SERUM: Simple, Efficient, Robust, and Unifying Marking for Diffusion-based Image Generation

Jan Kociszewski*, Hubert Jastrzębski*, Tymoteusz Stępkowski*, Filip Manijak*, Krzysztof Rojek*,
Franziska Boenisch, Adam Dziedzic

CISPA Helmholtz Center for Information Security

arXiv GitHub

TL;DR

We propose **SERUM**: a simple yet highly effective method for marking diffusion-based image generation by adding a unique watermark noise to the initial diffusion noise and training a lightweight detector to reliably identify watermarked images.

Main Contributions

SERUM our novel simple watermark for Diffusion Models:

- Adds a fixed, normalized watermark noise to the *initial diffusion noise*, which preserves the generation quality.
- Trains a lightweight and effective *latent-space detector* to avoid costly diffusion inversion or modifying the diffusion model.
- Achieves state-of-the-art robustness to perturbations, removal attacks, and is highly radioactive, i.e., outputs from models adapted on SERUM-marked data remain detectable with our detector.
- Supports thousands of users while retaining its efficiency, robustness, and the quality of generated images.
- Across Stable Diffusion 1.4 / 2.0 / 2.1, it achieves *state-of-the-art robustness* to 8 perturbations and 7 advanced removal attacks.

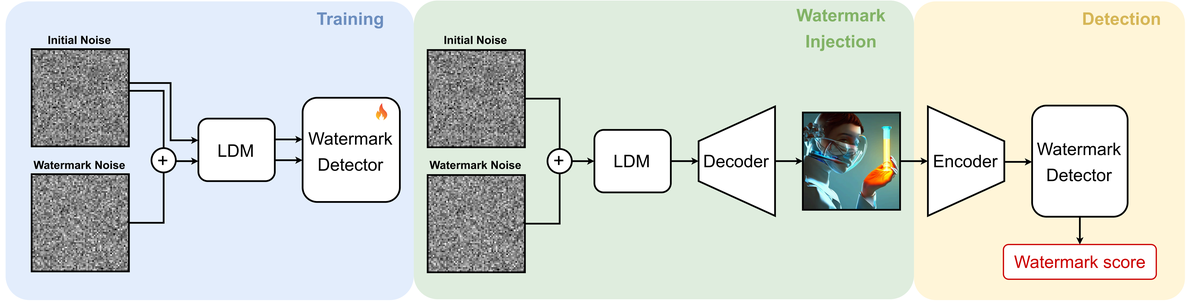
Watermark Injection

We sample a watermark tensor once, normalize it, and mix it with the initial Gaussian noise:

$$A' = \frac{A - \mathrm{mean}(A)}{\mathrm{std}(A)}, \qquad \eta' = \sqrt{1-\alpha}\,\eta + \sqrt{\alpha}\,A'.$$

- Varibale α controls the detectability–fidelity trade-off; the paper uses $\alpha = 0.5$ for the main results.
- Because the watermark is added *before* denoising, it is distributed throughout the generation process and hard to remove cleanly.
- Normalization keeps the injected noise close to the Gaussian prior, helping preserve image quality.

Overview of SERUM



1. **Train detector:** generate clean and watermarked images, encode them into latent space, and train a lightweight detector.
2. **Inject watermark:** add the watermark noise to the initial diffusion noise before sampling.
3. **Detect watermark:** encode a suspect image once and score it directly, **no DDIM inversion** required.

Training and Detection

A lightweight CNN detects the watermark from **latent encodings** of the image.

- Trained with augmented clean and watermarked samples.
- No DDIM inversion required at test time.

Robustness to Perturbations

TPR @ 1% FPR evaluated on 10,000 samples; FPR thresholds calculated per-perturbation.

Method	Avg	Clean	Rot.	JPEG	C&S	R.Drop	Blur	S.Noise	G.Noise	Bright
TrustMark	59.43	100.00	4.30	93.18	99.98	0.42	100.00	14.80	86.96	35.20
Stable Sig.	90.94	99.96	94.62	91.14	99.96	99.06	99.40	91.02	49.02	94.24
RingID	88.48	100.00	99.88	99.98	5.50	99.96	100.00	100.00	92.30	98.74
GaussMarker	97.87	100.00	98.94	99.48	88.72	100.00	97.72	99.90	97.80	98.24
SERUM (Ours)	**99.75**	100.00	99.34	**99.98**	**99.54**	99.86	100.00	100.00	**99.30**	**99.72**

SERUM achieves the best average robustness on SD 2.1 and remains strong across all perturbations, especially compared to prior failures on **rotation**, **crop & scale**, and **noise**.

Advanced Removal Attacks

TPR @ 1% FPR evaluated on 10,000 samples; Detector not trained on these attacks.

Method	Average	VAE	Regen-12	Rinse-4x8	Regen-30	Rinse-2x25	CtrlRegen	I2V
TrustMark	12.30	57.62	1.32	0.70	1.06	1.02	1.18	23.20
Stable Signature	1.18	2.70	1.30	0.62	0.42	0.76	0.86	1.60
RingID	65.40	**99.98**	98.22	84.12	32.58	24.88	77.82	40.20
GaussMarker	58.01	98.36	89.30	65.80	9.72	10.40	91.06	41.40
SERUM (Ours)	**90.87**	99.88	**99.72**	**99.38**	**88.50**	**90.76**	**99.64**	**58.20**

Even under strong unseen attacks, SERUM remains highly robust.

Quality and Efficiency

Image quality

Method	FID↓	CLIP↑
Clean	17.90	0.3143
TrustMark	17.92	0.3139
Stable Sig.	17.97	0.3123
GaussMarker	20.16	0.3140
SERUM	18.96	0.3138

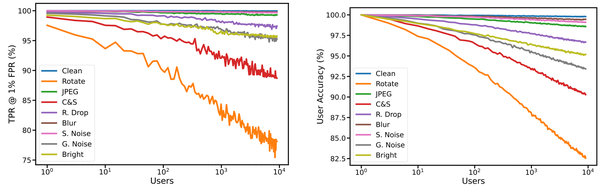
Runtime

Method	Train	Inj.	Det.
TrustMark	336 h	76 s	1.34 m
Stable Sig.	57.52 h	0 ms	1.1 m
GaussMarker	0.93 h	2022 ms	117.0 m
SERUM	9.48 h	17 ms	2.5 m

SERUM preserves near-clean image quality while enabling fast injection and much cheaper detection than inversion-based methods.

Multi-User Setting

- SERUM scales to **9,045 users** using 135 detectors and subset size $k = 2$ (detectors per user).
- At that scale, it achieves **99.96%** clean-image TPR and **99.79%** user accuracy.



(a) Watermark detection (TPR @ 1% FPR). **(b)** User identification accuracy.

Radioactivity

Adapted model	TPR@1%FPR
SD 1.4 (LoRA)	52.30%
SD 1.4 (FFT)	77.12%
SANA-0.6B (FFT)	96.76%

SERUM remains detectable after downstream model adaptation, enabling provenance tracing through later training stages.

Qualitative Results

The style and semantic content are preserved while our SERUM watermark remains highly detectable.