RESON: PERSISTENT WATERMARKING ACROSS HETEROGENEOUS GENERATIVE LINEAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated images are increasingly edited and re-synthesized by downstream generative models, yet existing watermarking evaluations largely focus on single-stage transformations rather than provenance across successive generations. We introduce RESON, a generation-time latent watermark for generative lineage persistence that embeds a keyed, message-bearing carrier once in the source noise and enables detection directly from pixels, without downstream reinsertion, generator modification, or diffusion inversion. We evaluate the method across heterogeneous multi-hop lineages, alternative lineage orderings, conventional perturbations, deliberate generative removal attacks, and varying payload capacities. The watermark remains detectable through repeated cross-model re-synthesis while retaining recoverable source-associated information at increasing lineage depths. Controlled carrier and strength ablations show that persistence is driven by structured carrier content rather than injection energy alone, while generation-quality analysis indicates limited degradation in semantic fidelity. Together, these findings establish generative lineage persistence as a distinct watermarking objective and demonstrate the feasibility of insert-once, detect-later provenance across heterogeneous generative transformations. The source code is available at `https://airesearchfolio.github.io/reson`.

1 INTRODUCTION

Diffusion models have become a dominant paradigm for photorealistic image synthesis, from latent diffusion and SDXL to newer high-resolution diffusion and rectified-flow systems Ho et al. (2020); Rombach et al. (2022); Dhariwal & Nichol (2021); Podell et al. (2024); Labs et al. (2025); Xie et al. (2024). Their realism and accessibility make reliable provenance increasingly important: once synthetic images are redistributed, edited, or used as inputs to later generative systems, the original source can become difficult to establish. Digital watermarking addresses this problem by embedding machine-detectable information into generated content, extending a long line of work on robust signal embedding and learned data hiding Cox et al. (1997); Chen & Wornell (2001); Zhu et al. (2018); Tancik et al. (2020); Wan et al. (2022).

Recent watermarking methods for diffusion models have made substantial progress on imperceptibility, payload recovery, and robustness to conventional distortions and deliberate removal attacks Wen et al. (2023); Ci et al. (2025); Yang et al. (2024); Arabi et al. (2025); Li et al. (2025); Kociszewski et al. (2026); Fernandez et al. (2023). Yet the prevailing robustness question remains largely one-step: can a watermark survive a transformation or attack applied to a marked image? Recent benchmarks confirm that generative reconstruction is substantially stronger than conventional corruptions An et al. (2024); Lu et al. (2025); Zhao et al. (2024); Liu et al. (2025); Hu et al. (2024), but a single reconstruction is not the same as a generative lineage. In realistic reuse, a synthetic image may be re-synthesized by one model, edited by another, and passed through still another generator, each replacing low-level evidence according to a different learned prior. A watermark that survives one reconstruction therefore need not preserve the identity or payload of the original source after several heterogeneous transitions.

This exposes a gap between robust watermark detection and persistent provenance. Existing work has studied generated/not-generated verification, multi-key or multi-user identification, proactive attribution, training-data traces, and model/data copyright signals Sablayrolles et al. (2020); Meintz

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.

et al. (2025); Asnani et al. (2024); Yu et al. (2021); Dubiński et al. (2025). These are complementary provenance objectives, but they do not directly address the setting studied here:

> If provenance information is inserted once at source generation, can the original payload remain recoverable after repeated re-synthesis by heterogeneous downstream generators, without reinsertion or downstream cooperation?

We formalize this setting as *persistent payload provenance*. A source generator produces G_0 with a keyed payload embedded once. The image then traverses a lineage

$$G_0 \to G_1 \to \cdots \to G_K,$$

where downstream generators are unmodified and watermark-unaware. There is no watermark refresh, no access to intermediate latent states, and no detector adaptation at later hops. A detector trained and calibrated at the source must recover the original source watermark directly from descendants. This formulation distinguishes persistence across successive generative transitions from robustness to a collection of independent perturbations.

We introduce RESON, a generation-time latent watermark for persistent provenance. A keyed, message-bearing carrier is injected once into the source latent under constant-energy normalization, and a frozen decoder separately recovers watermark presence and payload; downstream generators require no modification, key access, or reinsertion. We evaluate heterogeneous G_0–G_4 lineages, alternative generator orders, multi-bit payloads, conventional and generative attacks, carrier/strength ablations, and generation fidelity. On the canonical SD2.1 lineage, RESON retains 92.50% TPR at 1% FPR at G_4, while four-bit recovery remains 94.03% per bit. Controlled ablations further attribute persistence to carrier structure rather than injection energy alone.

Our contributions are:

- **A persistent-provenance problem formulation.** We define the insert-once, detect-later setting in which a source watermark must survive a heterogeneous generative lineage without downstream reinsertion, generator modification, intermediate-state access, or detector adaptation.
- **RESON, a source-side keyed payload watermark.** RESON embeds constant-energy, message-bearing keyed carriers at G_0 and uses a frozen decoder that separates calibrated presence detection from payload recovery, enabling source provenance to be queried from later descendants while leaving downstream generators unchanged.
- **A lineage-centered evaluation protocol.** We evaluate heterogeneous G_0–G_4 chains and lineage orders alongside payload-capacity, perturbation, deliberate-removal, carrier/strength, and fidelity experiments, separating multi-hop persistence from conventional one-step robustness while testing the mechanism and quality trade-offs underlying the approach.

