IRIS: Visual-Semantic Binding for Forgery-Resistant Watermarking of Diffusion Images

Xiaoyan Feng[1], Zheng Gao[2], Tong Guan[1,3], Rui Bao[2], Bokang Zeng[2], Xiaoyu Li[2], Jiaojiao Jiang[2]

[1]Griffith University, [2]University of New South Wales, [3]Zhejiang University

arXiv:2608.03539v1 [cs.CV] 4 Aug 2026

Abstract

Most in-generation diffusion watermarks embed patterns independent of the image that carries them, and attackers transplant the marks onto images the generator did not produce, resulting in forgery. Binding the mark to visual semantics prevents such transplantation, yet existing bindings anchor to a proxy image rather than the image they mark. Realizing visual-semantic binding inside generation faces two challenges. The mark derives from the image itself yet enters the sampling trajectory before that image exists, and may itself shift the semantics it binds. The binding also meets opposite sensitivity demands, breaking under semantic change while holding through common processing. We present **IRIS**, a training-free watermarking scheme that embeds an **I**ntrinsic **R**ing **I**dentifier from **S**emantics. IRIS reads a content code from the non-watermarked generated image, derives a one-time ring from the code and a secret key, returns to the final low-noise steps of the same trajectory and blends the ring in, after the semantics it binds are settled. To meet the opposite sensitivity demands, the code is read through a canonicalization shared between embedding and detection, holding through common distortions and mild regeneration while flipping under semantic change. Detection recomputes the ring from the query image and the key alone, and the mark therefore fails on a foreign or spliced image, with acceptance tracking semantic displacement. On three prompt datasets IRIS detects reliably and stays close to its same-seed non-watermarked counterpart, a fidelity prior in-generation marks do not reach. While forgeries transfer fixed-pattern marks and regeneration strips post-hoc marks, IRIS alone among the compared marks withstands both.

1 Introduction

Latent diffusion (Rombach et al. 2022) underlies most large-scale AI image generation, powering Stable Diffusion and DALL-E 3 (Betker et al. 2023), and flow models such as FLUX inherit its latent space (Labs et al. 2025). Diffusion images are hard for people and forensic detectors to tell from real images (Lu et al. 2023), and the unsafe imagery these models return (Qu et al. 2023) makes tracing images to their source a pressing concern for platforms, regulators, and the public (Coalition for Content Provenance and Authenticity 2026; European Parliament and Council of the European Union 2024). Watermarking has become a necessary response. Prevailing schemes embed a pattern into the latent along the DDIM trajectory, weaving the mark into the image (Wen et al. 2023; Yang et al. 2024b). Yet these marks are forgeable without the key (Yang et al. 2024a; Müller et al. 2025). Attackers can distill a mark by averaging watermarked examples and stamp it onto chosen content (Yang et al. 2024a), or imprint one from a single reference image without recovering the pattern (Müller et al. 2025).

Binding the mark to visual semantics promises to make the mark forgery-resistant by requiring verification to match the hidden mark against the visual semantics of the image, and the mark thereby cannot be transplanted to images it does not match. Binding the in-generation mark to visual semantics for diffusion images faces two challenges. First, it faces *causal entanglement*. The mark depends on the semantics of the image it binds to, yet it is injected mid-sampling before that image exists, and the perturbation it introduces may itself alter those semantics as it propagates along the trajectory, invalidating its own derivation. Second, it faces *opposite sensitivity* requirements for semantic and non-semantic changes. Verification should hold under common distortions such as JPEG compression and fail once the visual semantics change. This failure is what flags a forged or altered image.

The closest semantic binding, SEAL (Arabi et al. 2025), sidesteps the causal entanglement rather than solving it. It hashes the caption embedding of a pre-generated *proxy* image with SimHash, regenerates the delivered image from an initial noise latent constructed to carry the hash bits, and the mark is thereby bound to the proxy's caption rather than to the delivered image itself. Captions are a coarse summary, and images sharing one caption can share the mark. SEAL thereby remains vulnerable to mark transplantation.

We present **IRIS**, a training-free in-generation watermarking scheme that embeds an **I**ntrinsic **R**ing **I**dentifier from the **S**emantics of the delivered image. To address the causal entanglement, IRIS derives a phase ring from the CLIP embedding of the clean image at the end of the sampling trajectory and blends the ring back into the trajectory's later DDIM steps at an energy matched to the latent's own spectrum, leaving the visual semantics unchanged and letting the image itself carry the mark. To address the opposite sensitivity requirements, IRIS reads the CLIP embedding through a canonicalization pipeline at both embedding and verification, keeping the reading stable under common distortions such as JPEG compression, and derives the mark from a keyed hash of the embedding, making it sensitive to semantic changes. Measured along the CLIP image-embedding angle, IRIS's

acceptance falls as visual displacement grows, whereas a caption-level binding stays open to targeted edits that leave the caption unchanged. Our contributions are threefold.

- We identify two challenges in binding the in-generation mark to visual semantics for diffusion images, causal entanglement and opposite sensitivity requirements.
- We propose IRIS, a training-free watermark whose one-time ring is derived from the trajectory's endpoint by a keyed hash of a canonicalized CLIP embedding, blended back at an energy matched to the latent's spectrum.
- We evaluate IRIS on three prompt datasets, where it achieves detectability, fidelity, and security at once. It preserves the generated content at PSNR 30.2 while in-generation baselines redraw it at 14.4 or below, and holds the lowest attack success rate averaged over the four attacks. We trace the retention decay under semantic displacement to the code recomputed at verification, which drifts and no longer derives the embedded ring.

2 Related Work

In-Generation Watermarks for Diffusion Images. Watermarking for diffusion images can be divided into post-hoc marking and in-generation marking. Post-hoc marks (Fernandez et al. 2023; Cox et al. 2007) preserve the image content, but they sit shallow and are easily removed by add-noise-then-denoise regeneration. In-generation marks inject a pattern into a latent along the sampling trajectory and read it back by inversion, making them withstand regeneration removal (Zhao et al. 2024). Tree-Ring (Wen et al. 2023) writes a ring into the Fourier spectrum of the initial noise, and RingID (Ci et al. 2024) extends it with multi-channel keys. Gaussian Shading (Yang et al. 2024b) maps a secret message to a Gaussian-preserving latent. PRC-Watermark (Gunn, Zhao, and Song 2025) seeds the initial noise from pseudorandom codes. ROBIN (Huang, Wu, and Wang 2024) relocates the mark to an intermediate latent with a hidden guiding prompt. However, these patterns are independent of their carrier images, enabling forgeries which make attacker-chosen content read as watermarked. Müller et al. (2025) invert a watermarked image and regenerate attacker-chosen content from the recovered latent, carrying the mark into the new image. Saberi et al. (2024) query the system with a noise image and transfer the resulting mark to a natural image.

Content-Bound Watermarks. These forgeries exploit the same weakness, a mark that does not depend on the image carrying it. Content-bound watermarks counter them by binding the mark to its carrier image. Existing designs differ in when they embed the mark and where they anchor this binding. SEAL (Arabi et al. 2025), SemBind (Zhang et al. 2026), and SLICE (Gao et al. 2026) derive their bindings from prompt-conditioned auxiliary images rather than the delivered image, requiring an additional full generation. SEAL hashes the auxiliary image's caption embedding. SemBind trains a private semantic masker to assign similar codes to images from the same prompt. SLICE uses a VLM to extract four semantic descriptors from the auxiliary image. SafeTR and Dual-Guard instead embed during generation but rely on per-image records. SafeTR links a content-independent latent index to a prompt stored in an external record (Xie et al. 2026b). Dual-Guard writes a registered fingerprint into the initial noise and the final latent and stores a reference latent in its main setting (Xie et al. 2026a). MetaSeal derives a cryptographic signature from the completed image and embeds it post hoc with an invertible neural network (Zhou et al. 2026). IRIS derives its binding *in place* from the delivered image's visual semantics, along the trajectory that produces it, without watermark-specific training, an auxiliary generation, per-image records, or a post-hoc embedding pass, and its verification is therefore blind and stateless.

