BitMark: Watermarking Bitwise Autoregressive Image Generative Models

Louis Kerner, Michel Meintz, Bihe Zhao, Franziska Boenisch, Adam Dziedzic

CISPA Helmholtz Center for Information Security

NEURAL INFORMATION PROCESSING SYSTEMS

arXiv GitHub

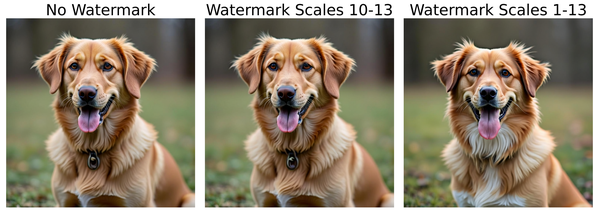
TL;DR Our new `BitMark` is the first watermark for image autoregressive models that operates on bits. It is robust against state-of-the-art attacks and empowers model owners to prevent model collapse.

Motivation

- Generated images are indistinguishable from real images.
- Training data is sourced from the internet.
- Training on generated data reinforces biases and decreases performance → **Model Collapse**.

Contributions

- We propose `BitMark`, the first watermarking scheme for bitwise image generative models preserving high quality outputs.
- We show that `BitMark` is **robust** against conventional and advanced removal attacks.
- `BitMark` is **highly radioactive**: models trained on watermarked images reproduce our watermark.

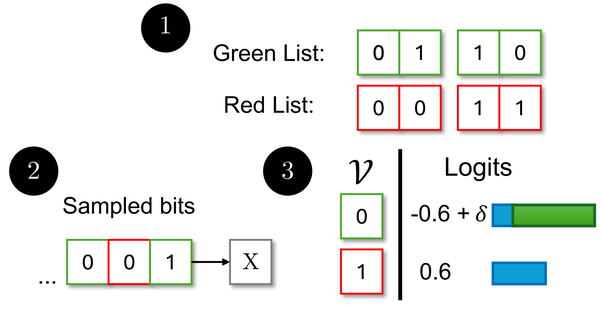


Soft Red/Green-List Biasing

We apply `BitMark` by adding a bias (δ) to the logits of the bit prediction, resulting in minimal impact on the generation process.

$$p_j = \begin{cases} \dfrac{\exp\left(l_j^{(b_j)} + \delta\right)}{\exp\left(l_j^{(\neg b_j)}\right) + \exp\left(l_j^{(b_j)} + \delta\right)}, & \text{if } pre + b_j \in G, \\ \dfrac{\exp\left(l_j^{(\neg b_j)}\right)}{\exp\left(l_j^{(\neg b_j)}\right) + \exp\left(l_j^{(b_j)} + \delta\right)}, & \text{if } pre + \neg b_j \in R. \end{cases}$$

Intuition of BitMark



1 Before generation, a disjoint green and red list of the same size is chosen.

2 The target bit [X] is determined via the previously sampled bits and the green list.

3 The logits of the current prediction are biased towards the target bit by the watermark strength δ.

Detecting BitMark

Algorithm 1: Watermark Detection

Inputs: raw image im, green list G, red list R, image encoder $\mathcal{E}$, quantizer $\mathcal{Q}$;

Hyperparameters: steps K (number of resolutions), resolutions $(h_i, w_i)_{i=1}^{K}$, the number of bits for resolution i is r_i, n - the length of the bit vector.;

$e = \mathcal{E}(im)$;

$C = 0$

for $i = 1, \ldots, K$ **do**

$u_i = \mathcal{Q}(\text{Interpolate}(e, h_i, w_i))$

$u_i = (b_1, \ldots, b_{r_i})$

$C = \text{Count}((b_1, \ldots, b_{r_i}), G)$

$z_i = \text{Lookup}(u_i)$;

$z_i = \text{Interpolate}(z_i, h_K, w_K)$;

$e = e - \phi_i(z_i)$;

Return: $\text{StatisticalTest}(\mathcal{H}_0, (C))$;

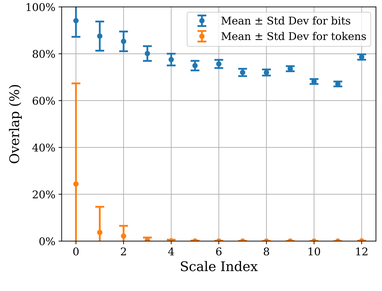
Evaluation of BitMark

We report the TPR@1%FPR (%) for the different conventional and reconstruction attacks.

Watermark	Conventional Attacks									Reconstruction Attacks	
	None	Noise	Blur	Color	Rotate	Crop	JPEG	Vertical	Horizontal	SD2.1-VAE	CtrlRegen+
RivaGAN [1]	99.7	98.3	99.7	99.4	96.7	99.4	99.7	0.0	0.0	98.5	1.6
StegaStamp [2]	100.0	100.0	100.0	98.7	32.1	1.0	100.0	1.0	33.8	100.0	44.2
TrustMark [3]	99.9	99.5	99.9	2.2	2.7	1.6	99.9	0.7	99.8	99.7	1.1
Infinity-2B ($\delta = 2$)	100.0	99.6	99.9	99.8	20.1	98.8	100.0	78.8	100.0	100.0	91.6
Infinity-8B ($\delta = 1.5$)	100.0	99.7	100.0	75.5	57.6	99.4	99.9	93.8	99.7	100.0	30.4
Instella IAR ($\delta = 1.5$)	100.0	96.0	100.0	75.3	2.7	7.5	100.0	9.3	12.1	100.0	93.6

Motivation for Watermarking Bits

Bits are more robust against the re-encoding process than tokens.



High Quality Generation

Applying `BitMark` has no negative impact on image quality if $\delta \leq 3$.

δ	FID↓	KID ($\times 10^{-2}$)↓	CLIP Score↑
0	33.36	1.42 (0.12)	**31.16 (0.28)**
1	32.61	1.38 (0.13)	**31.16 (0.28)**
2	31.05	1.26 (0.12)	31.15 (0.28)
3	**29.61**	**1.03 (0.08)**	31.03 (0.27)
4	42.98	1.78 (0.08)	29.78 (0.32)
5	127.44	11.16 (0.34)	26.13 (0.40)

Radioactivity

We test if `BitMark` is detectable from the output of a given model M_2 after finetuning it on generated and watermarked data from another model M_1. We finetune for 5 epochs on 1,000 watermarked images. We report the TPR@1%FPR (%).

Type of M_1	Type of M_2	Output of M_1	Output of M_2
Infinity-2B	VAR-16	100.0	24.2
Infinity-2B	VAR-20	100.0	25.8
Infinity-2B	VAR-24	100.0	25.7
Infinity-2B	VAR-30	100.0	25.6
Infinity-2B	RAR-B	100.0	4.3
Infinity-2B	RAR-L	100.0	3.3
Infinity-2B	RAR-XL	100.0	3.9
Infinity-2B	RAR-XXL	100.0	4.1
Infinity-2B	Infinity-2B	100.0	100.0
Infinity-2B	Stable Diffusion 2.1	100.0	98.9

References

[1] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. 2019.

[2] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020.

[3] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *IEEE International Conference on Computer Vision (ICCV)*, October 2025.