distribution shift contributes to the expectation of ℓ_1 distance in Eq. (14).

A.6 Distribution Shift in Real Scenarios

As mentioned above, in practice, many other factors affect the distribution shift of the ℓ_1 distance to reference. These factors include the matching of pattern, attacks, diffusion and inversion process, etc. To assess the distribution shift in real scenarios, we conduct a set of control experiments.

In Control 1, we follow the original setup of *Tree-Ring*. The original setup aims at distinguishing between the recovered watermark $\hat{w}$ (shifted) and the

Table 1: Control experiments to demonstrate the effect of distribution shift. We report AUC in verification setting. In Control 1, distribution shift helps distinguish. In Control 2, distribution shift doesn't help. We observe the big performance drop under Rotate and C&S attacks.

Experiment	Clean	Rotate	JPEG	C&S	Blur	Noise	Bright	Avg
Control 1	**1.000**	**0.935**	**0.999**	**0.961**	**0.999**	**0.944**	**0.983**	**0.975**
Control 2	**1.000**	0.728	**0.999**	0.746	0.998	0.940	0.978	0.913

Table 2: Average ℓ_1-to-reference distance in 2 control experiments. We use Δ to denote the difference between $\|\hat{w}_\varnothing - w\|_1$ and $\|\hat{w}_\varnothing - w\|_1$, $\|\hat{w}_\varnothing - w\|_1$ and $\|\hat{w}'_\varnothing - w\|_1$. Larger Δ means easier to distinguish.

Experiment	Target	Clean	Rotate	JPEG	C&S	Blur	Noise	Bright	Avg
-	$\|\hat{w} - w\|_1$	51.50	78.85	66.16	76.96	69.41	73.73	64.86	68.78
Control 1	$\|\hat{w}_\varnothing - w\|_1$	83.92	83.96	84.35	84.58	90.84	83.10	83.30	84.86
	Δ	32.43	5.11	18.18	7.63	21.43	9.37	18.44	**16.08**
Control 2	$\|\hat{w}'_\varnothing - w\|_1$	77.31	80.60	81.21	79.70	86.46	81.63	80.34	81.04
	Δ	25.81	1.75	15.05	2.74	17.05	7.90	15.48	<u>12.26</u>

null watermark $\hat{w}_\varnothing$ (not shifted) recovered from the unwatermarked images, *i.e.* $\|\hat{w} - w\|_1$ *v.s.* $\|\hat{w}_\varnothing - w\|_1$.

In Control 2, we shift null watermark $\hat{w}_\varnothing$ to the same extent as the operation of discarding imaginary part does. Then we get a shifted null watermark $\hat{w}'_\varnothing$. We distinguish between $\hat{w}$ and $\hat{w}'_\varnothing$, *i.e.* $\|\hat{w} - w\|_1$ *v.s.* $\|\hat{w}'_\varnothing - w\|_1$. Here, both $\hat{w}$ and $\hat{w}'_\varnothing$ are shifted to the same extent, so we eliminate the help of distribution shift.

We compare the results of Control 1 and 2 in Tab. 1. We can find general performance drop under all attacks in Control 2. The average AUC decreases from 0.975 to 0.913, making it more challenging to distinguish watermarked and non-watermarked images without the help of distribution shift. Further observation reveals that the major drop occurs in Rotation and Crop & Scale attacks, indicating that distribution shift contributes substantially to the robustness to Rotate and Crop & Scale attacks. This also implies that the original *Tree-Ring* watermark pattern cannot handle these attacks.

Meanwhile, we stat the average ℓ_1-to-reference distance in these two control experiments and compare them in Tab. 2. Without the help of distribution shift in Control 2, the expectations of $\|\hat{w}_\varnothing - w\|_1$ and $\|\hat{w}'_\varnothing - w\|_1$ are closer, especially under Rotation and C&S attacks, indicating overlapped distribution. This implies a lower AUC and more difficulty in distinguishing, consistent with the results in Tab. 1.

B Discarding Imaginary Part As Standalone Watermarking Approach

We demonstrate the operation of discarding the imaginary part can be used as a standalone watermarking approach. Concretely, for each initial noise instance intended for watermarking, rather than injecting a ring watermark into the frequency spectrum X, we opt to inject a random Gaussian noise $\sim \mathcal{N}_{\mathbb{C}}(0, N)$ into the same region. Note that the injected noise are *i.i.d.* sampled for each case thus different from each other. Although originating from the same distribution, the newly introduced Gaussian noise typically lacks conjugate symmetry. Consequently, when transforming to the spatial domain, it necessitates discarding the excess imaginary parts. This results in a distribution shift in watermarked noise, thus the actually injected noise is from $\mathcal{N}_{\mathbb{C}}(0, \frac{N}{2})$.