3 Preliminaries

Latent diffusion and DDIM inversion. A latent diffusion model generates an image by denoising in the compressed latent space of a VAE (Rombach et al. 2022). Given a text prompt p and an initial Gaussian latent z_T, the model's denoising network predicts the noise at each step, and a sampler uses these predictions to produce a sequence of T progressively cleaner latents ending at z_0. The model's VAE decoder $\mathcal{D}$ then maps z_0 to an image $x = \mathcal{D}(z_0)$. DDIM is a deterministic sampler that introduces no additional random noise between denoising steps. A fixed prompt p and initial latent z_T therefore determine the denoising trajectory.

DDIM inversion applies the DDIM updates in the opposite direction to estimate this trajectory from an image. Given an image x, the VAE encoder $\mathcal{E}$ maps it to $\hat{z}_0$. When the original prompt is unavailable, DDIM inversion conditions the denoising network on an empty prompt and proceeds from $\hat{z}_0$ through increasingly noisy states to estimate $\{\hat{z}_t\}_{t=0}^{T}$.

Ring watermarks in diffusion latents. Ring watermarks are in-generation watermarks that embed a pattern into the Fourier coefficients of a diffusion latent and recover it from the generated image through DDIM inversion (Wen et al. 2023; Huang, Wu, and Wang 2024; Ci et al. 2024). For one channel of a latent z_t, the shifted two-dimensional Fourier spectrum places the zero frequency at its center. Let ω index a Fourier coefficient, and let $|\omega|$ denote its distance from the zero-frequency center. An inner radius r_{lo} and an outer radius r_{hi} define the radial support

$$\mathcal{R} = \{\omega : r_{\mathrm{lo}} \leq |\omega| \leq r_{\mathrm{hi}}\}. \tag{1}$$

When $r_{\mathrm{lo}} = 0$, the support is a centered low-frequency disk, while $r_{\mathrm{lo}} > 0$ gives an annular band. A seed s determines a complex pattern W_s supported on $\mathcal{R}$. The band $\mathcal{R}$ specifies where the ring is written, while W_s specifies the coefficient values written there. During generation, the pattern is embedded on $\mathcal{R}$ at one or more latent states, and the remaining sampling steps map the modified latent to a watermarked image. Given a query image, DDIM inversion estimates the corresponding latent states, and the detector compares the coefficients observed on $\mathcal{R}$ with the expected pattern W_s.

Threat model. A provider uses a diffusion model and a secret key k to generate watermarked images, which are then released publicly. Given a query image and k, the provider verifies whether it generated that visual content. Verification is

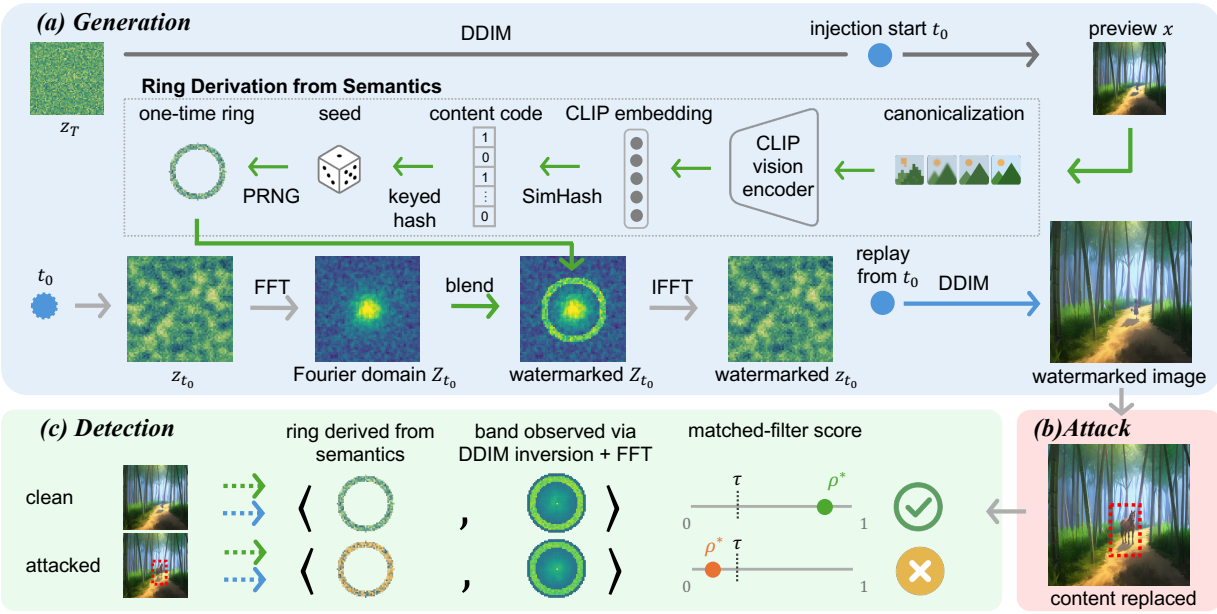

Figure 1: **The IRIS framework. (a)** Generation turns the canonicalized CLIP embedding of the preview into a keyed one-time ring and blends it into the late DDIM steps of the same trajectory. **(b)** An attack replaces part of the content. **(c)** Detection correlates the ring derived from the query's semantics against the observed band. The replacement moves the recomputed code, the keyed hash maps it to a different ring, and the score falls below τ, while the clean image verifies.

blind because it requires no corresponding non-watermarked image. It is also stateless because it retains no per-image prompt, sampling seed, latent state, or metadata record.

The adversary has access to the diffusion pipeline and the watermarking and verification algorithms, while only k remains secret and neither detector scores nor decisions are exposed. The adversary may analyze or modify released images through inversion, regeneration, averaging, or optimization.

A removal attack starts from a released watermarked image and attempts to make it fail verification while preserving its visual content. A forgery attack attempts to make an image pass verification even though the provider did not generate its visual content, the costlier failure. Watermarked images should remain accepted under content-preserving JPEG compression and moderate levels of blur and additive noise.

4 Method

IRIS is an in-generation watermarking scheme built on *visual-semantic binding*, where retention of the mark tracks displacement of the visual semantics of the image that carries it. From a released image and the secret key alone, the provider can verify that it generated this content. Verification requires the ring derived from the image's semantics to match the band observed through DDIM inversion in the Fourier domain. A forged image therefore fails verification, whether it carries a transplanted mark or replaces part of the content.

4.1 Ring Derivation from Semantics

IRIS derives a one-time ring from the visual semantics of the image, following the derivation chain in Figure 1. The image is canonicalized and read into a CLIP embedding, the embedding is quantized under the key into a content code, and a keyed hash of the code seeds the ring.

To ensure the ring stays derivable from a released image under common processing, the **CLIP embedding** (Radford et al. 2021) is read from a **canonicalized** image. The canonicalization normalizes the image's luminance, then blurs the image heavily and recompresses it at a low JPEG quality. The embedding is averaged over a few copies of the image that are randomly re-degraded and then canonicalized, smoothing residual jitter. The same canonicalization runs at embedding and verification, mapping an image and its JPEG-compressed or brightened copy to close CLIP embeddings.

A keyed **SimHash** (Charikar 2002) ϕ_k quantizes this embedding into a short binary **content code**

$$c = \phi_k(x) = \mathbb{1}[\, P_k(E(x) - \mu) > 0 \,] \tag{2}$$

where $E(x)$ is the canonicalized embedding of the image x, $\mathbb{1}[\cdot]$ applies elementwise, and the rows of P_k are hyperplane normals drawn pseudorandomly from the secret key k. Subtracting μ, the cone center precomputed and shared by embedding and verification, spreads the embeddings around the origin, and each hyperplane then splits the images into two balanced sides. SimHash keeps nearby embeddings on nearby codes, and altering the content shifts the embedding far and lands it on a distant code. The farther the embedding drifts, the larger the Hamming drift of the code c.

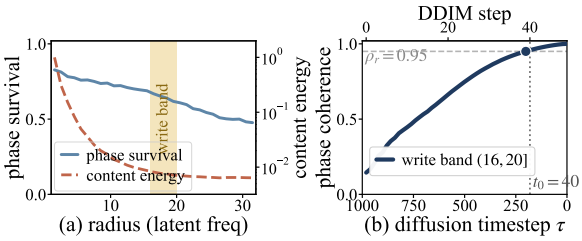

Figure 2: **(a)** Log-scaled content energy per coefficient and phase survival under Gaussian noise at σ=0.1 across radial frequency. **(b)** Write-band phase coherence ρ_r over DDIM steps, with the threshold dashed and injection start dotted.

