FREEZE THE WEIGHTS, NOT THE GRAPH: PROMPT-RETRIEVABLE WATERMARKS FOR FROZEN DIFFUSION MODELS WITH CERTIFIED RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt-retrievable watermarking asks a text-to-image model to embed a semantic feature of the prompt in the generated image so that a verifier holding a bank of N candidate prompts can later return the one that produced it. We realise this on a *frozen* Stable Diffusion 1.5 with a trainable token selector, latent injector and ResNet decoder around the fixed sampler, and show that two implementation defaults decide whether it works. The memory-cheap freeze, a no-grad denoise with a detached output latent, removes the only gradient into the injector and pins retrieval at chance, whereas a differentiable frozen-weight rollout with per-step gradient checkpointing keeps the weights fixed but the graph intact and lifts held-out Top-1 to 13 to 87 times chance across bank sizes. The latent clamp inherited from bounded-template watermarks saturates a third of a standard-normal latent and replaces the image with a colour field; clamp-free inference restores image content at unchanged retrieval, 0.357 ± 0.025 Top-1 at $N{=}128$, and turns the injection amplitude into a dial between retrieval and prompt fidelity. We state a randomized-smoothing certificate for retrieval with an explicit finite-sample ceiling $R_{\max} = \sigma\Phi^{-1}(\alpha^{1/n})$ and certify the model with 10^4 votes over three seeds. A control suite built around the unwatermarked image from the same seed establishes that the decoder reads the watermark rather than the image, that the signal survives the training attack suite, and that near-duplicate prompts and abstention remain weaknesses. Code and checkpoints are released.

1 INTRODUCTION

Provenance for text-to-image diffusion models (Rombach et al., 2022) is usually posed as *attribution*: embed a random bit-string in the generation and later decide whether an image carries it (Fernandez et al., 2023; Wen et al., 2023; Yang et al., 2024). Prompt-retrievable watermarking (PRW) asks for more: embed a *semantic* feature of the prompt so that a verifier holding a bank of N candidate prompts can return the one that produced the image. A prompt-level payload is what a platform needs to answer "which of these commissions produced this picture" without storing a key per generation, and it is a natural test of how much semantic information a frozen generator can carry from its initial latent to its output.

The architecture we study is small and frozen-backbone friendly (Figure 1). A selector pools CLIP token states of the prompt into a 768-d feature z; an injector g_ϕ maps z to a perturbation of the initial latent; the frozen sampler G renders the image; and a ResNet decoder D_ψ reads back $\hat{z}$ and retrieves $\arg\max_j \cos(\hat{z}, z_j)$ over the bank. Nothing in this description is exotic, and that is the point: the difficulty of PRW is not in the modules but in three choices about *how* a frozen generator is trained through, sampled from, and evaluated. Each choice has a default that practitioners copy, and each default quietly breaks the paradigm.

Training through a frozen sampler. Back-propagating through a multi-step sampler is memory-heavy, so "freeze the generator" is almost always implemented by running the denoise with gradient tracking disabled and detaching the final latent. The diffusion weights stay fixed, as intended, but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the only gradient path into g_ϕ is cut with them: the injector never leaves random initialisation, the decoder is asked to read a fixed random code through generation, and held-out retrieval sits at exactly chance for every bank size, batch size, negative count and amplitude, a plateau that reads as a capacity wall. Our first contribution is the *differentiable frozen-weight rollout*: weights fixed, DDIM graph kept, activations checkpointed per step. At the same 16k-step budget it moves held-out Top-1 from chance to 8–16$\times$ chance; at 120k steps with 64k real prompts it reaches 13–87$\times$ chance for $N \in \{16, \ldots, 512\}$ and is still rising (Section 4.1).

Sampling without a clamp. Latent-space watermarks clip the perturbed initial latent to $[-\gamma, \gamma]$, a range guard inherited from methods that add a bounded template. For a standard-normal latent and γ=1 the guard saturates 32% of the entries and lowers the latent standard deviation to 0.72 *before any watermark is added*; a few-step DDIM rollout then yields a smooth colour field with no recognisable content, watermark or not (Figure 2). Retrieval on such images can be excellent and is meaningless. Our second contribution is *clamp-free inference*: removing the clamp at test time, with no retraining, restores image content at unchanged retrieval (0.357 ± 0.025 vs. 0.359 ± 0.021 Top-1 at N=128) and makes the true cost of the watermark measurable against the unwatermarked image from the same seed. The injection amplitude λ then becomes a dial the operator sets after training: a quarter of the trained amplitude leaves CLIP score and CLIP zero-shot retrieval at the unwatermarked level and still retrieves $14\times$ chance (Section 4.2); a 2k-step clamp-free fine-tune lifts the curve by 3–4 Top-1 points at unchanged fidelity (Section 4.3).

Certifying retrieval. Randomized smoothing (Cohen et al., 2019) certifies any deterministic base classifier, and the bank-retrieval rule is one. Our third contribution is the certificate for retrieval stated in Cohen's base-classifier form, with the one-sided Clopper–Pearson bound written out and the *protocol ceiling* $R_{\max} = \sigma\Phi^{-1}(\alpha^{1/n})$ made explicit: no image can be certified beyond $R_{\max}$ at n votes, so a certified table without this column can silently report unreachable radii. We certify the model with n=10^4 votes ($R_{\max}$=3.20σ) over three seeds in both sampling regimes (Section 4.4).

Controls for learned watermarks. Once images have content, a retrieval number needs a second reading: does the decoder read the watermark, or the image? Our fourth contribution is a control suite that answers it (Section 4.5): the same decoder on the unwatermarked image from the same seed (λ=0), CLIP zero-shot image-to-prompt retrieval on both images, a feature-space hard-negative bank, rule-based prompt edits, and an absent-prompt test of whether the verifier can abstain. The signal is the watermark (44.7 ± 2.3 of 128 hits are watermark-only, 0.0 ± 0.0 content-only); it survives the training attack suite and a TRADES-style adversary does not improve on it; but the feature it carries is bag-of-words, degrades among near-duplicate prompts, and gives no abstention signal (AUROC 0.545 ± 0.005). We report these limits as part of the result.

Contributions. (1) A PRW pipeline on frozen SD-1.5 trained with a differentiable frozen-weight rollout, with the detached-rollout failure diagnosed and measured under matched budgets. (2) Clamp-free inference and an amplitude dial that expose and control the retrieval–fidelity trade-off, plus a short clamp-free fine-tune that lifts it. (3) A certified-retrieval statement and protocol with an explicit finite-sample ceiling, and n=10^4-vote certificates over three seeds. (4) A control suite for learned watermarks and its results, including the negative ones, together with a protocol map that ties every table to its checkpoint, gradient path, clamp, sampler and seeds (Appendix G). Everything is measured with real SD-1.5 in the loop on a single GPU; code and checkpoints are in the supplement.