As discussed in previous sections, this shift induces deviations in the ℓ_1 distance and energy of the watermarked noise from the non-watermarked noise. During verification, we explore three different methods to distinguish the watermarked and the non-watermarked. Specifically, we compute the ℓ_1 distance between the recovered noise and three different references: (1) random Gaussian noise $\sim \mathcal{N}_{\mathbb{C}}(0, N)$ (2) zero (3) zero but with ℓ_2 distance (corresponding to distinguishing the energy between the watermarked and the non-watermarked). Empirical results are presented in Tab. 3. We can find all the mentioned methods effectively detect the presence of the watermark. When we use zero as the reference, the distinction is most pronounced, highlighting its superior discriminatory performance.

Relation with *Tree-Ring$_{rand}$* [5] provides a variant called *Tree-Ring$_{rand}$* that also injects noise as the watermark. However, they inject the same noise pattern for all generated images and intend to rely on pattern matching for watermark verification. The proposed method in this section distinguishes itself from *Tree-Ring$_{rand}$* by injecting *i.i.d.* sampled noise for each generated image. The AUC for *Tree-Ring$_{rand}$* [5] and ours is 0.918 and 0.901, respectively. The closely matched performances indicate that the deviation introduced by discarding imaginary part offers very robust discriminative power. This suggests that discarding imaginary part can effectively distinguish between watermarked and non-watermarked images even without relying on the specific noise pattern.

C Failure Cases of Multi-Channel Rings

As discussed in the main text, we can imprint the ring watermarks onto multiple channels to increase the capacity. However, we find that this often leads to the generation of ring-like artifacts, evident in a substantial proportion of cases, illustrated in Fig. 2. So we only imprint the ring watermark on a single channel by default.

Table 3: Quantitative results when discarding imaginary part is used as a standalone watermarking approach. We calculate the distance between the recovered noise and different references for distinguishment between the watermarked and the non-watermarked. Note that when the reference is Zero and the metric is ℓ_2, it actually distinguishes by energy. AUC is reported.

Ref	Metric	Clean	Rotate	JPEG	C&S	Blur	Noise	Bright	Avg
Gaussian	ℓ_1	0.970	0.837	0.812	0.907	0.841	0.663	0.783	0.831
Zero	ℓ_1	0.998	**0.908**	**0.925**	**0.967**	0.879	**0.747**	**0.881**	**0.901**
Zero	ℓ_2	**0.999**	**0.908**	**0.925**	**0.967**	**0.883**	0.741	**0.881**	**0.901**

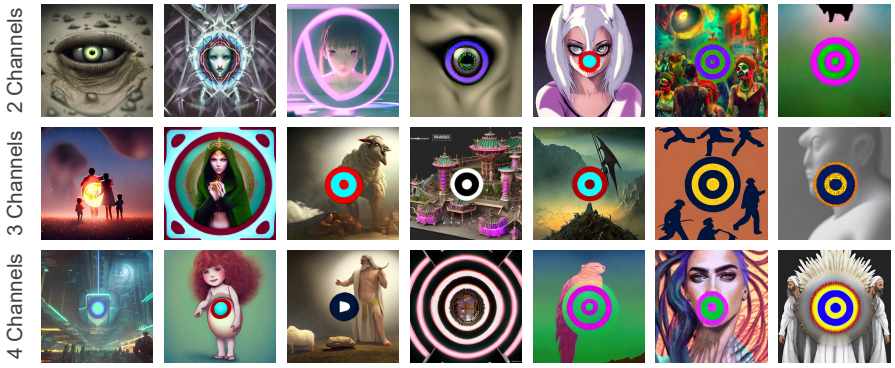

Fig. 2: Generated artifacts when we imprint the ring watermarks on multiple channels.

D More Ablations and Comparisons

D.1 Results on Different Diffusion Models

RingID is a universal method that can be applied to different diffusion models. Tab. 4 shows the results on more diffusion models. It is worth noting that the performance of *RingID* gradually improves from the older version of SD to the newer version of SD.

D.2 Comparison with More Methods

Tab. 5 compares *RingID* with more methods on the watermark verification task. We can find that *RingID* achieves the best **AUC** and **TPR@1%FPR** under both clean and adversarial settings, showcasing strong robustness.

E More Qualitatives

Fig. 3 gives more qualitative results.

Table 4: Empirical results of *RingID* on different diffusion models. Identification accuracy is reported.