To keep the mapping from an image to its ring unlearnable from released watermarked images, we map the code to the ring's **seed** through a **keyed hash** $g(k, c) = \mathrm{HMAC}_k(c)$. Flipping a single bit of the code produces an unrelated seed, and the output is unpredictable without the key. Averaging released images distills no common pattern, since each image carries a ring of its own. The seed s expands into the **ring**

$$W_s(\omega) \;=\; e^{i\theta_\omega}, \quad \omega \in \mathcal{R}, \qquad \theta_{-\omega} \;=\; -\theta_\omega, \tag{3}$$

where ω indexes the Fourier coefficients inside $\mathcal{R}$, a mid-frequency annulus of one latent channel, and a **PRNG** expands the seed s into the angles θ_ω. The unit modulus makes the ring phase-only, and the anti-symmetry $\theta_{-\omega} = -\theta_\omega$ makes it Hermitian. Because the spectrum of the real-valued latent is conjugate symmetric, a Hermitian ring survives the write whole, while a pattern without this symmetry loses its anti-symmetric half (Ci et al. 2024). The ring's magnitudes carry no information, and the phase survives common distortions better than the magnitude (Ruanaidh, Dowling, and Boland 1996). A Gaussian blur kernel is symmetric, and its transfer function is therefore zero-phase, attenuating the ring's magnitudes while leaving its phases in place. JPEG compression and contrast adjustment act on the spectrum as multiplicative scaling of the magnitudes. The ring therefore remains stable through the DDIM inversion at detection.

4.2 Ring Injection and Detection

Injecting a semantic-bound ring faces the challenge of causal entanglement. The ring derives from the final image, yet it enters the latent before sampling finishes, and the injection may itself shift the semantics of that image. IRIS therefore keeps the watermarked image on nearly the same code as its non-watermarked counterpart, by injecting faintly and late.

Band placement. The band $\mathcal{R}$ is a narrow annulus placed past the image's content yet within reach of robust detection. Figure 2(a) motivates the selection with two measurements on 500 images. Content energy per coefficient concentrates in the low bands, where a write would alter the image. Phase survival under additive white Gaussian noise falls as the radius grows, and the outer bands lose the ring first. The ring therefore sits between the two, in the annulus holding the least content energy. It also stays narrow. The ring is rescaled to the band's energy, and a wider band takes proportionally more injected energy, at a direct cost in PSNR.

Injection timing and strength. The ring is blended into the late DDIM steps at an energy matched to the band's content. Figure 2(b) tracks $\hat{z}_0$, the one-step prediction of the final latent, and reports its phase coherence ρ_r with that latent on the write band. The band does not reach $\rho_r \geq 0.95$ until DDIM step 39, and a ring written earlier would be overwritten by the steps that follow. After that point only low-energy detail is still forming there, and a late blend fuses the ring into it. IRIS therefore reuses the trajectory's endpoint, closing the causal loop. The clean pass is the generation itself, its latent stashed at the injection start t_0, just after the write band settles, and only the final $T{-}t_0$ steps are replayed with the ring blended in, a small sampling overhead, timed in Table 2. Each replayed step blends the ring with a weight α_t, which peaks at α_0 at the injection start and decays to zero at the final image, as Equation 4 shows.

$$Z_{\mathcal{R}} \leftarrow (1-\alpha_t) Z_{\mathcal{R}} + \alpha_t \gamma_t W_s, \quad \alpha_t = \alpha_0 \cdot \frac{T-t}{T-t_0}, \tag{4}$$

where $Z_{\mathcal{R}}$ holds the Fourier coefficients of the current latent inside the band $\mathcal{R}$, and γ_t is their root-mean-square modulus. Frequencies outside $\mathcal{R}$ stay untouched.

Matched-filter detection. Detection reads the ring derived from the semantics and correlates it against the band observed from the query image. The query runs through the same canonicalization as embedding, and the averaged embedding gives the recomputed code $\hat{c}$. Although canonicalization absorbs most of a distortion, the residual jitter can still flip a few bits of the recomputed code. The code therefore expands into a candidate set, and the keyed hash maps each candidate to its ring. Since the ring sits weakly below the band's own content, IRIS scores every candidate ring with the normalized matched filter of spread-spectrum watermarking (Cox et al. 1997, 2007), and the highest score is the detection score.

The ring is read back by DDIM inversion, retracing the $T{-}t_0$ injected steps and yielding $Z^{(t)}$, the band coefficients observed at step t. Each candidate c' is scored by

$$\rho(c') = \frac{1}{T-t_0} \sum_{t=t_0}^{T-1} \frac{\left|\langle W_{s'},\, Z^{(t)} \rangle\right|}{\|W_{s'}\|\, \|Z^{(t)}\|}, \quad s' = g(k,\, c'), \tag{5}$$

where g is the keyed hash, k is the secret key, and $\langle \cdot, \cdot \rangle$ is the complex inner product. The ring is blended at every replayed step, and the average pools these repeated views against the lossy inversion. The average $\rho(c')$ rises above the noise level only when the candidate's ring is the one present in the band.

The candidate set covers scattered and concentrated drift,

$$\mathcal{C}(\hat{c}) = \{c' : d_H(c', \hat{c}) \leq t_s\} \cup \{\hat{c} \oplus e_S : S \subseteq \mathcal{L}\}, \tag{6}$$

where the first part collects the codes within Hamming distance t_s of $\hat{c}$, covering scattered drift, and the second part flips every subset S of $\mathcal{L}$, the L bits whose projections lie closest to their hyperplanes, covering drift concentrated on these low-margin bits, with e_S the indicator vector of S and $\oplus$ the bitwise XOR. The candidate set raises detection at little cost in false positives, since the keyed hash maps distinct

	Method	Detectability	Generation Quality			Forgery ASR ↓		Removal ASR ↓		ASR↓
		TPR@1% FPR↑	PSNR↑	SSIM↑	FID↓	Reprompt	PGD	Regeneration	PGD	Mean
Post-hoc	StableSig	1.00	29.6	0.87	23.8	0.01	0.04	0.98	0.00	0.26
	DwtDctSvd	1.00	38.6	0.98	23.5	0.01	0.00	0.75	0.00	0.19
In-generation	Tree-Ring	1.00	14.4	0.50	25.2	0.94	0.10	0.05	0.19	0.32
	GS	1.00	9.8	0.28	26.8	1.00	1.00	0.00	0.00	0.50
	SEAL	0.99	9.5	0.25	24.2	0.70	0.01	0.29	0.39	0.35
	IRIS (ours)	**0.99**	**30.2**	**0.92**	**23.7**	**0.01**	**0.01**	**0.03**	**0.09**	**0.04**

Table 1: Results pooled across three prompt datasets. ASR: forgery acceptance/watermarked-image rejection; PSNR/SSIM: same-seed non-watermarked counterparts; FID (Heusel et al. 2017) against the 5,000-image MS-COCO validation set.

Figure 3: Fidelity under one DiffusionDB prompt. IRIS stays pixel-close to its same-seed counterpart, per-channel difference amplified $10\times$ with hue marking the channel, while the in-generation baselines redraw the image.

codes to unrelated rings, and a mismatched ring stays at the noise level. The final statistic is $\rho^\star = \max_{c' \in \mathcal{C}(\hat{c})} \rho(c')$, and the maximizing candidate is the recovered code. The image is declared watermarked when $\rho^\star$ clears the threshold τ.

5 Experiments

Setup. We generate with Stable Diffusion 2.1-base (Rombach et al. 2022) at T=50 DDIM steps and guidance 7.5 on three prompt datasets, Stable-Diffusion-Prompts (Gustavosta 2022), DiffusionDB (Wang et al. 2023), and MS-COCO (Lin et al. 2014), covering curated prompts, user prompts, and image captions, 500 prompts each. Each prompt produces a watermarked image and its non-watermarked counterpart from the same seed. The threshold τ is the 1%-FPR quantile of detection scores on non-watermarked images.