2 Related work

Watermarks for diffusion models. Learned steganography (HiDDeN, StegaStamp; Zhu et al., 2018; Tancik et al., 2020) inspired watermarks that live in the diffusion pipeline: a fine-tuned VAE decoder that writes a fixed payload (Stable Signature; Fernandez et al., 2023), Fourier rings in the initial noise revealed by DDIM inversion (Tree-Ring, RingID; Wen et al., 2023; Ci et al., 2024), and per-key Gaussian templates (Gaussian Shading; Yang et al., 2024). These carry *identifiers*. Concept attribution moves toward semantics: CustomMark customises a model per creator (Asnani et al., 2025), and TokenTrace embeds secrets by perturbing both the prompt embedding and the initial

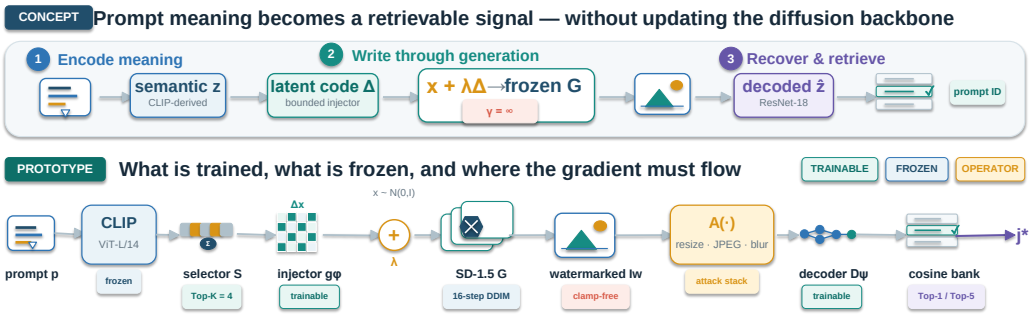


Figure 1: Concept and prototype of PRW. The concept layer follows a semantic feature from the prompt to bank retrieval; the prototype layer expands the executable path through the frozen CLIP and SD-1.5 modules, trainable selector, injector and decoder, clamp-free latent composition, and attack stack. The backward path highlights the implementation rule: freeze the generator weights, not the graph through which the injector learns.

latent and recovers them with a query-conditioned module (Zhang et al., 2026). PRW shares TokenTrace's initial-latent channel but embeds a *continuous* CLIP-derived feature retrieved against an open bank rather than a per-concept secret; the controls of Section 4.5 are meant to be applicable to any method in this family. Initial-noise watermarks change the generated image rather than perturbing a fixed one, so quality is judged distributionally (FID, CLIP score) rather than by PSNR to an unwatermarked reference (Wen et al., 2023); we report both, against the unwatermarked image from the same seed, precisely because the clamp collapse of Section 3.3 is invisible to distributional metrics computed on watermarked images alone.

Attacks and audits. Regeneration and purification attacks strip low-budget marks (Zhao et al., 2024; Saberi et al., 2024); WAVES shows headline numbers collapsing under stress tests absent from the original papers (An et al., 2024); obfuscated-gradient and adaptive attack studies teach that a robustness claim is only as strong as its adversary (Athalye et al., 2018; Tramer et al., 2020). Our control suite (Section 4.5) is in that tradition but aimed at the *clean* number rather than the attacker: whether the decoder reads the watermark or the image, and what the watermark costs against the unwatermarked seed, are questions no attack benchmark asks.

Certified robustness. Randomized smoothing gives ℓ_2 certificates for any base classifier (Lecuyer et al., 2019; Cohen et al., 2019); the practical certificate depends on a one-sided Clopper–Pearson bound whose finite-sample ceiling is rarely stated. Denoised smoothing prepends a denoiser to a frozen model (Salman et al., 2020; Carlini et al., 2023); we used a one-step SD denoiser the same way and state its guarantee as conditional (Appendix D). Adversarially trained smoothing (Salman et al., 2019) and TRADES (Zhang et al., 2019) motivate the mini-max fine-tune we compare against in Section 4.6.

3 Method

3.1 Prompt-retrievable watermarking

A prompt p is encoded by the frozen CLIP ViT-L/14 text tower into token states $H \in \mathbb{R}^{T\times d}$ (d=768). A selector S pools them into one feature $z = \mathsf{S}(H) \in \mathbb{R}^d$ by attention-weighted Top-K pooling (K=4) with a global EOT gate. An injector g_ϕ (a linear layer and four transposed-conv blocks, Tanh output) maps z to a perturbation $\Delta \in \mathbb{R}^{4\times 32\times 32}$ of the SD-1.5 initial latent $x \sim \mathcal{N}(0, I)$, and the frozen sampler G renders

$$I_w = G\big(\mathrm{clip}(x + \lambda\Delta,\, -\gamma, \gamma);\, p\big), \qquad I_0 = G(x;\, p), \tag{1}$$

where I_0 is the unwatermarked image from the *same* seed and prompt, which we use as the fidelity reference throughout. A ResNet-18 decoder D_ψ reads $\hat{z} = D_\psi(\tilde{I})$ from the possibly post-processed image $\tilde{I} = A(I_w)$, and the verifier returns $\arg\max_{j\le N} \cos(\hat{z}, z_j)$ over a bank $\{z_j\}$ of N candidate prompts. Training minimises $\mathcal{L} = \beta\,\mathcal{L}_{\mathrm{reg}} + (1-\beta)\,\mathcal{L}_{\mathrm{NCE}}$ over (ϕ, ψ, S) with β=0.35, where $\mathcal{L}_{\mathrm{reg}}$

is $1 - \cos(\hat{z}, z)$ and $\mathcal{L}_{\text{NCE}}$ an InfoNCE term over 127 in-batch and queued negatives at temperature 0.07; an attack stack A is drawn per step from a hyperbox of resize, JPEG, brightness, blur and salt-and-pepper parameters (Appendix E). Only G and CLIP are frozen. Two scalars in Equation (1), the amplitude λ and the clamp γ, turn out to matter more than any module (Section 3.3).

3.2 Training through a frozen sampler

The gradient $\partial\mathcal{L}/\partial\phi$ exists only through the sampler: Δ enters the initial latent, and the loss is computed on the rendered image. The memory-cheap implementation of "freeze G" (denoise with gradient tracking disabled, detach the final latent, decode with the VAE) therefore removes the only path by which g_ϕ can learn. The injector stays at random initialisation; the decoder, which does receive a gradient, learns to match the selector on *clean* features, so the loss decreases while held-out retrieval sits at $1/N$ and the contrastive term at its $\log(B{+}K)$ ceiling. A direct probe on the real SD-1.5 pipeline confirms the diagnosis: the injector gradient norm is 0 under the detached rollout and $\mathcal{O}(1)$ on every parameter tensor once the graph is kept.

We instead run a *differentiable frozen-weight rollout*: the UNet and VAE weights are excluded from the set of trainable parameters, so $\nabla_\theta{=}0$ exactly as before, but the few-step DDIM graph is retained and each denoising step is wrapped in gradient checkpointing (Chen et al., 2016), so activation memory grows with the number of steps only through recomputation. At 16 DDIM steps, classifier-free guidance 7.5 and 256^2 output this trains at batch 8 on one 96 GB GPU in fp16 at ≈ 2.6 s/step. The rule is short: *freeze the weights, not the graph*.

3.3 Clamp-free inference and the amplitude dial

Latent-space watermarks that add a bounded template clip the perturbed latent to $[-\gamma, \gamma]$ to keep it in range, and $\gamma{=}1$ is the value the configuration we inherited carried. For $x \sim \mathcal{N}(0, I)$ this is not a guard: it saturates 31.7% of the entries of x before any watermark is added and lowers the latent standard deviation from 1 to 0.72 (measured: 32.4% and 0.72 with the watermark, 31.7% and 0.72 without). A DDIM rollout that expects unit-variance noise then produces a low-frequency image with no object structure (Figure 2, row 3). The injector, whose gradient now flows, does exactly what the loss asks of a renderer that outputs colour fields: it writes z into the field, and every retrieval metric looks healthy.

We therefore evaluate (and, in Section 4.3, fine-tune) with the clamp removed ($\gamma{=}\infty$), so that $I_w = G(x + \lambda\Delta; p)$. Because Δ is Tanh-bounded, the watermarked latent is a bounded shift of x with amplitude λ, and λ can be changed at inference without retraining. This gives the operator a one-parameter trade-off: larger λ buys retrieval, smaller λ buys fidelity to the unwatermarked rendering I_0. We measure fidelity against I_0 from the same seed, using CLIP score to the prompt (Hessel et al., 2021), CLIP zero-shot retrieval of the prompt from the image, CLIP image–image cosine and PSNR, because an initial-noise watermark changes the generated image rather than perturbing a fixed one, and distributional metrics on I_w alone would not have exposed the colour fields either (Wen et al., 2023).