Models	#Assigned Keys	Clean	Rotate	JPEG	C&S	Blur	Noise	Bright	Avg
SD 1.4	128	0.950	0.920	0.950	**0.350**	0.950	0.930	0.910	0.851
SD 1.5	128	0.960	0.940	0.960	0.340	0.950	0.940	0.910	0.857
SD 2.1	128	**1.000**	**0.980**	**1.000**	0.280	**0.980**	**1.000**	**0.940**	**0.883**
SD 1.4	2048	0.970	0.810	0.950	0.080	0.970	0.900	0.820	0.786
SD 1.5	2048	0.990	0.800	0.970	**0.110**	0.950	**0.950**	0.850	0.803
SD 2.1	2048	**1.000**	**0.860**	**1.000**	0.080	**0.970**	**0.950**	**0.870**	**0.819**

Table 5: Comparison with more methods on verification.

Methods	AUC/T@1%F (Clean)↑	AUC/T@1%F (Adv)↑	FID↓	CLIP Score ↑
DwtDctSvd [1]	**1.0** / **1.0**	0.702 / 0.262	25.01	0.359
RivaGAN [6]	0.999 / 0.999	0.854 / 0.448	**24.51**	0.361
Tree-Ring [5]	**1.0** / **1.0**	0.975 / 0.694	25.93	0.364
RingID	**1.0** / **1.0**	**0.995** / **0.926**	26.13	**0.365**

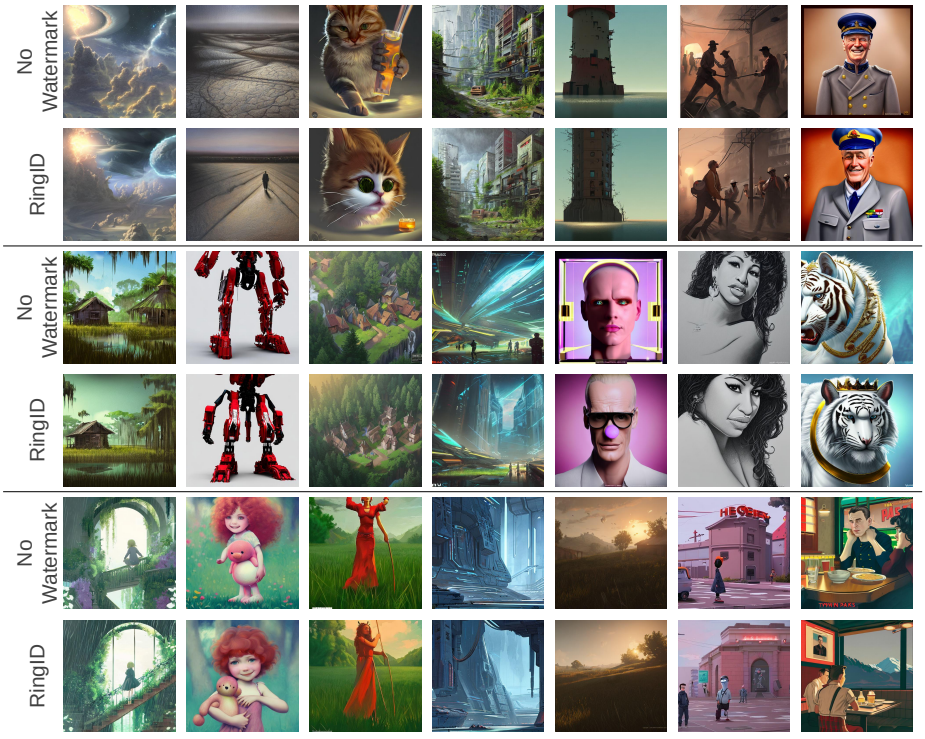

Fig. 3: More qualitative results. Images are generated by SD 2.1 with Stable-Diffusion-Prompts [3].

References

1. Cox, I.: Digital watermarking and steganography. Morgan Kaufmann google schola **2**, 893–914 (2007)
2. Gonzalez, R.C., Woods, R.E.: Digital image processing. Pearson/Prentice Hall, Upper Saddle River, N.J;Harlow;, 3rd edn. (2008)
3. Gustavosta: Stable-diffusion-prompts. `https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts`
4. Mitra, S.K.: Digital signal processing: a computer-based approach. McGraw-Hill Higher Education, London;New York;, 4th edn. (2011)
5. Wen, Y., Kirchenbauer, J., Geiping, J., Goldstein, T.: Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030 (2023)
6. Zhang, K.A., Xu, L., Cuesta-Infante, A., Veeramachaneni, K.: Robust invisible video watermarking with attention. arXiv preprint arXiv:1909.01285 (2019)