IRIS configuration. The ring occupies band $(16, 20]$ of latent channel 0, injected from t_0=40 with peak strength α_0=0.3. The content code is a 24-bit keyed SimHash on CLIP ViT-B/32 with the frozen cone-center. Canonicalization applies luminance normalization, Gaussian blur with σ=2, JPEG at quality 25, and K=8 test-time augmentations. Detection uses radius t_s=3 with L=7 low-margin bits and inverts only the 10 steps down to t_0.

Baselines. We compare against three in-generation watermarks, Tree-Ring (Wen et al. 2023), Gaussian Shading (GS) (Yang et al. 2024b), and SEAL (Arabi et al. 2025). SEAL is the closest baseline and uses its content-aware detector. We include two post-hoc watermarks, Stable Signature (Fernandez et al. 2023) and DwtDctSvd (Cox et al. 2007; Navas et al. 2008). Each baseline runs on the same backbone and prompts, with its threshold at the same 1% FPR.

5.1 Detection, Fidelity, and Security at Once

Table 1 reports detectability, fidelity, forgery resistance, and removal resistance. A forgery attack makes an image the provider did not watermark pass verification, and a removal attack makes a watermarked image fail it. Reprompt follows the inversion-reuse forgery of Müller et al. (2025). The released image is DDIM-inverted to its initial latent and regenerated with a different prompt, carrying the mark onto new content. PGD forgery (Madry et al. 2018) adapts the single-image latent-matching attack of Jain et al. (2025) to a bounded form, perturbing an MS-COCO cover image within an ℓ_∞ budget of $8/255$ for 1000 iterations, pushing its VAE latent toward a watermarked image's latent while a fidelity term keeps the cover's content. SEAL (Arabi et al. 2025) evaluates the same attack. Regeneration follows the add-noise-then-denoise removal attack of Zhao et al. (2024). The watermarked image is noised to step 50 and denoised, redrawing the fine detail carrying the mark. PGD removal reuses that budget and pushes a watermarked image's latent toward that of its same-seed non-watermarked counterpart, an oracle target that makes the ASR an upper bound.

IRIS achieves all four properties at once, while each baseline fails at least one. IRIS detects at TPR 0.99 and keeps PSNR 30.2 and SSIM 0.92 against its same-seed non-watermarked counterpart, with the lowest FID among the in-generation marks. In Figure 3 the IRIS image is visually indistinguishable from the counterpart and their amplified difference is nearly black, while each in-generation baseline delivers a different image under the same prompt. Gaussian Shading and SEAL have no same-seed counterpart to hold fixed, because their marks construct the initial noise itself. The Tree-Ring sample starts from the counterpart's initial

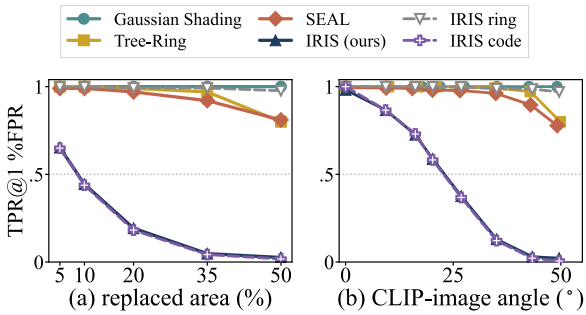

Figure 4: Binding mechanism. **(a)** TPR@1%FPR versus the replaced fraction. **(b)** TPR@1%FPR of the same detectors versus the CLIP-image displacement of the edits.

noise, verified bit-identical before injection, and still moves the lighthouse at PSNR 18.9. IRIS holds both forgery attacks at ASR 0.01 and averages ASR 0.04 across the four attacks, the lowest among the compared marks.

The post-hoc marks preserve content, and PGD removal leaves them untouched because their signatures do not enter the inverted latent, but regeneration strips them at ASR 0.75 or above. The three in-generation baselines redraw the image at PSNR 14.4 or below, and Gaussian Shading and Tree-Ring let reprompt forgeries pass at ASR 0.94 or above. SEAL binds to a proxy caption instead of the image, lets reprompt forgeries pass at ASR 0.70, and loses the most to regeneration and PGD removal among the in-generation marks.

5.2 Measuring Visual-Semantic Binding

IRIS resists forgery with a one-time mark bound to the visual semantics of the delivered image. Verification requires the ring derived from those semantics to match the band observed through DDIM inversion. We quantify that binding by measuring whether the derived ring or the observed band does the letting go as replacement grows.

Figure 4(a) shows TPR@1%FPR as a patch from an MS-COCO image replaces 5% to 50% of the watermarked image. IRIS rejects more than half the images at a 10% replaced area, while the other in-generation baselines let nearly all pass. We evaluate ring survival by granting the detector the oracle code, testing whether the embedded ring still clears the matched-filter threshold against the observed band. Code survival is evaluated by recomputing the code from the attacked image. The code survival curve overlaps the IRIS TPR curve, while ring survival stays nearly unchanged. The derived ring therefore does the letting go. The observed band stays stable because every coefficient aggregates the entire frame, and a replaced patch shifts the correlation by only about the replaced fraction. The recomputed code drifts because content replacement moves the CLIP embedding and with it the code. Once the code changes, the keyed hash maps it to a different derived ring, and IRIS rejects the image.

Figure 4(b) replots the same replacements against the CLIP-image angle they induce between the original and the attacked image. The code survival curve again overlaps

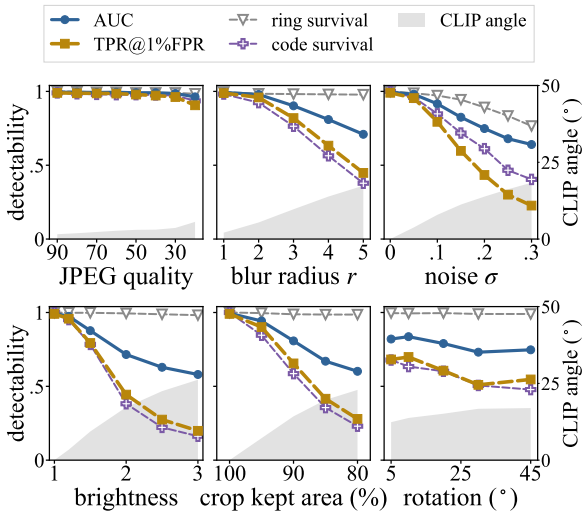

Figure 5: IRIS detectability across six distortions, with ring and code survival and the canonicalized CLIP angle overlaid.

the IRIS TPR curve as the angle grows, while the other in-generation baselines stay high across the sweep. IRIS's acceptance falls below one half beyond ≈23° of displacement, the half-acceptance angle θ_{50}. The two panels together confirm that the code binds the mark to the visual semantics, with acceptance tracking the semantic displacement itself.

The cone is therefore the scheme's security parameter, and the candidate set that absorbs benign drift sets its angle. Two unrelated generated images sit about 49° apart on this axis, and the cone spans roughly half that separation. Our forgery attacks reach the mark from far outside the cone, and an adaptive search for a semantically distinct image whose code still lands in the candidate set remains uncharacterized. We propose the cost of that search as the security notion for content-bound watermarks, comparable across schemes.

5.3 Watermark Robustness

We evaluate the robustness of IRIS under four valuemetric and two geometric distortions, JPEG, blur, noise, brightness, crop, and rotation. For crop, the kept region stays at its original position and the removed border is filled by replicating the region's edge pixels. For rotation, we search over a grid of candidate angles and apply the same search to non-watermarked images so that both sides of the detection threshold go through the same procedure. Figure 5 reports TPR and AUC under the six distortions.