3.4 Certified retrieval

Fix the bank and let $f(I) = \arg\max_j \cos(D_\psi(I), z_j)$ be the *base* retrieval rule, a deterministic map from images to bank indices. Its Gaussian smoothing is $g_\sigma(I) = \arg\max_c \Pr_{\eta\sim\mathcal{N}(0,\sigma^2 I)}[f(I + \eta) = c]$. The following is Cohen et al. (2019)'s Theorem 1 specialised to $p_B = 1 - p_A$, stated in their base-classifier form.

Theorem 1 (Certified retrieval)**.** *Let $c_A \in [N]$ and $\underline{p_A} \in (\frac{1}{2}, 1]$ satisfy $\Pr_\eta[f(I + \eta) = c_A] \geq \underline{p_A}$. Then $g_\sigma(I + \delta) = c_A$ for every δ with $\|\delta\|_2 < R(I) := \sigma\,\Phi^{-1}(\underline{p_A})$.*

Estimating $\underline{p_A}$ and the protocol ceiling. Following the certification procedure of Cohen et al. (2019), n_0 noisy votes select c_A, n fresh votes give k successes, and

$$\underline{p_A} \;=\; \text{Beta}^{-1}\big(\alpha;\; k,\; n - k + 1\big), \tag{2}$$

the one-sided Clopper–Pearson lower bound at level α; the procedure abstains if $\underline{p_A} \leq \frac{1}{2}$, and with probability at least $1 - \alpha$ over the votes it either abstains or returns a valid radius. Because $k \leq n$,

Table 1: Held-out Top-1/Top-5 by bank size. *Detached*: no-grad denoise with detached latent; every N we measured sat at chance (the N=128 cell is measured, the others list the chance level the runs pinned to). *Differentiable, 16k*: same 16k-step budget, 4-step DDIM, synthetic prompts. *Base*: 120k steps, 16-step DDIM, 64k real prompts; mean of five held-out bank draws (N=128: .384 $\pm$ .029). All rows are evaluated with the inherited γ=1 clamp; Table 2 shows retrieval is unchanged without it.

	detached	differentiable, 16k		base (differentiable, 120k)	
N	Top-1	Top-1	Top-5	Top-1	Top-5
16	chance (.062)	.500	.896	**.787**	**1.000**
32	chance (.031)	.365	.740	**.662**	**.969**
64	chance (.016)	.240	.661	**.506**	**.922**
128	.008 (chance .008)	.122	.432	**.384**	**.805**
512	chance (.002)	.031	.163	**.170**	**.502**

Equation (2) is at most $\alpha^{1/n}$, so *no image can ever be certified beyond*

$$R_{\max}(\sigma, n, \alpha) \;=\; \sigma\,\Phi^{-1}\big(\alpha^{1/n}\big), \qquad R_{\max} = 1.83\sigma\ (n{=}200), \quad 3.20\sigma\ (n{=}10^4), \tag{3}$$

at α=10^{-3}. We make $R_{\max}$ a column of every certified table and mark cells beyond it "–": such a cell is not "zero robustness", it is unreachable by the protocol. The ceiling is easy to violate in practice: a bound computed from the point estimate rather than the Beta quantile returns $\underline{p_A}$=1 whenever all votes agree and certifies every radius (Appendix C), so we state it. Certified accuracy at (σ, R) is the fraction of images with $g_\sigma(I_w)$ equal to the true prompt and $R(I_w) \geq R$; it certifies Top-1 only.

4 Experiments

Setup. Generator: SD-1.5, 16-step DDIM, guidance 7.5, 256^2, fp16, frozen. Training prompts: 64k real prompts, 8k each from COCO captions, Parti, DiffusionDB, DrawBench, LAION-Aesthetics, Midjourney, an art-domain and a photo-realistic set. The main model ("base") is trained for 120k steps at batch 8 in three stages (Appendix E). Unless stated otherwise, every retrieval number is held-out Top-1 at bank size N=128 (chance $1/128$=.0078): 128 prompts are drawn from a 40,811-prompt pool disjoint from the training prompts, one image is generated per prompt, and each image is retrieved against the 128 features. Seeds $\{42, 43, 44\}$ index the prompt draw and the base latents and are shared by every table; we report mean$\pm$std over seeds. Wilson intervals for a single seed at N=128 are about $\pm$.08 around .35, so we never compare single-seed cells. "Top-1 att." is the mean over three draws of the training attack stack.

4.1 Training through the frozen sampler

Table 1 gives the effect of the rollout at matched budget and at the largest budget we could afford. With the detached rollout, held-out Top-1 is $1/N$ for every $N \in \{16, \ldots, 512\}$, batch size, negative count, regulariser, β and λ we tried. With the differentiable frozen-weight rollout and the *same* 16k steps, Top-1 moves to 8–16$\times$ chance with a train/held-out gap under .04; the base reaches 13–87$\times$ chance and its learning curve is still rising when we stop (N=128: .174 $\rightarrow$ .384 from 32k to 120k steps, monotone; Appendix A).

4.2 Clamp-free inference and the retrieval–fidelity trade-off

Removing the clamp restores the image at unchanged retrieval. Figure 2 shows the same seeds under both regimes and Table 2 the numbers. Clamp-free inference, with no retraining, gives 0.357 ± 0.025 vs. 0.359 ± 0.021 Top-1 clean and 0.376 ± 0.018 vs. 0.364 ± 0.007 under the attack suite. The decoder was trained on colour fields yet reads the same z from images with content: the watermark it learnt is a low-frequency signal that survives the sampler in both regimes. With content restored, the cost of the watermark becomes measurable against I_0 from the same seed. At the trained λ=0.18, CLIP score drops from 0.190 ± 0.002 to 0.159 ± 0.005, CLIP zero-shot retrieval from 0.344 ± 0.036 to 0.135 ± 0.030, the CLIP image–image cosine is 0.746 ± 0.007 and PSNR 6.0 ± 0.1 dB:

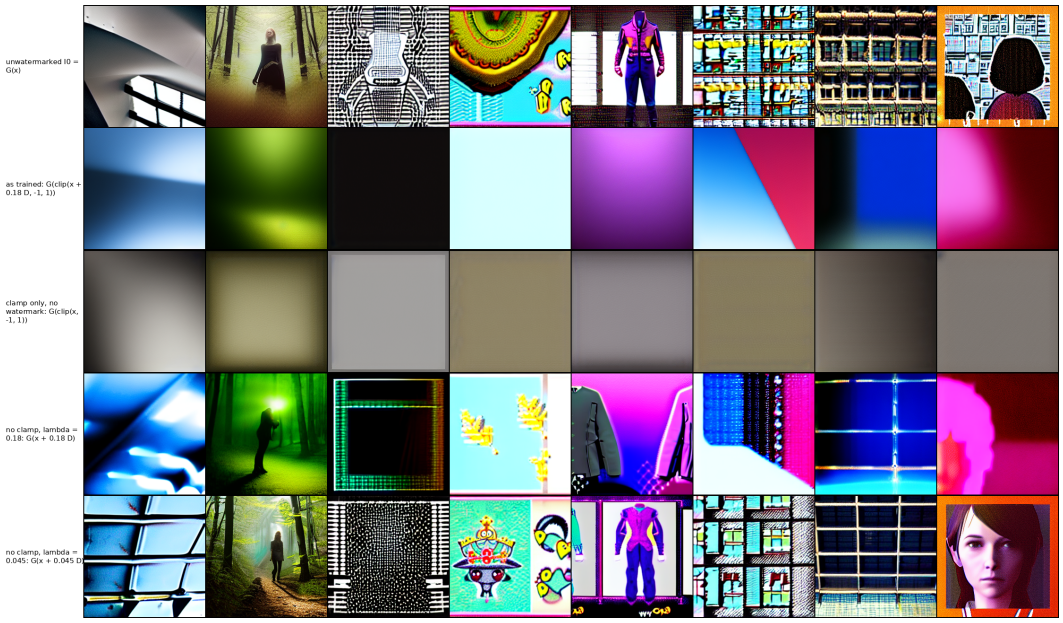


Figure 2: Eight held-out prompts, same seed per column. Row 1: unwatermarked $I_0 = G(x)$. Row 2: with the inherited clamp, $G(\text{clip}(x+\lambda\Delta, -1, 1))$. Row 3: $\text{clip}(x, -1, 1)$ with *no* watermark: the clamp alone produces the collapse. Row 4: clamp-free, $G(x+\lambda\Delta)$ at the trained λ=0.18: content returns with a visible colour cast. Row 5: clamp-free at λ=0.045: visually close to row 1. Retrieval for each regime is in Tables 2 and 3.