2 Related Work

Diffusion watermarking. Diffusion watermarking spans post-processing, model-modifying, and generation-time approaches. Learned image watermarks such as HiDDeN and StegaStamp embed signals directly in image space Zhu et al. (2018); Tancik et al. (2020), while Stable Signature and related methods modify a diffusion decoder or model so that generated images carry a recoverable signature Fernandez et al. (2023); Zhao et al. (2023); Cui et al. (2025b;a); Feng et al. (2024); Wang et al. (2025). Tuning-free methods instead exploit the initial diffusion noise. Tree-Ring embeds a Fourier pattern and detects it through diffusion inversion Wen et al. (2023); RingID improves multi-key identification Ci et al. (2025); Gaussian Shading encodes high-capacity messages while preserving the Gaussian latent distribution Yang et al. (2024); WIND uses pseudorandom initial noises with two-stage retrieval to improve resistance to removal and forgery Arabi et al. (2025); and GaussMarker combines spatial- and frequency-domain marking with a learned Gaussian-noise restorer Li et al. (2025). More recently, SERUM adds a watermark vector to the initial noise but replaces inversion-based verification with a lightweight learned detector, enabling fast detection and multi-user marking Kociszewski et al. (2026). RESON is closest to this generation-time family. Its distinguishing objective is persistent provenance: a keyed payload is inserted only at source generation and must remain recoverable after successive heterogeneous re-generations, without downstream reinsertion, cooperation, or detector adaptation.

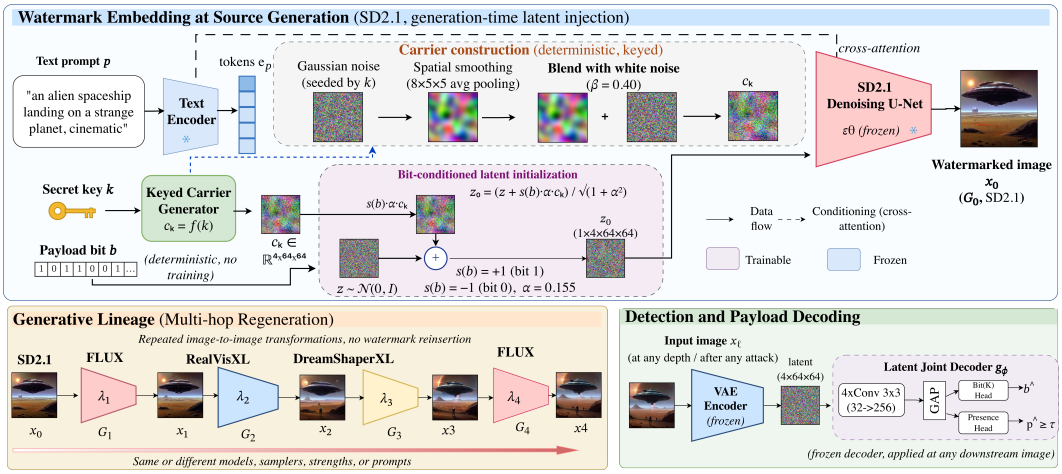

Figure 1: **Overview of RESON.** A keyed payload is embedded once through source latent initialization, propagated through multi-hop generative transformations without reinsertion, and recovered by a frozen latent decoder at any lineage depth.

Robustness, removal, and generative reconstruction. Conventional watermark evaluation considers transformations such as compression, crop, blur, noise, and geometric edits. WAVES and related benchmarks broaden this evaluation and show that robustness can vary substantially across attack families An et al. (2024); Lu et al. (2025). Of particular relevance to RESON, generative reconstruction can remove or weaken invisible watermarks while maintaining perceptual content Zhao et al. (2024); Liu et al. (2025); Hu et al. (2024); watermark security also includes forgery and tamper-aware settings Müller et al. (2025); Chen et al. (2025); Zhang et al. (2024). These studies motivate treating a generator itself as a destructive watermark channel. Prior work also considers repeated regeneration in selected settings, but our focus is heterogeneous generative lineage: provenance is inserted only at G_0 and evaluated after successive transitions across different generator families, without reinsertion or detector adaptation. Thus, multi-hop lineage persistence is complementary to—rather than a replacement for—standard perturbation and deliberate-removal robustness.

Provenance, attribution, and persistent signals. Prior work has studied several forms of provenance beyond binary detection. Multi-key and multi-user watermarks support attribution Ci et al. (2025); Kociszewski et al. (2026); proactive fingerprints and marks associate content with a source or identity Yu et al. (2021); Asnani et al. (2024); Fernandez et al. (2022); and radioactive-data methods test whether marked training examples leave detectable traces in subsequently trained models Sablayrolles et al. (2020); Meintz et al. (2025). Other work studies ownership signals that survive training or fine-tuning Cui et al. (2025b;a); Dubiński et al. (2025); Wang et al. (2025). RESON studies a distinct persistence mechanism: *sample-to-sample generative inheritance*. Downstream models are not trained on the marked data, need not know the key, and are never asked to re-embed the mark. We therefore separate calibrated watermark presence from payload recovery and evaluate both lineage depth and lineage order. Multi-bit payload experiments further test whether the surviving signal carries usable provenance information rather than merely a binary watermark trace Chang et al. (2026); Zhang et al. (2026). Image fidelity is evaluated separately with standard distributional and semantic measures Heusel et al. (2017); Radford et al. (2021); Hessel et al. (2021).

3 Method

Problem setup and generative lineage. Let G_0 denote the source generator and $G_1, \ldots, G_L$ downstream image-to-image generators. RESON modifies only the initial latent used by G_0; subsequent images are generated from their immediate predecessors without access to the original watermarking operation. If x_0 denotes the source image, its descendants satisfy

$$x_\ell = G_\ell(x_{\ell-1}; \xi_\ell), \qquad \ell = 1, \ldots, L, \tag{1}$$

where ξ_ℓ represents the generation configuration at hop ℓ, such as the model, sampler, regeneration strength, or prompt. As illustrated in the lower-left portion of Fig. 1, the resulting lineage follows

$$x_0 \to x_1 \to \cdots \to x_L, \tag{2}$$

and may therefore traverse heterogeneous generators. Crucially, the watermark is inserted only once during source generation and is never reinserted into its descendants.