IRIS is robust to JPEG compression, keeping TPR at 0.96 down to quality 30, because both the code and the ring survive. Blur, noise, brightness, and crop keep TPR above 0.9 at moderate strengths such as blur σ=2, noise σ=0.05, and brightness ×1.2. Beyond these levels the TPR decays gradually. Blur, crop, and noise drift the code, brightness drifts it heavily, and noise also erodes the ring. The ring stays flat under blur and brightness, as the phase-only coding predicts. The rotation curve is relatively flat because the angle search

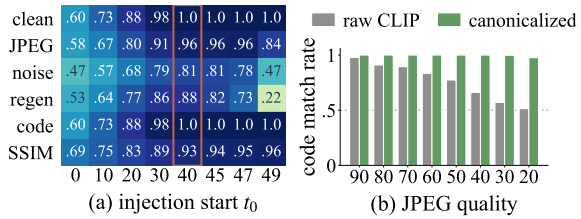

Figure 6: **(a)** Injection-start ablation. **(b)** Canonicalization ablation, raw versus canonicalized readout.

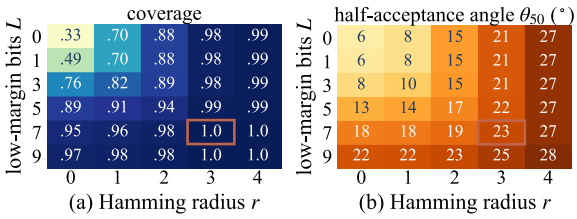

Figure 7: Candidate-set grid, boxed cell = deployed.

re-aligns the image, and the remaining gap comes from interpolation and the corners lost to rotation.

Geometric misalignment is a long-standing synchronization problem for blind watermarking (Petitcolas, Anderson, and Kuhn 1998; Licks and Jordan 2005) rather than a binding failure specific to IRIS. Ring-based marks such as Tree-Ring hold under rotation as-is because their fixed rings are rotationally symmetric, and the same content-independent symmetry lets reprompt forgeries transplant them. A one-time ring keyed to the image gives up that symmetry, and IRIS recovers alignment by the angle search instead.

5.4 Ablations and Efficiency

Injection start and canonicalization. Figure 6(a) sweeps the injection start t_0, pooled over the three prompt datasets, with thresholds recalibrated per t_0 on the non-watermarked pools. Rows track TPR clean and under JPEG quality 25, noise σ=0.1, and regeneration at step 150, along with code survival and SSIM. Starting the injection early hurts detection and fidelity together. The ring is overwritten by the content still forming, and the injection perturbs the semantics the code reads. Every perturbed row peaks at t_0=40. Starting past t_0=45 keeps clean TPR, code survival, and SSIM high, but the TPR collapses under noise and regeneration. We select t_0=40 because it is the deepest start that still leaves the code intact, matching the settling of the write band in Figure 2(b). In Figure 6(b) the code recomputed without canonicalization drifts out of the candidate set under compression alone, falling to 0.51 code survival at JPEG quality 20, while the deployed canonicalized reading holds 0.98 at the same level. The compression robustness reported in Figure 5 is bought by this component.

Candidate set. Figure 7 sweeps the candidate set over the Hamming radius r and the low-margin count L, re-scoring the cached images, with drifted-code coverage in (a) and the half-acceptance angle θ_{50} in (b). The margin mechanism is the workhorse. It targets the bits where drift happens, those with the smallest projection margins, and alone lifts coverage from 0.33 to 0.95 at the deployed L=7. The ball covers drift at arbitrary positions and adds the tail, 0.98 coverage alone at the deployed r=3 and 0.998 jointly. The union stays close to the ball alone, the margin subsets are few and half of them lie inside the ball. Oversizing either axis widens the acceptance cone, by 4° per radius step and ≈1° per additional bit past the deployed sizes. The deployed pair therefore sits at the joint knee, and a deployment that caps the cone at a specified angle reads its maximum (r, L) off panel (b).

Method	UNet steps emb. / det.	per image (s) embed	detect	total (s)
Tree-Ring	50 / 50	1.30 (1.0×)	1.24	2.54
Gaussian Shading	50 / 50	1.30 (1.0×)	1.24	2.54
SEAL	100 / 50	3.89 (3.0×)	2.44	6.33
IRIS (ours)	**60 / 10**	**1.87** (1.4×)	**0.63**	**2.51**

Table 2: Per-image wall-clock cost (seconds, batch 1, RTX 4090, T=50; × relative to one plain generation).

Computational cost. Table 2 times every in-generation mark end to end, each with all the components its scheme requires. The 10-step replay is what binding the mark to the delivered content costs, and among content-aware schemes IRIS pays the least while running cheapest overall.

6 Conclusion

We presented IRIS, a training-free in-generation watermark that binds a one-time ring to the visual semantics of its carrier image. The ring derives from a keyed hash of the canonicalized CLIP embedding at the trajectory's endpoint, enters the late DDIM steps by replay, and is recomputed at verification from the query image and the key alone. Across three prompt datasets IRIS detects reliably, preserves the generated content that the other in-generation marks redraw, and alone among the compared marks withstands both forgery and regeneration. Retention holds under common distortions and decays monotonically as the visual semantics move. Past those levels the limit sits in the code reading, where heavy valuemetric distortion moves the CLIP embedding along the same axis that flags semantic change. A stronger canonicalization or an embedding trained for valuemetric invariance would widen the retained range without touching the binding. The construction needs a deterministic invertible sampler and a frozen perceptual encoder. Deterministic flow samplers invert like DDIM along their ODE trajectories, and injection into them is a further direction. The acceptance cone is the security parameter this framing identifies, and a proxy binding stays above one half under reprompting.

References

Arabi, K.; Witter, R. T.; Hegde, C.; and Cohen, N. 2025. Seal: Semantic aware image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16196–16205.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving Image Generation with Better Captions. Technical report, OpenAI. https://cdn.openai.com/papers/dall-e-3.pdf.

Charikar, M. S. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 380–388.

Ci, H.; Yang, P.; Song, Y.; and Shou, M. Z. 2024. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European conference on computer vision*, 338–354. Springer.

Coalition for Content Provenance and Authenticity. 2026. Content Credentials: C2PA Technical Specification, Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html.

Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; and Kalker, T. 2007. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2 edition.

Cox, I. J.; Kilian, J.; Leighton, F. T.; and Shamoon, T. 1997. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 6(12): 1673–1687.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; and Furon, T. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22466–22477.

Gao, Z.; Yang, Y.; Li, X.; Feng, X.; Fan, H.; Song, Y.; and Jiang, J. 2026. SLICE: Semantic Latent Injection via Compartmentalized Embedding for Image Watermarking. *arXiv preprint arXiv:2603.12749*.

Gunn, S.; Zhao, X.; and Song, D. 2025. An Undetectable Watermark for Generative Image Models. In *The Thirteenth International Conference on Learning Representations*.

Gustavosta. 2022. Stable Diffusion Prompts. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts. Accessed: 2026-07-28.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, H.; Wu, Y.; and Wang, Q. 2024. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963.

Jain, A.; Kobayashi, Y.; Murata, N.; Takida, Y.; Shibuya, T.; Mitsufuji, Y.; Cohen, N.; Memon, N.; and Togelius, J. 2025. Forging and Removing Latent-Noise Diffusion Watermarks Using a Single Image. arXiv:2504.20111.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Licks, V.; and Jordan, R. 2005. Geometric attacks on image watermarking systems. *IEEE multimedia*, 12(3): 68–78.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, Z.; Huang, D.; Bai, L.; Qu, J.; Wu, C.; Liu, X.; and Ouyang, W. 2023. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in neural information processing systems*, 36: 25435–25447.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Müller, A.; Lukovnikov, D.; Thietke, J.; Fischer, A.; and Quiring, E. 2025. Black-Box Forgery Attacks on Semantic Watermarks for Diffusion Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20937–20946. IEEE Computer Society.

Navas, K. A.; Ajay, M. C.; Lekshmi, M.; Archana, T. S.; and Sasikumar, M. 2008. DWT-DCT-SVD based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE '08)*, 271–274.

Petitcolas, F. A.; Anderson, R. J.; and Kuhn, M. G. 1998. Attacks on copyright marking systems. In *International workshop on information hiding*, 218–238. Sprx'ringer.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruanaidh, J.; Dowling, W. J.; and Boland, F. M. 1996. Phase watermarking of digital images. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, 239–242. IEEE.

Saberi, M.; Sadasivan, V. S.; Rezaei, K.; Kumar, A.; Chegini, A.; Wang, W.; and Feizi, S. 2024. Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks. In *The*

Twelfth International Conference on Learning Representations.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2023. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 893–911.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063.