Table 2: Sampling regime and controls at N=128 (chance Top-1 .008, Top-5 .039), three seeds. *CLIP score*: cosine between the CLIP image embedding and the CLIP text embedding of the image's own prompt. *CLIP-ZS*: Top-1 of CLIP zero-shot image-to-prompt retrieval over the same bank. $\cos(I_w, I_0)$: CLIP image–image cosine to the unwatermarked image from the same seed.

	Top-1	Top-1 att.	Top-5	CLIP score	CLIP-ZS	$\cos(I_w, I_0)$
PRW, as trained (γ=1)	0.359 ±0.021	0.364 ±0.007	0.823 ±0.052	0.132 ±0.001	0.039 ±0.021	0.694 ±0.004
PRW, no clamp (γ=∞)	0.357 ±0.025	0.376 ±0.018	0.836 ±0.056	0.159 ±0.005	0.135 ±0.030	0.746 ±0.007
λ=0 control (same decoder, I_0)	0.008 ±0.008	0.012 ±0.007	0.034 ±0.009	0.190 ±0.002	0.344 ±0.036	–
CLIP zero-shot on I_w (no decoder)	0.135 ±0.030	0.102 ±0.013	0.279 ±0.023	–	–	–
CLIP zero-shot on I_0 (no watermark)	0.344 ±0.036	–	0.542 ±0.025	–	–	–

the watermarked image is a different rendering of the prompt with a strong tint, not an imperceptible edit. All numbers from here on are clamp-free unless marked.

The amplitude dial. Table 3 sweeps λ at inference with the same checkpoint; the trade-off is monotone on every column. A quarter of the trained amplitude (0.045) leaves CLIP score (.191 vs. .190) and CLIP zero-shot retrieval (.339 vs. .344) at the unwatermarked level and still retrieves 14× chance; half (0.09) costs 2% of CLIP score for 32× chance; 0.135 matches the trained retrieval (.362 vs. .357) at .179 CLIP score instead of .159, so the trained amplitude was already past the knee; doubling it buys .04 Top-1 and removes almost all prompt content CLIP can see (.013 zero-shot). We do not prescribe an operating point; we report the curve, and hold that a PRW number is uninterpretable without the fidelity column beside it.

4.3 Clamp-free fine-tuning

The base was trained with the clamp, i.e. against a renderer that output colour fields; its decoder has never seen a watermarked image with content during training. We fine-tune the base for 2k steps with the clamp removed, once at the trained amplitude (λ=0.18) and once at half of it (λ=0.09), all other settings unchanged, and evaluate with the same three-seed controls (Table 4). Two thousand clamp-free steps add 3–4 Top-1 points at both amplitudes while every fidelity column (CLIP score, CLIP zero-shot, cosine, PSNR) stays within seed noise of the base: at λ=0.18 Top-1 goes 0.357 ±0.025

Table 3: Amplitude–fidelity trade-off at inference (clamp-free, no retraining; N=128; three seeds). At λ=0.045 CLIP score and CLIP zero-shot retrieval match the unwatermarked image and Top-1 is 14× chance; at 0.135 retrieval equals the trained amplitude at a smaller fidelity cost; beyond 0.27 retrieval saturates while CLIP zero-shot retrieval from the image collapses.

λ	Top-1	Top-1 att.	Top-5	CLIP score	CLIP-ZS	$\cos(I_w, I_0)$	PSNR
0.045	0.109 ±0.034	0.096 ±0.020	0.352 ±0.061	0.191 ±0.003	0.339 ±0.018	0.800 ±0.011	9.9 ±0.1
0.090	0.250 ±0.008	0.237 ±0.023	0.695 ±0.064	0.186 ±0.003	0.305 ±0.016	0.772 ±0.007	7.9 ±0.2
0.135	0.362 ±0.027	0.352 ±0.008	0.807 ±0.063	0.179 ±0.004	0.229 ±0.012	0.757 ±0.007	6.7 ±0.1
0.18 (trained)	0.357 ±0.025	0.376 ±0.018	0.836 ±0.056	0.159 ±0.005	0.135 ±0.030	0.746 ±0.007	6.0 ±0.1
0.270	0.391 ±0.014	0.405 ±0.013	0.836 ±0.055	0.130 ±0.005	0.039 ±0.021	0.724 ±0.007	5.4 ±0.1
0.360	0.396 ±0.012	0.411 ±0.014	0.836 ±0.062	0.123 ±0.004	0.013 ±0.005	0.714 ±0.009	5.2 ±0.1
0 (no watermark)	0.008 ±0.008	0.012 ±0.007	0.034 ±0.009	0.190 ±0.002	0.344 ±0.036	1	∞

Table 4: Clamp-free fine-tuning of the base (2k steps, batch 8) at two amplitudes, against inference-only amplitude changes of the base (N=128, three seeds, clamp-free).

model	Top-1	Top-1 att.	Top-5	CLIP score	CLIP-ZS	$\cos(I_w, I_0)$	PSNR
base, λ=0.18 (trained amplitude)	0.357 ±0.025	0.376 ±0.018	0.836 ±0.056	0.159 ±0.005	0.135 ±0.030	0.746 ±0.007	6.0 ±0.1
+ clamp-free fine-tune, λ=0.18	0.393 ±0.052	0.386 ±0.032	0.859 ±0.068	0.160 ±0.004	0.151 ±0.016	0.744 ±0.006	6.1 ±0.1
base, λ=0.09 (inference only)	0.250 ±0.008	0.237 ±0.023	0.695 ±0.064	0.186 ±0.003	0.305 ±0.016	0.772 ±0.007	7.9 ±0.2
+ clamp-free fine-tune, λ=0.09	0.281 ±0.008	0.264 ±0.006	0.682 ±0.039	0.186 ±0.002	0.310 ±0.036	0.771 ±0.004	8.1 ±0.1

→ 0.393 ±0.052 (attacked 0.386 ±0.032), at λ=0.09 it goes .250 → 0.281 ±0.008, the latter gain well outside the seed spread. The fine-tune therefore shifts the retrieval–fidelity curve upward rather than moving along it, but it does not remove the trade-off; the λ=0.09 operating point remains at 0.28 Top-1 with prompt fidelity near the unwatermarked level.

4.4 Certified retrieval

Table 5 certifies the base with the protocol of Section 3.4, with independent noise for the selection and certification stages. The frontier is non-flat in σ and monotone in R: certified Top-1 stays within a few points of smoothed accuracy up to $R \approx R_{\max}/2$ and then falls towards the ceiling; the largest radius with non-trivial mass is R=0.80 at σ=0.30 (0.263 ±0.025) and σ=0.50 (0.177 ±0.039). The clamp-free images certify at least as well as the colour fields at every $\sigma \geq 0.10$ (0.388 ±0.025 vs. 0.367 ±0.014 at (σ, R)=(0.10, 0.20); 0.315 ±0.016 vs. 0.284 ±0.045 at (0.25, 0.40)). We make no claim that these radii are large: at σ=0.25 a certified ℓ_2 radius of 0.40 on a $[0,1]^{3\cdot 256^2}$ image is a per-pixel RMS of 9×10^{-4}. The claim is that the curve is computed with the correct bound and its ceiling stated; bank-size sweeps, other checkpoints, the proof and a unit test of the bound are in Appendix C.