Keyed carrier construction. The upper portion of Fig. 1 illustrates the complete source-generation procedure. A secret key k deterministically defines the watermark carrier. Specifically, the key seeds a Gaussian noise field, from which a spatially structured component c_s is obtained by smoothing. This structured component is combined with an independently generated white component c_w using the carrier blend ratio β. The resulting carrier is normalized as

$$c_k = \mathrm{Renorm}(\beta c_s + (1-\beta)c_w)\,. \tag{3}$$

Here, $\mathrm{Renorm}(x) = \|c_s\|_2\, x/\|x\|_2$, so the blended carrier preserves the ℓ_2 norm of the standardized smoothed carrier. The canonical configuration uses $\beta = 0.40$. Because carrier construction is deterministic with respect to the secret key, the same carrier can be reconstructed when required without storing an image-specific perturbation.

Payload-conditioned latent initialization. The payload determines the sign with which the keyed carrier is introduced into the source latent. For a one-bit payload $b \in \{0, 1\}$, we define

$$s(b) = 2b - 1, \tag{4}$$

so that $s(b) = +1$ for bit 1 and $s(b) = -1$ for bit 0. Given the standard source noise

$$z \sim \mathcal{N}(0, I), \tag{5}$$

RESON constructs the watermarked source latent z_0 as

$$z_0 = \frac{z + \alpha s(b) c_k}{\sqrt{1+\alpha^2}}\,, \tag{6}$$

where α controls watermark strength. The canonical setting uses $\alpha = 0.155$. The normalization in Eq. 6 keeps the injected latent at a controlled energy scale rather than simply increasing its magnitude as watermark strength is added. As shown in Fig. 1, z_0 replaces the ordinary source initialization before denoising. The text prompt is encoded normally and supplies the usual conditioning to the frozen SD2.1 denoising U-Net. Thus, RESON does not repeatedly modify intermediate U-Net layers or reinsert the carrier during denoising; the watermark enters through the generation-time latent initialization. The denoising process then produces the watermarked source image x_0. For a K-bit payload, independently keyed, low-correlation carriers are signed according to their corresponding payload bits, summed, and renormalized to the norm of a single carrier before applying the same injection rule in Eq. 6. Consequently, increasing the payload length does not intentionally increase the total watermark energy.

Persistence through generative lineage. After x_0 has been generated, no additional watermarking operation is performed. Each subsequent image is obtained only by applying the next image-to-image transformation according to Eq. 1. The lower-left portion of Fig. 1 illustrates this process across multiple generations, where different downstream models may be used at different lineage depths. More generally, the models, samplers, regeneration strengths, and prompts may vary between hops. This setting deliberately tests whether information introduced only at the source latent survives repeated generative transformations. In particular, a descendant x_ℓ does not receive the original carrier c_k or watermarked latent z_0 directly; any recoverable watermark evidence at depth ℓ must therefore have persisted through the preceding sequence of generative transformations.

Detection and payload decoding. The lower-right portion of Fig. 1 shows the decoding pipeline. At any lineage depth ℓ, or after an evaluated attack, an input image x_ℓ is first mapped by the frozen source VAE encoder to a latent representation. A compact latent decoder then processes this representation and produces two outputs: a presence score $\hat{p}$ and payload prediction $\hat{b}$ (or K payload predictions in the multi-bit setting). The presence head determines whether watermark evidence exceeds the calibrated detection threshold, whereas the payload head recovers the embedded message.

Presence is trained using both clean and watermarked source-generation examples, while the payload objective is evaluated on watermarked examples. The decoder is trained and calibrated using source-generation data and is then frozen for all downstream lineage and robustness evaluations. The same frozen decoder is therefore applied to $x_0, x_1, \ldots, x_L$ without adaptation to a particular lineage depth or downstream generator. We report ROC-AUC and true-positive rate at 1% false-positive rate (TPR@1%FPR) for watermark presence, together with mean bit accuracy and exact-message recovery for payload decoding. Presence detection and payload recovery are reported separately because they measure different aspects of watermark persistence: repeated regeneration may reduce the confidence of a conservative presence detector while the encoded payload remains recoverable.

4 Experiments and Results

Datasets and evaluation protocol. We construct a canonical benchmark of 10,000 paired clean/watermarked SD2.1 generations at 512×512 resolution using 10,000 unique prompts from Stable-Diffusion-Prompts. Each pair shares its prompt and generation seed, with RESON applied only to the watermarked branch. The benchmark is partitioned into 8,000 training, 1,000 validation, and 1,000 held-out test pairs, with balanced one-bit payloads. The decoder is trained on the training split; validation data determine checkpoint selection and calibrate the operating threshold to target 1% FPR. The selected checkpoint and threshold are then frozen for all held-out evaluations, with no training or adaptation on downstream generators or evaluated attacks. Accordingly, unless explicitly stated otherwise, reported TPR values use this validation-calibrated operating point; the realized FPR on a held-out or shifted distribution may differ from exactly 1%.

The canonical multi-hop protocol propagates each held-out source image sequentially through SD2.1 (G0) $\rightarrow$FLUX.1-dev (G1) $\rightarrow$RealVisXL V5.0 (G2) $\rightarrow$DreamShaperXL (G3) $\rightarrow$FLUX.1-dev (G4), with each generator operating on its immediate predecessor, using the source prompt, and without watermark reinsertion. Lineage-order experiments additionally include Juggernaut XL. Matched baseline comparisons, controlled ablations, and specialized robustness experiments report their evaluation set sizes with the corresponding results. Generation quality is evaluated independently on 10,000 COCO-caption generations (Lin et al., 2014), using FID (Heusel et al., 2017) against real COCO images and CLIP similarity (Radford et al., 2021) for prompt–image alignment.