Xie, J.; Ou, C.; Yu, P.; Zhou, X.; Huang, D.; Fei, J.; Shen, Z.; and Xia, Z. 2026a. Dual-Guard: Dual-Channel Latent Watermarking for Provenance and Tamper Localization in Diffusion Images. *arXiv preprint arXiv:2604.19090*.

Xie, J.; Yu, P.; Fei, J.; Zhou, X.; and Xia, Z. 2026b. SafeTR: Verifiable Semantic Tree-Ring Watermark for Diffusion Model Against Forgery Attacks. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9505–9511. IEEE.

Yang, P.; Ci, H.; Song, Y.; and Shou, M. Z. 2024a. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37: 56644–56673.

Yang, Z.; Zeng, K.; Chen, K.; Fang, H.; Zhang, W.; and Yu, N. 2024b. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12162–12171.

Zhang, X.; Yang, Z.; Chen, K.; Ma, L.; Zhang, W.; and Yu, N. 2026. SemBind: Binding Diffusion Watermarks to Semantics Against Black-Box Forgery Attacks. In *Forty-third International Conference on Machine Learning*.

Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.-X.; and Li, L. 2024. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37: 8643–8672.

Zhou, T.; Ding, R.; Liu, G.; Fleming, C.; Kompella, R. R.; Fei, Y.; Xu, X.; and Ren, S. 2026. MetaSeal: Defending against image attribution forgery through content-dependent cryptographic watermarks. *Transactions on Machine Learning Research*.

A Implementation Details

Experimental environment. Generation, inversion, and the matched filter run in half precision, and the gradient forgery takes its steps through a full-precision VAE. Measurements run on a single NVIDIA RTX 4090 under CUDA 12.8, PyTorch 2.7.0, and Python 3.9.

Detection scores and threshold. Each method scores an image on its own scale. We derive each threshold at a 1% false positive rate over a pool of 1500 non-watermarked images in our runs. The IRIS phase correlation lies in $[0, 1]$ with threshold 0.209. The Tree-Ring score is a negative L_1 distance of the inverted noise to the fixed ring, in $(-\infty, 0]$ with threshold -50.5. The Gaussian Shading bit accuracy lies in $[0, 1]$ at a chance level of 0.5 with threshold 0.57. The SEAL score counts the patches whose recovered noise matches the caption-derived hash, in $\{0, \ldots, 1024\}$, and accepts an image when more than one patch matches.

B Examples of Watermarked Images

B.1 Fidelity

Figure 8 places IRIS beside its same-seed non-watermarked images and beside three baselines across six prompts. The non-watermarked counterpart and the IRIS image share an initial latent, leaving them close at the pixel level. The third column shows the pixel-wise difference amplified $10\times$, and the residual stays near black. Tree-Ring, Gaussian Shading, and SEAL start from the same latent yet return a different image, because each writes its mark into the initial noise and lets the trajectory redraw the frame. This shows that IRIS is an in-generation watermark that keeps the main content of the image unchanged.

B.2 Distortion, Forgery, and Removal

Figure 9 tracks the IRIS image through one distortion, three forgeries, and one removal across six further prompts, with the detection score $\rho^\star$ under each image, green above the operating threshold of 0.209 and red below it. The watermarked image scores near 0.85, additive noise leaves it above the threshold on most prompts, and regeneration leaves the mark intact, since the ring occupies a band that the redrawn detail leaves mostly untouched. The three forgeries fall below the threshold on each prompt. Content replacement over a tenth of the frame and reprompt both move the semantics, and the code recomputed at verification derives a different ring, and the correlation drops to the non-watermarked floor. The gradient forgery matches a cover latent to a watermarked latent yet recovers neither the code nor the ring. IRIS answers whether the delivered image is the generated image, rather than whether a fixed pattern is present.

Figures 10, 11, and 12 repeat the six operations for Tree-Ring, Gaussian Shading, and SEAL, each scored by its own detector against its own threshold. The three forgeries separate the methods. Tree-Ring writes a fixed ring into the initial noise, and the ring rides through a spliced patch and through regeneration under a new prompt, and its score therefore stays above the threshold on those images, where the transplanted mark verifies. Gaussian Shading recovers its message from the forgeries as well as the distortions because it is so robust that it verifies content it was not issued for. SEAL binds to the caption, and a donor patch that leaves the caption unchanged keeps its score above the threshold, while reprompt lowers the score without clearing it on most prompts. IRIS drops to the non-watermarked floor under the forgeries, because it reads the code from the delivered image at verification, and the code is sensitive to these attacks.

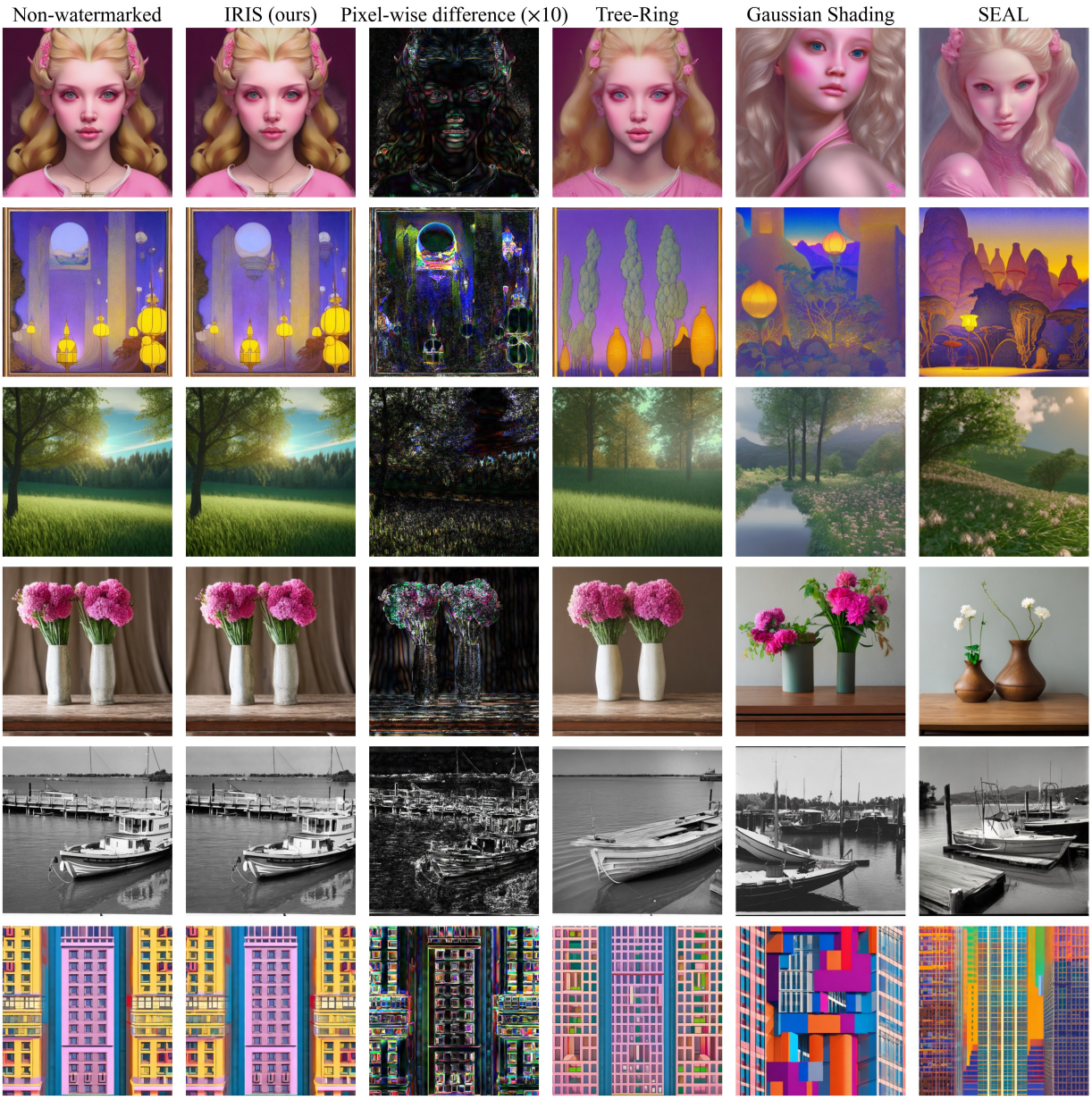

Figure 8: Fidelity across prompts from the three evaluation sources. Each row shows the non-watermarked counterpart, the IRIS image, the amplified pixel-wise difference, and the Tree-Ring, Gaussian Shading, and SEAL images from the same initial latent.