4.5 What the watermark carries

The signal is the watermark, not the content. The λ=0 row of Table 2 passes the *unwatermarked* image I_0 from the same seed through the same decoder: Top-1 is 0.008 ±0.008, i.e. chance. Pairing hits per image, 44.7 ±2.3 of 128 prompts are retrieved from I_w but not from I_0 and 0.0 ±0.0 the other way round. The decoder has learnt nothing about image content that helps retrieval; whatever it reads is put there by g_ϕ. CLIP zero-shot retrieval, the strongest content-only baseline we know of, reaches 0.344 ±0.036 Top-1 on the *unwatermarked* 256^2 image (few-step SD-1.5 images are semantically weak at this resolution) and only 0.135 ±0.030 on the watermarked one, because the tint hurts CLIP. PRW retrieves what content cannot, and Table 3 prices what it spends to do so.

Near-duplicate prompts are the hard case. A uniform bank of 128 prompts is a soft test: real banks contain prompts that differ in one word. We build a hard bank by adding, for each true prompt, its k=3 nearest neighbours in z-space from the 40k disjoint pool (bank size 507, mean cosine 0.88 ±0.02). Top-1 falls to 0.096 ±0.016 and Top-5 to 0.365 ±0.012; the λ=0 control and CLIP zero-shot stay at or below their uniform-bank levels (Table 6). Rule-based edits of the true prompt (Table 7) locate the weakness: PRW separates the true prompt from a colour/size/time swap or a word-order shuffle in 82–84% of pairs but from an object swap or a dropped clause in only 55–61%,

Table 5: Certified Top-1 of the base (N=128, chance .008, n=10^4 votes, n_0=100, α=10^{-3}, mean±std over seeds). "–" marks radii beyond the protocol ceiling $R_{\max} = 3.20\sigma$, which no vote count of 10^4 can certify. The R=0 column is smoothed (not certified) accuracy.

σ	$R_{\max}$	R=0	.05	.10	.20	.40	.80	1.60
clamp-free (γ=∞; images with content)								
0.05	0.16	0.385 ±0.032	0.380 ±0.032	0.380 ±0.032	–	–	–	–
0.10	0.32	0.401 ±0.032	0.396 ±0.027	0.396 ±0.027	0.388 ±0.025	–	–	–
0.25	0.80	0.328 ±0.016	0.328 ±0.016	0.328 ±0.016	0.320 ±0.014	0.315 ±0.016	–	–
0.30	0.96	0.305 ±0.000	0.305 ±0.000	0.299 ±0.005	0.299 ±0.005	0.294 ±0.005	0.263 ±0.025	–
0.50	1.60	0.216 ±0.039	0.216 ±0.039	0.214 ±0.035	0.208 ±0.035	0.193 ±0.039	0.177 ±0.039	–
with clamp (γ=1; colour-field images)								
0.05	0.16	0.401 ±0.012	0.401 ±0.012	0.398 ±0.008	–	–	–	–
0.10	0.32	0.378 ±0.016	0.375 ±0.021	0.372 ±0.018	0.367 ±0.014	–	–	–
0.25	0.80	0.310 ±0.044	0.310 ±0.044	0.310 ±0.044	0.302 ±0.050	0.284 ±0.045	–	–
0.30	0.96	0.281 ±0.034	0.279 ±0.037	0.279 ±0.037	0.268 ±0.035	0.247 ±0.036	0.214 ±0.039	–
0.50	1.60	0.135 ±0.040	0.135 ±0.040	0.135 ±0.040	0.135 ±0.040	0.133 ±0.041	0.115 ±0.036	–

Table 6: Hard-negative bank: the true prompts plus each prompt's k=3 nearest neighbours in z-space from the disjoint pool (bank size 507, mean cosine 0.88 ±0.02 between a true prompt and its neighbours), against the uniform bank (N=128, three seeds).

bank	PRW Top-1	PRW Top-5	λ=0 Top-1	CLIP-ZS Top-1
uniform bank (N=128)	0.357 ±0.025	0.836 ±0.056	0.008 ±0.008	0.135 ±0.030
z-NN hard bank (k=3)	0.096 ±0.016	0.365 ±0.012	0.005 ±0.005	0.068 ±0.025

tracking how far the edit moves z ($\cos$ =.62 vs. .90–.94). Top-K token pooling is bag-of-words with a preference for rare tokens; the watermark inherits that.

No abstention signal. A verifier should also say "none of these". Removing the true prompt from the bank and comparing the maximum cosine with and without it gives an AUROC of 0.545 ±0.005 for PRW (CLIP: 0.550 ±0.007); the mean margin between the true prompt and the best impostor is −0.005 ±0.006. PRW retrieves well when the answer is present and cannot tell when it is not; any threshold-based abstention rule, including the one in the certification procedure, inherits this.

4.6 Robustness and adversarial fine-tuning

The training attack stack (resize, JPEG, brightness, blur, salt-and-pepper) leaves Top-1 unchanged at every amplitude (Table 3, "Top-1 att."). We also test whether TRADES-style mini-max fine-tuning (Zhang et al., 2019) improves on the base: an instance-conditioned attack policy predicts the five attack-stack parameters, an SPSA inner step maximises $1 - \cos(D_\psi(A(I_w)), z)$ under an ℓ_2 budget, and the outer step mixes clean and attacked images (55% clean) for 2k steps at batch 8 with the differentiable rollout (Appendix F). Under the paired three-seed protocol against a 5-step PGD attacker (Table 8) the paired difference in PGD Top-1 is −0.013 ±0.071 and in clean Top-1 −0.081 ±0.061: the fine-tune costs clean retrieval and buys no robustness beyond seed noise. The base's robustness comes from the attack stack it was trained with, not from an adversary.

5 Discussion and limitations

What we now believe about PRW. The paradigm works in a precise sense: a small injector can write a 768-d prompt feature into the initial latent of a frozen SD-1.5 such that a ResNet reads it back from the rendered image at 45× chance for N=128, through the training attack suite, with a correctly computed randomized-smoothing certificate, but only if the sampler is differentiated through and the latent is not clamped. It works by changing the image: at the trained amplitude the watermark is a tint that costs 16% of CLIP score; at a quarter of the amplitude the cost vanishes and retrieval drops to 14× chance; clamp-free fine-tuning lifts the curve by a few points but does not remove the trade-off. The feature it carries is bag-of-words, which makes near-duplicate prompts hard and object swaps nearly invisible, and it offers no abstention signal.

Table 7: Rule-based edits of the true prompt. *Coverage*: fraction of prompts the rule applies to. *Pairwise*: fraction of covered images whose decoded feature is closer to the true prompt than to its edit (chance .5), for PRW and for CLIP zero-shot. *Top-1*: PRW retrieval when every edit is added to the bank.

edit	coverage	$\cos(z, z')$	PRW pairwise	CLIP-ZS pairwise	PRW Top-1
attribute swap (colour/size/time)	0.23 ± 0.01	0.62 ± 0.02	0.825 ± 0.077	0.702 ± 0.075	0.331 ± 0.134
object swap (noun)	0.50 ± 0.06	0.94 ± 0.01	0.554 ± 0.062	0.472 ± 0.010	0.189 ± 0.044
word-order shuffle	0.94 ± 0.02	0.62 ± 0.01	0.843 ± 0.054	0.357 ± 0.055	0.251 ± 0.031
drop last clause	0.66 ± 0.07	0.90 ± 0.02	0.610 ± 0.083	0.406 ± 0.047	0.226 ± 0.026

Table 8: TRADES-style mini-max fine-tuning of the base (clamp-free, $N{=}128$, three seeds, paired prompts and seeds).

	clean Top-1	clean Top-5	PGD-5 Top-1	PGD-5 Top-5
base (no fine-tune)	0.396 ± 0.030	0.844 ± 0.051	0.372 ± 0.030	0.799 ± 0.055
+ TRADES-style mini-max, 2k steps	0.315 ± 0.032	0.766 ± 0.041	0.359 ± 0.048	0.810 ± 0.037

Why the defaults fail silently. Each of the three defaults we replace passes the checks a careful practitioner runs: the detached rollout gives a decreasing loss and a retrieval curve flat at chance, which reads as a limit; the clamp gives a healthy learning curve, a small train/held-out gap and robustness to the attack suite, on images that are not images; a certificate without its ceiling gives a curve that is non-flat and monotone, with the unreachable cells inflated consistently across seeds. The diagnostic that catches each lies outside the loop the metric is computed in: a gradient-norm probe per module, a side-by-side look at the unwatermarked seed, and the finite-sample ceiling of the confidence bound. We suggest all three as defaults for learned watermarks on frozen generators (Appendix G).