Baselines and metrics. We compare RESON with Tree-Ring (Wen et al., 2023), Gaussian Shading (Yang et al., 2024), WIND (Arabi et al., 2025), and SERUM (Kociszewski et al., 2026), using each method's native detector. Watermark presence is evaluated by ROC-AUC and TPR at the validation-calibrated 1%-FPR operating point, while payload recovery is measured by mean bit accuracy and exact-message accuracy.

Computational cost. Embedding requires only deterministic carrier construction and element-wise latent injection. On an A100-SXM4-80GB, detection takes 29.2 ms/image (batch size 1); 30-epoch decoder training on 8,000 pairs requires approximately 2.7 GPU-hours.

4.1 Multi-hop persistence and matched baselines

Table 1 is the primary test of the paper's central claim. With SD2.1, RESON maintains 100.00% TPR at G1, 98.30% at G2, 93.10% at G3, and 92.50% at G4. At later depths, RESON achieves the highest observed TPR among the evaluated methods under the matched heterogeneous-lineage protocol. The same trend is observed with SD2.0: RESON reaches 94.00% TPR at G4 versus 86.00% for SERUM. ROC-AUC remains high at later depths for both source models, showing that the result is not confined to one source checkpoint. Figure 2 illustrates representative lineages. For each prompt, the clean and RESON-watermarked G0 images are generated with the same prompt and seed, so the only difference between them is the watermark carrier added to the initial latent. Because the diffusion sampler is sensitive to its initial latent, this small perturbation yields a different but semantically faithful realization of the prompt rather than a pixel-level edit of the clean image, which explains the changes in pose and composition already visible at G0. Each branch is then regenerated independently through the canonical lineage without watermark reinsertion. Every generator substantially re-renders the image, altering pose, lighting, and fine detail, yet the watermarked images remain faithful

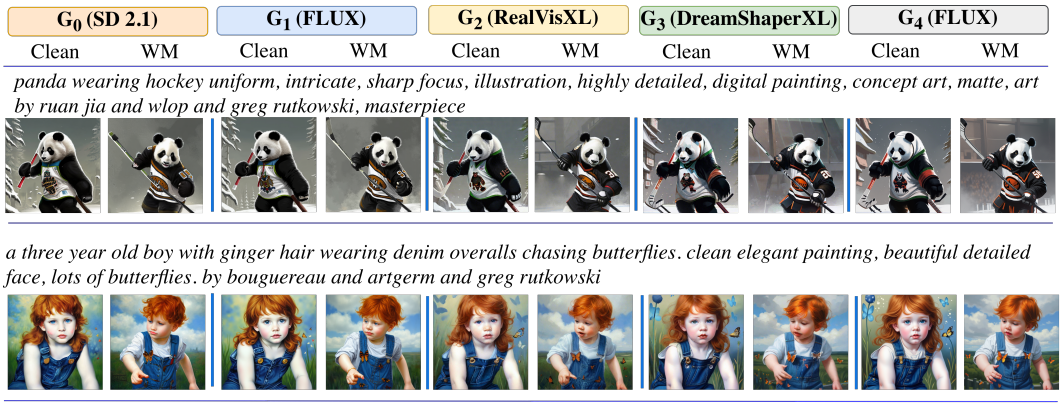

Figure 2: **Qualitative G0–G4 lineages.** Clean and RESON-watermarked images for two prompts at each depth of the canonical lineage (SD2.1→FLUX→RealVisXL→DreamShaperXL→FLUX). No watermark is reinserted after G0.

Table 1: Matched watermark robustness under the canonical heterogeneous lineage for different base diffusion models. All methods are evaluated on the same prompt subset and G0–G4 regeneration protocol, while source generations follow each method's native watermarking procedure. G0 denotes source generation; G1–G4 denote sequential regeneration through FLUX, RealVisXL, DreamShaperXL, and FLUX, respectively. TPR is reported using the validation-calibrated operating threshold (target 1% FPR), which is frozen before held-out lineage evaluation. Higher is better; the best result at each depth within each base-model block is shown in bold.

		G0		G1		G2		G3		G4	
Base	Method	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
SD V2.1	Gaussian Shading	**100.00**	**1.0000**	99.67	0.9996	34.00	0.9036	6.33	0.7843	12.33	0.7815
	WIND	**100.00**	**1.0000**	36.88	0.7784	1.18	0.5436	0.65	0.5293	2.45	0.4806
	SERUM	**100.00**	**1.0000**	**100.00**	**0.9999**	97.67	0.9968	92.00	0.9941	89.33	0.9916
	Tree-Ring	**100.00**	**1.0000**	71.67	0.9584	26.67	0.8669	10.83	0.7500	14.17	0.7372
	RESON (ours)	**100.00**	**1.0000**	**100.00**	0.9996	**98.30**	**0.9980**	**93.10**	**0.9967**	**92.50**	**0.9963**
SD V2.0	Gaussian Shading	**100.00**	**1.0000**	99.33	0.9994	35.00	0.9075	6.67	0.7891	12.67	0.7868
	WIND	**100.00**	**1.0000**	37.33	0.7816	1.33	0.5479	0.67	0.5326	2.33	0.4895
	SERUM	**100.00**	**1.0000**	**100.00**	0.9969	93.53	0.9934	91.82	0.9722	86.00	0.9799
	Tree-Ring	**100.00**	**1.0000**	72.00	0.9590	27.33	0.8708	11.33	0.7569	14.67	0.7426
	RESON (ours)	**100.00**	**1.0000**	98.33	**0.9988**	**94.33**	**0.9951**	**93.67**	**0.9830**	**94.00**	**0.9824**

to the prompt at every depth. A qualitative comparison with baseline watermarks is provided in Appendix A.2.1.