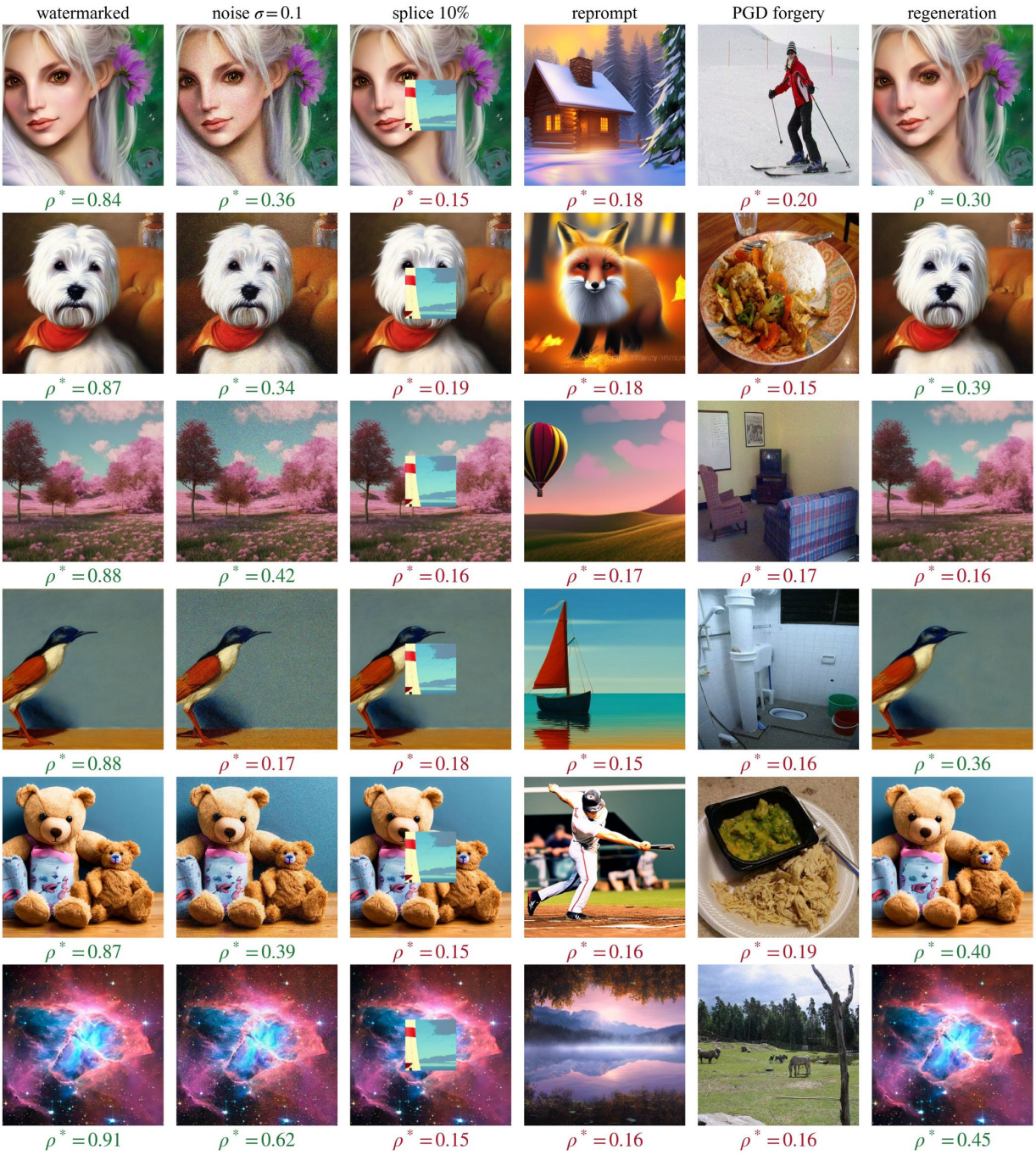

Figure 9: IRIS under distortion, forgery, and removal across prompts. Each row shows the watermarked image, additive noise, content replacement, reprompt, gradient forgery, and regeneration, with the detection score $\rho^\star$ printed under each image, green when the mark is retained and red when the image is rejected.

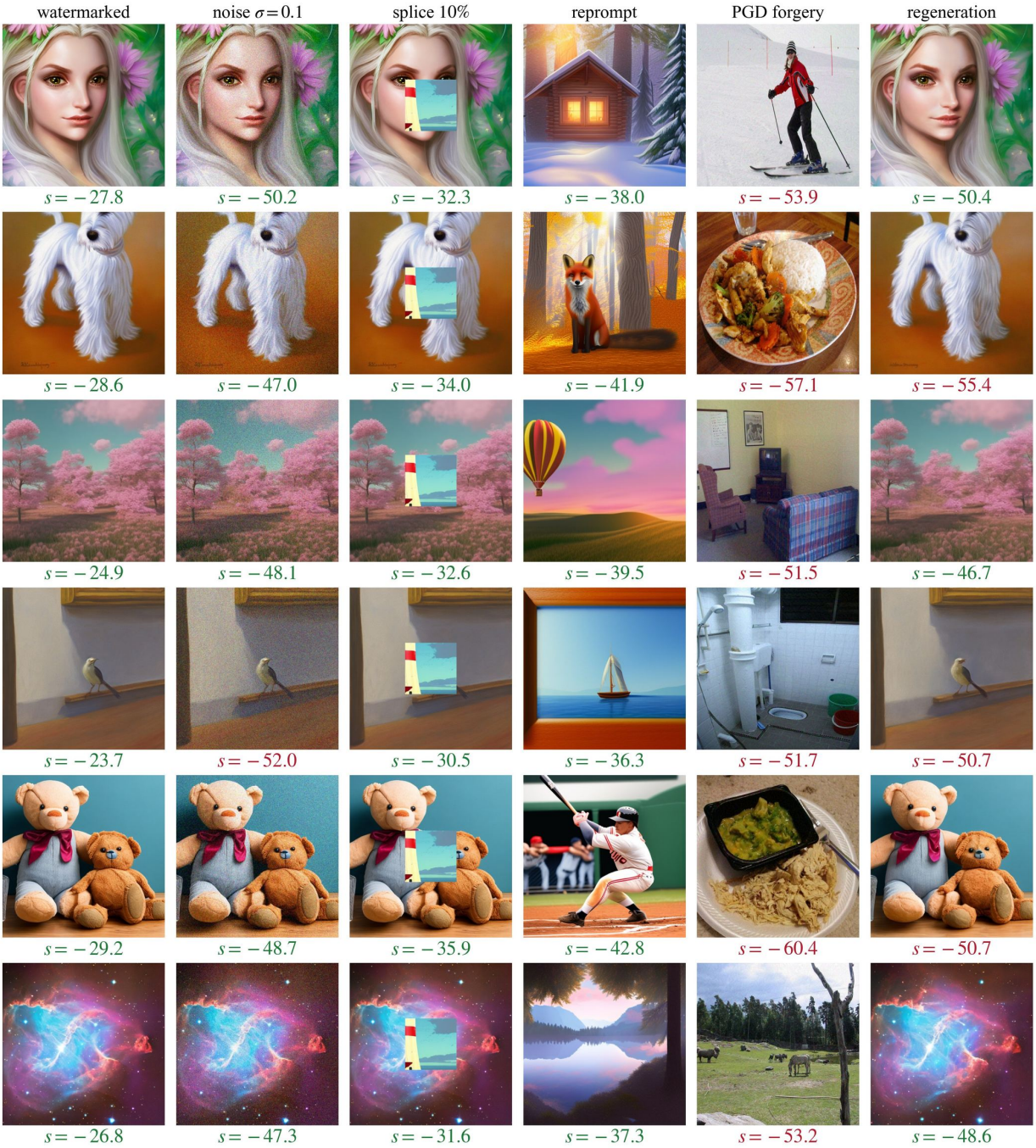

Figure 10: Tree-Ring under the same operations, scored by its own detector, green above its threshold and red below it.