Limitations. One backbone (SD-1.5) at 256^2, one sampler, one decoder architecture, and a single training seed for the base (evaluation seeds vary prompts and latents). The clamp-free fine-tune is short (2k steps) and has no explicit fidelity term, so whether a fidelity-constrained injector can hold $45\times$ chance at zero CLIP-score cost is open. The hard-negative and edit banks are built by rules and feature-space nearest neighbours, not human paraphrase. The mini-max comparison uses one recipe and one budget. Certified radii are ℓ_2 in pixel space and small in absolute terms; we do not claim they protect against realistic editing attacks.

ETHICS STATEMENT

This work proposes a provenance watermark and the controls needed to evaluate it honestly. It reduces, rather than creates, the risk of over-claiming watermark robustness; no new attack capability is introduced beyond the standard attack stack and PGD/SPSA solvers already in the literature. All prompts come from public prompt corpora; no personal data is used. We note that a prompt-retrieval verifier that cannot abstain (Section 4.5) could produce confident false attributions if deployed against an incomplete bank, and we report that limitation explicitly.

REPRODUCIBILITY STATEMENT

The supplementary ZIP contains the full training and evaluation code (injector, selector, decoder, differentiable frozen-weight rollout, attack stack, mini-max fine-tuning, the certification module and the control suite), the fp16 base checkpoint and the clamp-free fine-tuned checkpoint, the exact held-out prompt banks and seeds, the shell scripts that produced every table, and the script that regenerates all tables in this paper from the raw per-seed outputs; the README maps each table to the script that produced it. Appendix E lists hyper-parameters; Table 13 ties every table to its checkpoint and protocol. A unit test for the Clopper–Pearson bound (Appendix C) checks the shortcut and the correct quantile. All runs use one 96 GB GPU; wall-clock is given in Appendix E.

AI USE STATEMENT

Large language models were used as writing and coding assistants throughout: to polish prose, to draft sections that the authors then edited, to search for and summarise related work, and to help design, implement and debug experiments (including the control suite and the certification code). All experimental results were produced by code the authors ran and inspected; no result, figure or number was generated by a language model, and no mathematical claim rests on model-generated reasoning. The authors take full responsibility for the content.

REFERENCES

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. WAVES: Benchmarking the robustness of image watermarks. In *International Conference on Machine Learning (ICML)*, 2024.

Vishal Asnani, John Collomosse, Xiaoming Liu, and Shruti Agarwal. CustomMark: Customization of diffusion models for proactive attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1523–1533, 2025.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

Nicholas Carlini, Florian Tramèr, Krishnamurthy Dvijotham, Leslie Rice, Mingjie Sun, and J. Zico Kolter. (Certified!!) Adversarial Robustness for Free! In *International Conference on Learning Representations (ICLR)*, 2023.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv:1604.06174*, 2016.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. RingID: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision (ECCV)*, 2024.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

P. Fernandez, G. Couairon, H. Jegou, M. Douze, and T. Furon. Stable signature: Rooting watermarks in latent diffusion models. In *ICCV*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (S&P)*, 2019.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of AI-image detectors: Fundamental limits and practical attacks. In *ICLR*, 2024.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *NeurIPS*, 2020.

M. Tancik, B. Mildenhall, and R. Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *CVPR*, 2020.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *NeurIPS*, 2020.

Y. Wen et al. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *NeurIPS*, 2023.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

Li Zhang, Shruti Agarwal, John Collomosse, Pengtao Xie, and Vishal Asnani. TokenTrace: Multi-concept attribution through watermarked token recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42942–42952, 2026.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. In *NeurIPS*, 2024.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*, 2018.

A Learning curve of the base

Table 9 gives the held-out learning curve of the base (resumed from a 30k-step differentiable-rollout run, then a 60k "strong" stage, then the 120k final stage whose final checkpoint is the base used throughout). The curve is monotone at every N and still rising when we stop. All rows are in the as-trained γ=1 regime; Table 2 shows the final checkpoint retrieves identically without the clamp.

Table 9: Held-out Top-1 of the base over training. Rows up to 60k average 3 held-out bank draws, later rows 5.

step	N=16	N=64	N=128	N=512
32k	.562	.292	.174	.053
40k	.646	.380	.232	.078
50k	.667	.443	.253	.109
60k	.646	.453	.299	.118
70k	.750	.491	.348	.129
80k	.800	.516	.350	.132
90k	.825	.516	.350	.154
100k	.775	.497	.362	.157
110k	.825	.528	.386	.160
120k	.787	.506	.384	.170

B Controls by bank size

Table 10 repeats the main controls of Table 2 at $N \in \{16, 32, 64, 128, 512\}$ (three seeds). The as-trained and no-clamp retrieval columns agree at every N; the λ=0 column stays at chance; CLIP zero-shot on the watermarked image stays well below its unwatermarked counterpart.

Table 10: Controls by bank size (three seeds). PRW columns are the same decoder on the same checkpoint, with and without the clamp; the remaining columns are in the no-clamp regime.

		PRW Top-1					CLIP-ZS Top-1	
N	chance	γ=1	no clamp	Top-1 att.	Top-5	λ=0	I_w	I_0
16	0.062	0.792 $\pm$0.072	0.812 $\pm$0.062	0.847 $\pm$0.024	1.000 $\pm$0.000	0.083 $\pm$0.072	0.250 $\pm$0.125	0.542 $\pm$0.130
32	0.031	0.646 $\pm$0.036	0.677 $\pm$0.048	0.691 $\pm$0.016	1.000 $\pm$0.000	0.031 $\pm$0.031	0.229 $\pm$0.065	0.417 $\pm$0.065
64	0.016	0.521 $\pm$0.048	0.516 $\pm$0.027	0.535 $\pm$0.016	0.969 $\pm$0.054	0.021 $\pm$0.009	0.167 $\pm$0.009	0.417 $\pm$0.009
128	0.008	0.359 $\pm$0.021	0.357 $\pm$0.025	0.376 $\pm$0.018	0.836 $\pm$0.056	0.008 $\pm$0.008	0.135 $\pm$0.030	0.344 $\pm$0.036
512	0.002	0.185 $\pm$0.011	0.174 $\pm$0.003	0.176 $\pm$0.003	0.503 $\pm$0.015	0.003 $\pm$0.001	0.072 $\pm$0.005	0.221 $\pm$0.013

C CERTIFICATION: PROOF, BOUND TEST, AND FULL TABLES

Proof of Theorem 1. f is a deterministic measurable map from $\mathbb{R}^{3HW}$ to $[N]$, so g_σ is the Gaussian smoothing of a classifier and Theorem 1 of Cohen et al. (2019) applies with p_B replaced by its worst case $1 - \underline{p_A}$, giving $R = \frac{\sigma}{2}\big(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(1 - \underline{p_A})\big) = \sigma\,\Phi^{-1}(\underline{p_A})$. The bank enters only through f; it must be fixed before certification. □

Clopper–Pearson unit test. A tempting shortcut for the lower bound is `binom.ppf(α, n, k/n)/n`, the α-quantile of a binomial with the point estimate as success probability. It is not a confidence bound. For k=n=200, α=10^{-3} it returns 1.0, whence $\Phi^{-1}(1) = \infty$ and a certificate at every radius; the correct `beta.ppf(0.001, 200, 1)` returns $0.96605 = \alpha^{1/n}$, and $\Pr[\mathrm{Bin}(200, 0.96605) \geq 200] = 10^{-3}$ confirms it is the exact one-sided bound. For k=190, n=200 the shortcut gives 0.895 (below the point estimate, so it looks like a bound) and the correct value is 0.8835; the two are close whenever votes disagree, so the error is visible only on unanimous images—exactly the ones that set the large-radius cells, which is how a table can report radii beyond $R_{\max}$ without any cell looking wrong. A unit test covering both cases is in the supplement.