4.2 Lineage-order robustness

The canonical lineage should not depend on a favorable ordering of downstream generators. We therefore evaluate five heterogeneous regeneration sequences, all originating from the same SD2.1 watermarked source and without watermark reinsertion. As shown in Table 2, persistence is consistent across generator orderings: G4 TPR at the validation-calibrated operating point remains between 87.3% and 92.5%, while final payload recovery remains 99.8–100.0%. This indicates that lineage persistence is not specific to a single downstream regeneration path.

4.3 Robustness to perturbations and deliberate removal

We consider watermark-unaware downstream transformations that may modify or regenerate marked content but do not access the secret carrier key or adapt against the RESON detector. Adaptive white-box adversaries with access to either are outside the present threat model.

Table 2: **Robustness to lineage order.** Watermark persistence across heterogeneous regeneration sequences originating from the same SD2.1 watermarked source. ROC-AUC and TPR at the validation-calibrated operating point are reported at each depth; Bit@G4 denotes final payload accuracy and CLIP_{avg} denotes average semantic similarity over G_1–G_4. No watermark is reinserted after G_0. All experiments use $N = 1{,}000$ held-out pairs. M2 = FLUX, M3 = RealVisXL, M4 = DreamShaperXL, and M5 = Juggernaut.

	G_1		G_2		G_3		G_4		**Bit@G4**	**CLIP_{avg}**
Lineage	**AUC**	**TPR**	**AUC**	**TPR**	**AUC**	**TPR**	**AUC**	**TPR**	**(%)**	**↑**
M2→M3→M4→M2	.9996	100.00	.9980	98.30	.9967	93.10	.9963	92.50	99.9	0.3309
M4→M3→M2→M5	.9825	91.0	.9711	85.7	.9622	80.2	.9410	87.3	99.8	0.3229
M3→M2→M4→M2	.9928	93.3	.9835	92.3	.9822	88.0	.9688	90.7	100.0	0.3261
M4→M2→M3→M2	.9920	95.3	.9865	92.0	.9859	91.3	.9760	92.3	100.0	0.3297
M2→M4→M2→M3	.9935	97.7	.9949	95.7	.9876	89.0	.9795	89.3	100.0	0.3294

Conventional perturbations. Table 3 summarizes conventional-perturbation robustness across RESON and representative baselines. Complete RESON severity sweeps are provided in Appendix A.3.1.

Table 3: **Conventional-perturbation robustness.** Results summarize robustness under standard image perturbations. Values are TPR (%); higher is better.

Method	JPEG	Blur	Gaussian Noise	Brightness	Rotation	C&S
Gaussian Shading	99.99	**100.00**	92.30	99.80	5.20	0.81
$\text{WIND}_{\text{full}}$-128	**100.00**	**100.00**	95.00	78.40	**88.00**	93.00
SERUM	**100.00**	**100.00**	95.30	99.72	59.80	**99.50**
Tree-Ring	79.00	61.00	53.00	42.00	3.00	4.00
RESON (ours, SD2.1)	**100.00**	**100.00**	96.49	**99.88**	70.30	99.40
RESON (ours, SD2.0)	**100.00**	99.67	**96.67**	99.00	72.47	98.33

Deliberate generative removal. Generative reconstruction is a stronger threat than conventional corruption, because it re-synthesizes image content through a learned prior. Table 4 evaluates VAE and diffusion regeneration, including repeated RINSE-style reconstruction, using the validation-frozen threshold. Realized FPR stays at or below 0.9% under every attack, so calibration transfers to the attacked distribution, and frozen TPR ranges from 96.4% to 98.6%. Payload recovery is essentially unaffected (bit accuracy $\geq$99.9%). Samples missed by the presence detector still carry recoverable payload information, with 100% bit accuracy on misses for four of five attacks. Thus, failure of the calibrated presence decision does not necessarily imply loss of the encoded payload. Oracle-threshold diagnostics are reported in Appendix A.3.2.

Table 4: **Deliberate generative removal** using WatermarkAttacker (Zhao et al., 2024); configurations are detailed in Appendix A.3.2.

		Presence		Payload		
Attack	AUROC ↑	TPR (%) ↑	FPR (%) ↓	Bit acc. (%) ↑	Misses	On misses (%) ↑
Regen-VAE q4	0.9943	96.4	0.8	100.0	36	100.0
Regen-Diff t60	0.9997	98.6	0.6	100.0	14	100.0
Regen-Diff t100	0.9996	98.4	0.9	100.0	16	100.0
RINSE-2× t60	0.9985	98.3	0.6	99.9	17	94.1
RINSE-4× t60	0.9948	96.4	0.3	100.0	36	100.0

4.4 Carrier Design and Operating-Point Analysis

Our watermark combines injection strength α with carrier mixing β. We first select the canonical operating point from the detection–fidelity trade-off, then use controlled ablations to isolate the contribution of carrier structure to multi-hop persistence.