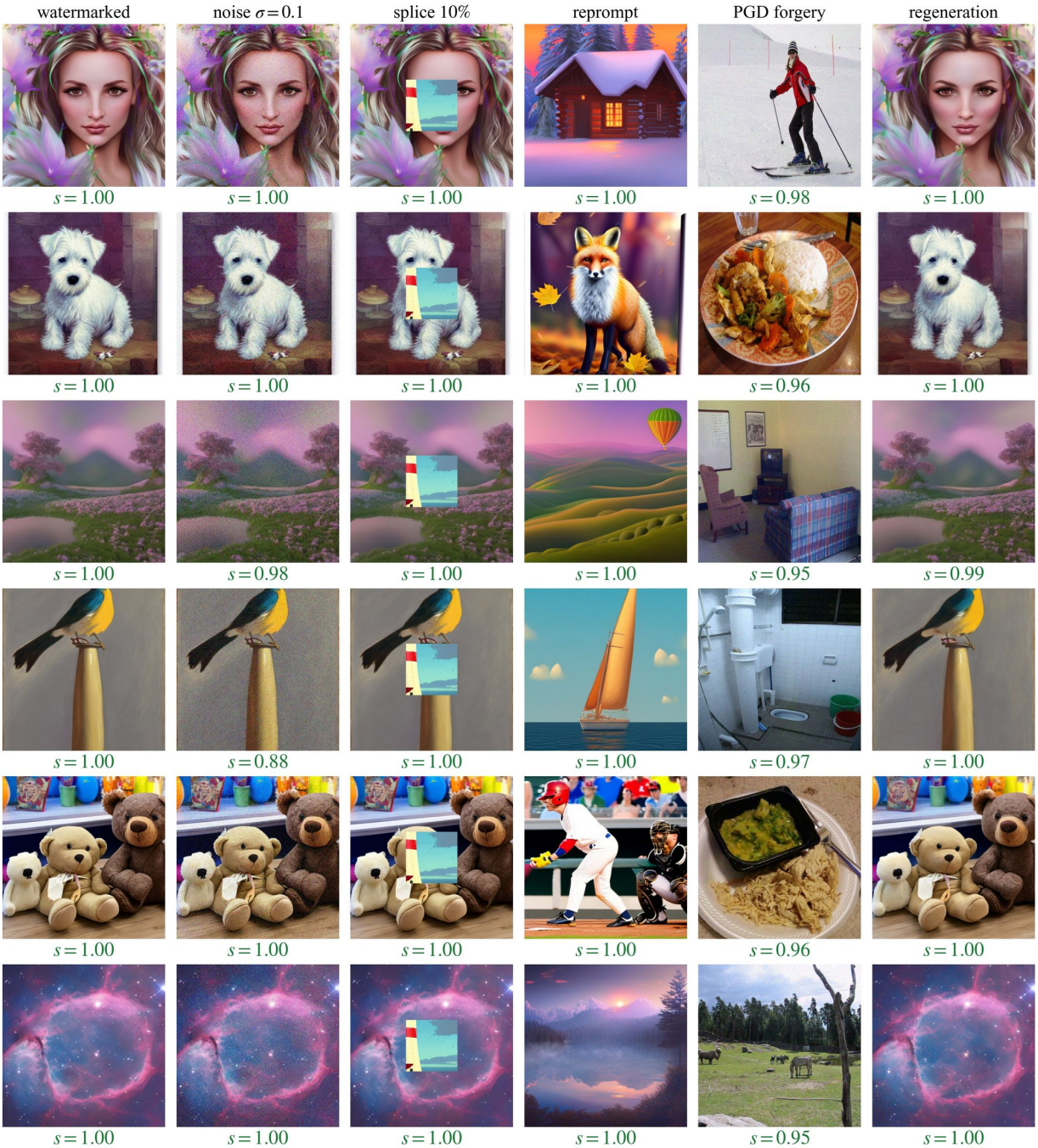

Figure 11: Gaussian Shading under the same operations, scored by its own detector, green above its threshold and red below it.

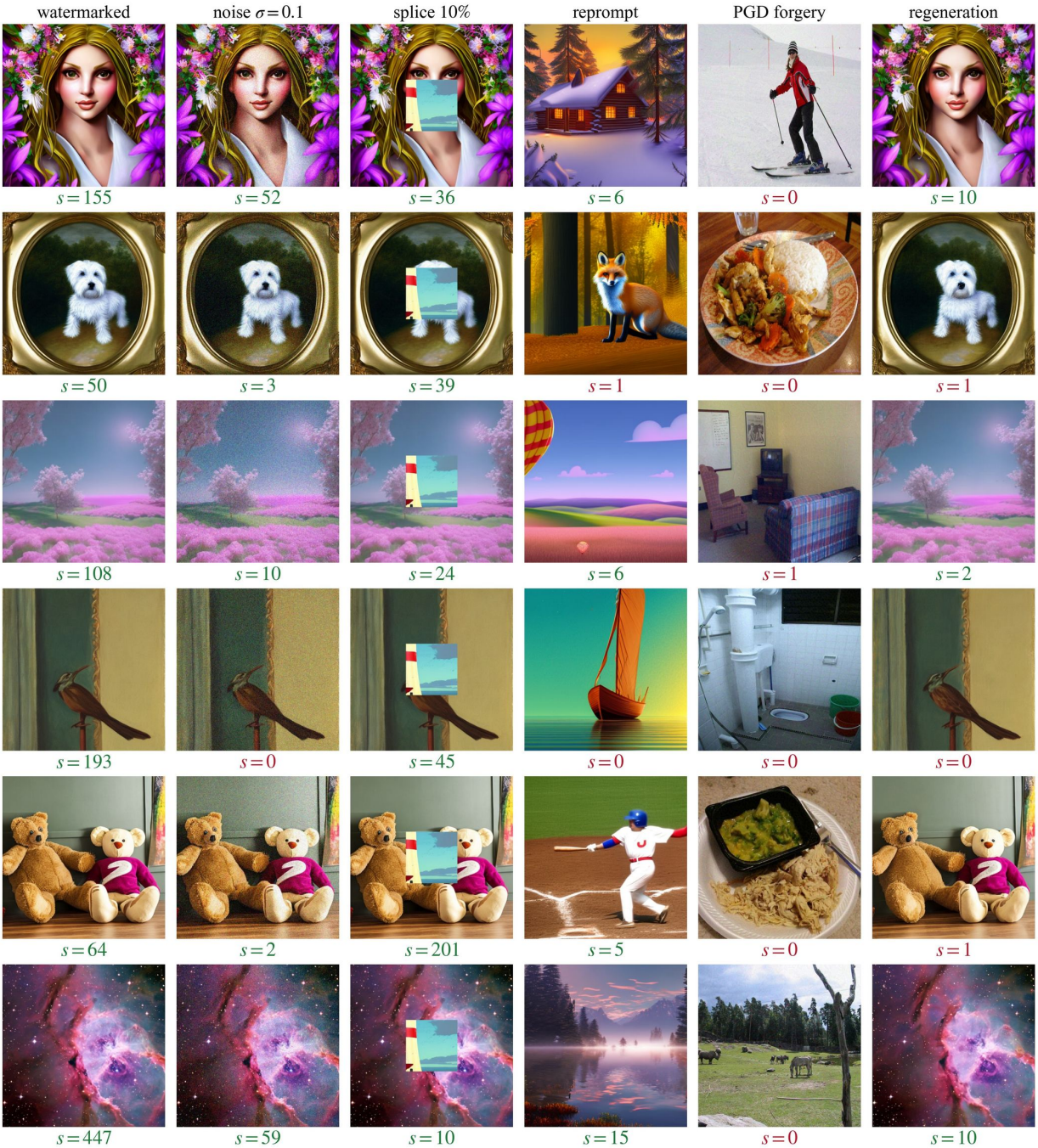

Figure 12: SEAL under the same operations, scored by its own detector, green above its threshold and red below it.