Bank-size sweep. Table 11 certifies the base at $N \in \{32, 64, 128, 512\}$ (seed 42) in both regimes.

Table 11: Certified Top-1 by bank size (seed 42; n=10^4, α=10^{-3}). Columns are (σ, R) pairs; all radii are below $R_{\max}$.

N	chance	(.10, 0)	(.10, .20)	(.25, .40)	(.30, .80)
as trained (γ=1)					
32	0.031	0.594	0.594	0.562	0.469
64	0.016	0.516	0.516	0.406	0.328
128	0.008	0.391	0.375	0.258	0.195
512	0.002	0.156	0.137	0.094	0.070
no clamp					
32	0.031	0.656	0.656	0.531	0.500
64	0.016	0.562	0.562	0.438	0.375
128	0.008	0.430	0.406	0.328	0.273
512	0.002	0.180	0.170	0.137	0.117

Other checkpoints. Table 12 certifies the intermediate and ablation checkpoints under the same protocol: the 60k "strong" stage of the base, the 16k differentiable-rollout run on synthetic prompts (Table 1, middle), the detached-rollout run whose injector never trained, and the hybrid latent-plus-feature variant of Appendix D. All are evaluated with the γ=1 clamp they were trained with (colour-field certificates); the hybrid is also certified clamp-free.

D TWO EXTENSIONS WE EXPLORED AND DO NOT CLAIM

Diffusion-as-denoiser smoothing (DDS). Prepending a one-step SD-1.5 denoiser D_t (t=100) to the decoder gives a base map $f_{\mathrm{DDS}} = f \circ D_t$ that is still deterministic, so Theorem 1 applies to it with $\underline{p_A}$ estimated under f_{DDS}. One would like a "radius domination" statement: if D_t is an L-contraction with $L < 1$ and $D_t(I) \approx I$ on watermarked images, then $p_A^{\mathrm{DDS}}(I) \geq p_A^{\mathrm{van}}(I)$ and

Table 12: Certificates for the other checkpoints (N=128, three seeds, same held-out banks). Same layout as Table 5. The 16k run was trained on synthetic prompts, so the real held-out bank is out of distribution for it (its in-distribution Top-1 is .122, Table 1).

σ	$R_{\max}$	R=0	.05	.10	.20	.40	.80	1.60
base, 60k stage								
0.05	0.16	0.286 ±0.024	0.276 ±0.018	0.268 ±0.025	–	–	–	–
0.10	0.32	0.302 ±0.024	0.286 ±0.016	0.279 ±0.005	0.268 ±0.012	–	–	–
0.25	0.80	0.276 ±0.020	0.268 ±0.025	0.268 ±0.025	0.263 ±0.024	0.247 ±0.025	–	–
0.30	0.96	0.266 ±0.000	0.263 ±0.005	0.260 ±0.009	0.253 ±0.005	0.237 ±0.009	0.214 ±0.024	–
0.50	1.60	0.201 ±0.037	0.201 ±0.037	0.198 ±0.033	0.193 ±0.030	0.182 ±0.025	0.161 ±0.037	–
differentiable rollout, 16k, synthetic prompts (Table 1, middle)								
0.05	0.16	0.089 ±0.012	0.086 ±0.014	0.078 ±0.016	–	–	–	–
0.10	0.32	0.078 ±0.027	0.078 ±0.027	0.070 ±0.014	0.068 ±0.009	–	–	–
0.25	0.80	0.070 ±0.008	0.070 ±0.008	0.070 ±0.008	0.062 ±0.008	0.052 ±0.009	–	–
0.30	0.96	0.060 ±0.005	0.057 ±0.009	0.055 ±0.008	0.052 ±0.012	0.042 ±0.012	0.031 ±0.016	–
0.50	1.60	0.036 ±0.012	0.036 ±0.012	0.036 ±0.012	0.031 ±0.008	0.031 ±0.008	0.026 ±0.005	–
detached rollout, 30k (injector never trained)								
0.05	0.16	0.031 ±0.000	0.023 ±0.008	0.018 ±0.005	–	–	–	–
0.10	0.32	0.021 ±0.012	0.013 ±0.009	0.013 ±0.009	0.013 ±0.009	–	–	–
0.25	0.80	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	–	–
0.30	0.96	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	0.010 ±0.005	0.008 ±0.000	–
0.50	1.60	0.008 ±0.000	0.008 ±0.000	0.008 ±0.000	0.008 ±0.000	0.008 ±0.000	0.008 ±0.000	–
hybrid latent+feature (feature head certified), γ=1								
0.05	0.16	0.240 ±0.033	0.232 ±0.027	0.224 ±0.025	–	–	–	–
0.10	0.32	0.229 ±0.018	0.224 ±0.023	0.221 ±0.025	0.216 ±0.030	–	–	–
0.25	0.80	0.211 ±0.016	0.206 ±0.020	0.203 ±0.021	0.195 ±0.027	0.177 ±0.037	–	–
0.30	0.96	0.203 ±0.027	0.203 ±0.027	0.198 ±0.025	0.193 ±0.030	0.180 ±0.041	0.159 ±0.032	–
0.50	1.60	0.156 ±0.036	0.156 ±0.036	0.156 ±0.036	0.154 ±0.037	0.148 ±0.034	0.133 ±0.021	–
hybrid latent+feature, clamp-free								
0.05	0.16	0.234 ±0.034	0.232 ±0.037	0.227 ±0.028	–	–	–	–
0.10	0.32	0.234 ±0.028	0.234 ±0.028	0.229 ±0.025	0.221 ±0.025	–	–	–
0.25	0.80	0.216 ±0.037	0.214 ±0.039	0.208 ±0.037	0.206 ±0.040	0.190 ±0.039	–	–
0.30	0.96	0.206 ±0.040	0.206 ±0.040	0.195 ±0.036	0.195 ±0.036	0.188 ±0.036	0.167 ±0.025	–
0.50	1.60	0.172 ±0.036	0.167 ±0.043	0.167 ±0.043	0.161 ±0.035	0.159 ±0.032	0.154 ±0.027	–

hence $R_{\mathrm{DDS}}(I) \geq R_{\mathrm{van}}(I)$. The implication from the probability ordering to the radius ordering is immediate from the monotonicity of Φ^{-1}; the premise is the substantive part, and we could only verify the contraction hypothesis on *average* over images and noise draws, not as a per-image supremum. The statement we can make is therefore conditional: for any image at which $p_A^{\mathrm{DDS}} \geq p_A^{\mathrm{van}}$, the DDS certificate is at least as large. Whether the condition holds is an empirical question per image, and our measurements of it were on colour-field images. We retain DDS as a zero-parameter heuristic and make no domination claim.