Detection–fidelity trade-off. Figure 3 summarizes the sequential parameter selection. We first vary α while holding $\beta = 0.34$ fixed (Fig. 3a). Increasing α from 0.135 to 0.155 raises TPR@1% FPR from 85.0% to 94.0%, while the corresponding prompt-fidelity loss increases more gradually. We therefore select $\alpha = 0.155$ for the subsequent carrier-mixing sweep. With $\alpha = 0.155$ fixed, increasing β progressively strengthens detection (Fig. 3b), reaching 100% TPR@1% FPR at $\beta = 0.40$, while simultaneously increasing $-\Delta\text{CLIP}$. This exposes the expected robustness–fidelity trade-off: stronger carrier mixing improves watermark detectability but incurs greater prompt-fidelity cost. Based on this sequential sweep, we use

$$(\alpha, \beta) = (0.155,\ 0.40)$$

as the canonical configuration in the main experiments.

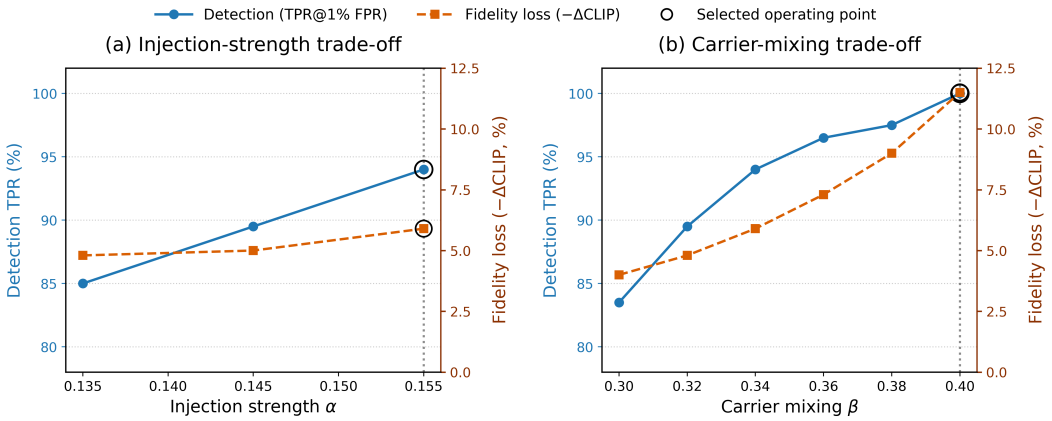

Figure 3: **Detection–fidelity trade-off and sequential operating-point selection.** (a) At fixed $\beta = 0.34$, increasing injection strength α improves TPR@1% FPR while incurring a moderate increase in prompt-fidelity loss, motivating $\alpha = 0.155$. (b) Fixing $\alpha = 0.155$ and increasing carrier mixing β further improves detection, reaching 100% TPR at $\beta = 0.40$, at the cost of increased $-\Delta\text{CLIP}$. Circled markers denote the value selected at each tuning stage; the resulting canonical configuration is $(\alpha, \beta) = (0.155, 0.40)$. Higher TPR and lower $-\Delta\text{CLIP}$ are preferred.

Controlled carrier-mechanism ablation. To isolate the role of carrier structure, we compare matched $\beta \in \{0, 0.4, 1\}$ carriers on the same $N = 300$ prompts, with a separately trained detector and validation-frozen threshold for each condition. Table 5 shows that the white carrier ($\beta = 0$) collapses from 19.0% TPR at G0 to 1.7% at G4, whereas the canonical blend ($\beta = 0.4$) retains 64.3% and the fully structured carrier ($\beta = 1$) retains 81.7% at G4. G4 bit accuracy similarly increases from 64.7% to 98.3% and 99.7%, respectively, providing strong evidence that carrier structure is a major contributor to multi-hop persistence. We retain $\beta = 0.40$ as the canonical detection–fidelity operating point rather than optimizing persistence alone. Full depth-wise results are provided in Appendix A.5.1.

4.5 Payload Capacity and Persistent Information

RESON preserves multi-bit information across G0–G4 despite a capacity–robustness trade-off. At G4, the 4-bit configuration retains 94.03% mean bit accuracy and 78.4% exact-message recovery, showing that persistence extends beyond binary presence. Larger-payload and payload-specific quality results are provided in Appendix A.4.

Table 5: **Controlled carrier-mechanism ablation.** This $N = 300$ controlled ablation uses separately trained matched detectors and is intended for within-table comparison of carrier structure; its absolute TPR values should therefore not be directly compared with the canonical Table 1 experiment.

Carrier	β	G0 AUC	G0 TPR	G4 TPR	Bit@G4
White	0.00	.7914	19.0	1.7	64.7
RESON blend	0.40	.9940	93.3	64.3	98.3
Structured	1.00	.9997	98.3	**81.7**	**99.7**

4.6 Generation Quality

On the full 10K canonical paired benchmark, mean prompt–image CLIP similarity changes from 0.2750 for clean generations to 0.2699 for RESON. Across matched pairs, the mean relative change is -1.27%, and RESON scores higher than its matched clean generation for 42.96% of prompts. Separate COCO-reference FID/CLIP evaluation across payload lengths and additional paired fidelity analyses are reported in Appendix A.7 and Appendix A.4.3, respectively.

5 Discussion and Limitations

Detection and information persistence. RESON separates calibrated watermark presence from message recovery. Repeated regeneration can reduce the presence score below its operating threshold while payload bits remain decodable on known-watermarked samples, as shown by the multi-bit results in Sec. 4.5. This demonstrates residual encoded information beyond the binary presence decision; because payload recovery is evaluated conditionally on known-watermarked samples, however, it should not be interpreted as an independent FPR-controlled provenance detector.

Why does structure persist? The carrier ablation indicates that persistence arises primarily from the structured component rather than injection energy alone: white carriers rapidly collapse, whereas increasingly structured carriers retain substantially more signal across regeneration. One plausible explanation is that spatially smooth structure overlaps with information preserved during re-synthesis, while unstructured noise is discarded. Establishing this mechanism causally remains future work.

Limitations. Our evaluation covers SD2.1/SD2.0 sources, four downstream generators, lineages through G4, and 512×512 images; persistence under unseen future generators, higher resolutions, and longer lineages therefore remains unestablished. The source-trained detector is never adapted downstream; we do not evaluate adaptive key-aware attacks or multi-sample carrier-estimation and averaging attacks against the fixed keyed carrier. Multi-user attribution across generative lineage depth is not evaluated, leaving scalable identity-level provenance under repeated re-synthesis unresolved. Future work will extend RESON to broader generator families, resolutions, and longer lineages, while evaluating adaptive attacks and scalable multi-user attribution under repeated re-synthesis.

6 Conclusion

We introduce generative lineage persistence and RESON, a source-side latent watermark designed to preserve keyed provenance across heterogeneous generative re-synthesis without downstream reinsertion or cooperation. Experiments across heterogeneous and reordered lineages, conventional perturbations, and generative reconstruction demonstrate that source-associated information can survive substantial re-synthesis. Multi-bit results further show that payload information can remain recoverable even as conservative presence detection weakens, while carrier ablations show substantially stronger persistence for structured carriers under matched watermark energy. These findings distinguish watermark presence from provenance-information persistence and motivate evaluating both across evolving generative lineages rather than only isolated transformations. Overall, RESON demonstrates the feasibility of insert-once, detect-later provenance and establishes generative lineage persistence as a complementary axis for evaluating generative-media provenance.

AI USE STATEMENT

Generative AI tools (Claude, Anthropic) were used to assist with manuscript drafting, language editing, LaTeX formatting, figure layout, and critical feedback on the presentation of methodology and experiments. All experimental design decisions, carrier and detector implementations, hyperparameter selections, reported numerical results, mathematical formulations, figure content, citations, and scientific claims were produced and verified by the authors. No generative AI tools were used to produce experimental results, generate training or evaluation data, or design the RESON watermarking method. The authors take full responsibility for the content of this submission.

ETHICS STATEMENT

This work studies defensive provenance marking of AI-generated images. No new human-subject data were collected; all experiments use prompt-conditioned synthetic images generated by the authors. The prompts are drawn from an existing public collection (Stable-Diffusion-Prompts). The deliberate-removal and regeneration-attack evaluations are included to characterize watermark failure modes under realistic adversarial conditions, not to facilitate evasion of deployed provenance systems. We report these failure modes and detection limits to support more reliable evaluation of generative watermarking. The watermarking method itself is designed to support responsible provenance and does not enable surveillance, content filtering, or restriction of legitimate use.

REPRODUCIBILITY STATEMENT

The canonical dataset construction, source model, downstream generators, and evaluation protocol are specified in Section 4 and Appendix A. The detector architecture (`LatentJointDecoder`), training objectives, optimization settings, and checkpoint-selection procedure are detailed in Appendix A.1; the operating threshold is calibrated on validation data to target $\leq 1\%$ FPR. Carrier construction and constant-energy normalization are defined in Equations 3–6, while the canonical hyperparameters ($\alpha = 0.155$, $\beta = 0.40$) are reported in Section 4. Perturbation severities, attack configurations, and lineage generator checkpoints are documented on the project page. Generation quality metrics and their measurement protocols are specified in Appendix A.7. Evaluation-set sizes are reported with the corresponding experiments; the canonical lineage evaluation uses $N = 1{,}000$ held-out test pairs, while controlled ablations use the sample sizes stated in their respective sections. Anonymized code, evaluation scripts, and full reproducibility details are available at `https://airesearchfolio.github.io/reson`. The project page also provides additional qualitative examples, including matched clean/watermarked generation pairs across diverse prompts and extended G0–G4 lineage trajectories beyond those shown in the paper.

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, 2024. URL `https://arxiv.org/abs/2401.08573`.

Kasra Arabi, Benjamin Feuer, R. Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 61271–61304, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/file/999fcab97007ebef0cda9949550b4a9e-Paper-Conference.pdf`.

Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. Promark: Proactive diffusion watermarking for causal attribution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10802–10811, 2024. doi: 10.1109/CVPR52733.2024.01027.

Xuanhang Chang, Zhonghao Yang, Cheng Zhuo, and Yu Li. Maxmark: High-capacity diffusion-native watermarking via robust and invertible latent embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9394–9403, June 2026.

B. Chen and G.W. Wornell. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47 (4):1423–1443, 2001. doi: 10.1109/18.923725.

Yuzhuo Chen, Zehua Ma, Han Fang, Weiming Zhang, and Nenghai Yu. Tag-wm: Tamper-aware generative image watermarking via diffusion inversion sensitivity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16723–16732, 2025. doi: 10.1109/ICCV51701.2025.01553.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 338–354, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73390-1.

I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997. doi: 10.1109/83.650120.

Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Lingjuan Lyu, Wenqi Fan, Hui Liu, and Jiliang Tang. Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *SIGKDD Explor. Newsl.*, 26(2):76–88, January 2025a. ISSN 1931-0145. doi: 10.1145/3715073.3715080. URL https://doi.org/10.1145/3715073.3715080.

Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for data copyright protection against generative diffusion models. *SIGKDD Explor. Newsl.*, 26(2):60–75, January 2025b. ISSN 1931-0145. doi: 10.1145/3715073.3715079. URL https://doi.org/10.1145/3715073.3715079.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. Cdi: Copyrighted data identification in diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18674–18684, 2025. doi: 10.1109/CVPR52734.2025.01740.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora, 2024. URL https://arxiv.org/abs/2405.11135.