Hybrid latent-plus-feature payload (HLFW). Adding a per-prompt Gaussian template to the initial latent alongside $g_\phi(z)$ and fusing the two verifier scores by a z-normalised max made the bank-aware L2-PGD that closes the feature head and the template swap that closes the template head each less effective than against its own single head. A blind-spot bound of the form $\beta_{\mathrm{HLFW}} \geq \sqrt{\gamma_{\mathrm{f}}^2 + \gamma_{\mathrm{t}}^2 - 2\rho\gamma_{\mathrm{f}}\gamma_{\mathrm{t}}}$ would follow from a ρ-incoherence and a unit-rate decrease assumption; the latter is not verified for z-scored scores, whose normalisation couples all bank entries, so we report only the empirical observation and only the feature head's certificate for the hybrid checkpoint (Table 12).

E Hyper-parameters and compute

Model. Selector: attention-weighted Top-K pooling over CLIP ViT-L/14 token states (K=4, threshold τ=0.14, global EOT gate). Injector: linear $768 \to 512 \cdot 2 \cdot 2$, four transposed-conv blocks $512 \to 256 \to 128 \to 64 \to 4$ (kernel 4, stride 2), `Tanh` output so $\|\Delta\|_\infty \leq 1$; λ=0.18; γ=1 (training) / ∞ (no-clamp inference). Decoder: ResNet-18 (ImageNet init) with a 768-d ℓ_2-

normalised head. Generator: SD-1.5, 16-step DDIM, guidance 7.5, 256^2, fp16, frozen weights, differentiable rollout with per-step checkpointing.

Training (base). 120k steps in three stages (30k, 60k, 120k), batch 8, AdamW lr 2.5×10^{-5} (final stage), weight decay 10^{-5}, β=0.35, InfoNCE temperature 0.07 with 127 negatives. Attack stack at p=0.5 per image: resize scale $[0.6, 1.4]$, JPEG quality $[50, 95]$, brightness $[0.75, 1.25]$, Gaussian blur $\sigma \in [0.1, 1.5]$ (kernel 5), salt-and-pepper rate $[0, 0.01]$. Prompts: 8,000 each from COCO captions, Parti, DiffusionDB, DrawBench, LAION-Aesthetics, Midjourney, an art-domain and a photo-realistic set (64k); held-out pool 40,811 prompts drawn with a different seed and de-duplicated against training.

Evaluation. Prompt seeds $\{42, 43, 44\}$ draw N=128 prompts and the base latents. Controls (Section 4.5): CLIP ViT-L/14 image and text towers for CLIP score and zero-shot retrieval (images bicubically resized to 224); attack-suite numbers average three draws of the training attack stack; hard-negative neighbours by cosine in z-space over the held-out pool with the 128 bank prompts excluded. Certification: n_0=100, n=10^4, α=10^{-3}, noise batches of 512, independent noise for selection and certification.

Compute. One 96 GB GPU. Base training: $\approx$ 2.6 s/step at 16 DDIM steps with checkpointing on a shared GPU ($\approx$ 43 h for the final 60k-step stage). Certification at n=10^4: $\approx$ 17 min per 128-image seed. Control suite: $\approx$ 1 min per seed at N=128. Clamp-free fine-tuning (Section 4.3): 2k steps at $\approx$ 2.9 s/step (two concurrent jobs), same optimiser, learning rate, attack stack and data as the final stage of the base, with γ=∞ and $\lambda \in \{0.18, 0.09\}$ (`configs/prw_ultra_ft_*.yaml`). Mini-max fine-tuning: $\approx$ 4 s/step (three concurrent jobs on the GPU).

F MINI-MAX FINE-TUNING RECIPE

Inner player: an instance-conditioned policy (three stride-2 conv blocks with GroupNorm and SiLU, global pool, linear head) outputs five raw parameters mapped by `tanh` to the physical ranges resize $[0.4, 1.6]$, JPEG $[10, 95]$, brightness $[0.5, 1.5]$, blur $\sigma \in [0.1, 4.0]$, salt-and-pepper $[0, 0.1]$; the attack is applied with differentiable surrogates (JPEG proxy, straight-through mask) and the policy maximises $1 - \cos(D_\psi(A(I_w)), z)$ minus a penalty on relative ℓ_2 distortion above a 0.05 budget. Outer player: the PRW model minimises the training loss on the attacked image with probability 0.45 and on the clean image with probability 0.55; 50 warm-up steps without the attacker; 2,000 steps at batch 8; lr 3×10^{-5} for the PRW parameters and 10^{-3} for the policy; gradients of the two players are computed separately with `autograd.grad` so neither leaks into the other's parameters. The text encoder is inherited from the checkpoint (CLIP ViT-L/14); the generator is the differentiable frozen-weight rollout with the clamp disabled (γ=∞), matching the evaluation regime. Evaluation: 5-step PGD in the attack-parameter space against the verifier's Top-1, paired prompts and latents between the base and the fine-tuned model, three seeds.

G PROTOCOL MAP AND DIAGNOSTICS

Table 13: Protocol map: which checkpoint, gradient path, clamp, sampler and seed semantics each table was measured under. "Prompt draw": seeds $\{42, 43, 44\}$ draw the same held-out N=128 banks in every table; "bank draw": held-out bank redrawn per evaluation. Rows with different clamp or gradient settings are not comparable.

checkpoint	injector grad	clamp at eval	sampler	seeds index	tables
detached rollout, 30k	cut	γ=1	SD-1.5, 4-step	bank draw	Table 1 (col. 1); Table 12
differentiable, 16k	flows	γ=1	SD-1.5, 4-step, 256^2	bank draw	Table 1 (col. 2); Table 12
base, 120k	flows	γ=1	SD-1.5, 16-step, 256^2	bank draw	Table 1 col. 3; Table 9
base, 120k	flows	γ=1	SD-1.5, 16-step, 256^2	prompt draw + latents	Table 5 bottom; Table 2 row 1; Table 11
base, 120k	flows	none	SD-1.5, 16-step, 256^2	prompt draw + latents	Table 5 top; Tables 2, 3, 6 and 7; Table 8 row 1; Table 4
+ clamp-free fine-tune 2k	flows	none	SD-1.5, 16-step, 256^2	prompt draw + latents	Table 4
+ mini-max 2k	flows	none	SD-1.5, 16-step, 256^2	prompt draw + latents	Table 8 row 2

Three one-line diagnostics. (i) *Gradient probe*: after one training step, print $\|\nabla_\phi \mathcal{L}\|$ per module; a frozen generator implemented with a detached rollout gives exactly 0 for the injector while the

decoder's norm is $\mathcal{O}(1)$. (ii) *Unwatermarked seed*: render $I_0 = G(x)$ next to I_w for the same x; a clamp or any other latent transform that destroys content is visible in one figure and in no retrieval metric. (iii) *Certificate ceiling*: print $R_{\max} = \sigma\Phi^{-1}(\alpha^{1/n})$ next to every certified table; any populated cell beyond it means the bound is wrong. Our own first implementation failed all three, which is why we report them.

H PROMPTS IN FIGURE 2

1. noisy photograph of outside tiny homes in liminal underground, office ceiling panels, retrofuturism, brutalism, minimalist, cinematic, soft vintage glow
2. eleanor tomlinson in the dark forest. andreas achenbach, artgerm, mikko lagerstedt, zack snyder, tokujin yoshioka
3. doodle vector art of a white electric guitar against a fully black background, simple. A k_euler_ancestral
4. a cute magical girl with a crown in the style of sakura card captor
5. Clothes concept by Moebius, full body, trending on artstation, artstationHD, artstationHQ, 4k, 8k
6. frustrated student studying, in style of studio ghibli,
7. inside a huge palace, futuristic architecture, many windows, with sun and sea and trees, realistic, ultra-detailed, octane 4k render, hdr shot, bright and clear
8. A portrait photo of Max Caulfield, from Life is Strange, on the Ravenclaw common room in Hogwarts. in game capture