Pierre Fernandez, Matthijs Douze, Hervé Jégou, and Teddy Furon. Active image indexing, 2022. URL https://arxiv.org/abs/2210.10620.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22409–22420, 2023. doi: 10.1109/ICCV51070.2023.02053.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, and Neil Gong. Stable signature is unstable: Removing image watermark from diffusion models, 2024. URL https://arxiv.org/abs/2405.07145.

Jan Kociszewski, Hubert Jastrzębski, Tymoteusz Stępkowski, Filip Manijak, Krzysztof Rojek, Franziska Boenisch, and Adam Dziedzic. Serum: Simple, efficient, robust, and unifying marking for diffusion-based image generation, 2026. URL https://arxiv.org/abs/2603.13396.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. Gaussmarker: Robust dual-domain watermark for diffusion models, 2025. URL https://arxiv.org/abs/2506.11444.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87310–87327, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9750da8aec3b79cf14dd29e7ab6605a-Paper-Conference.pdf.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83902–83936, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d077bc9ea82a2998ca6b2d0158b5ac6e-Paper-Conference.pdf.

Michel Meintz, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Radioactive watermarks in diffusion and autoregressive image generative models, 2025. URL https://arxiv.org/abs/2506.23731.

Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20937–20946, 2025. doi: 10.1109/CVPR52734.2025.01950.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Herve Jegou. Radioactive data: tracing through training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8326–8335. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sablayrolles20a.html.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2114–2123, 2020. doi: 10.1109/CVPR42600.2020.00219.

Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. A comprehensive survey on robust image watermarking. *Neurocomputing*, 488:226–247, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.02.083. URL https://www.sciencedirect.com/science/article/pii/S0925231222002533.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8213–8224, 2025. doi: 10.1109/CVPR52734.2025.00769.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023. URL https://arxiv.org/abs/2305.20030.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL https://arxiv.org/abs/2410.10629.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models, 2024. URL https://arxiv.org/abs/2404.04956.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14428–14437, 2021. doi: 10.1109/ICCV48922.2021.01418.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11964–11974, 2024. doi: 10.1109/CVPR52733.2024.01137.

Yi Zhang, Hongbo Huang, and Liang-Jie Zhang. Gaussian shannon: High-precision diffusion model watermarking based on communication, 2026. URL https://arxiv.org/abs/2603.26167.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8643–8672. Curran Associates, Inc., 2024. doi: 10.52202/079017-0276. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10272bfd0371ef960ec557ed6c866058-Paper-Conference.pdf.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models, 2023. URL https://arxiv.org/abs/2303.10137.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 682–697, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01267-0.

A APPENDIX

This appendix provides supplementary implementation details, extended robustness evaluations, payload-capacity experiments, ablations, and qualitative analyses supporting the main paper. Unless otherwise stated, experiments use the canonical Stable Diffusion 2.1 (SD2.1) RESON configuration and the evaluation protocol described in the main text. Heterogeneous downstream generators are introduced specifically to evaluate watermark persistence under generative lineage transformations; they do not change the canonical source-model setting.

A.1 IMPLEMENTATION AND TRAINING DETAILS

Detector architecture. RESON performs detection and payload decoding in VAE latent space. Images are encoded into $4 \times 64 \times 64$ latents using the frozen SD2.1 VAE encoder, consistent with the source-model latent space used for watermark insertion. The `LatentJointDecoder` comprises four 3×3 convolutional blocks with 32, 64, 128, and 256 channels, respectively, each followed by batch normalization, ReLU, and 2×2 max pooling.

Training. The detector is optimized with AdamW using a learning rate of 5×10^{-5}, weight decay 10^{-5}, and batch size 32 for at most 30 epochs. Presence detection uses a margin loss with margin 1.5, while payload recovery uses binary cross-entropy evaluated only on watermarked samples. The joint objective weights the presence and payload losses by 1.0 and 0.5, respectively. Training uses only source-generation data; no downstream regeneration examples are used.

Model selection. After each epoch, performance is evaluated on the validation set. Checkpoint selection prioritizes worst-condition validation AUC, followed by worst-condition bit accuracy and mean validation AUC, with early stopping after eight epochs without improvement. The selected detector is then frozen for held-out lineage evaluation.

Canonical lineage settings. All G1–G4 transitions operate at 512×512 resolution with an image-to-image strength of 0.50 and 30 inference steps. Guidance is 3.5 for FLUX.1-dev and 7.5 for the SDXL-derived generators. The canonical sequence is FLUX.1-dev $\rightarrow$ RealVisXL V5.0 $\rightarrow$ DreamShaperXL $\rightarrow$ FLUX.1-dev. At every hop, matched clean and watermarked branches use the same source prompt, transition seed, and generation settings. G0 is generated with SD2.1-base using 50 inference steps, guidance 7.5, and the default PNDM scheduler. No watermark is reinserted after G0.

A.2 ADDITIONAL QUALITATIVE RESULTS

We provide additional qualitative examples complementing the main-paper evaluation. Clean and RESON generations are shown under matched prompts and seeds, together with representative G0–G4 trajectories illustrating how marked images evolve under heterogeneous generative transitions.

A.2.1 MULTI-HOP LINEAGE EXAMPLES

Figure 4 compares G0–G4 lineages for RESON and the four baseline watermarks under the same prompt and the canonical lineage (SD2.1→FLUX→RealVisXL→DreamShaperXL→FLUX). Each method generates its clean and watermarked G0 images using its own generation pipeline; within each row, the clean and watermarked images share the same prompt and seed and differ only by the method's watermark. Each branch is then regenerated independently through the lineage without watermark reinsertion. Across all methods, the downstream generators substantially re-render the image while preserving the prompt semantics, illustrating why lineage persistence is challenging: any watermark signal must survive repeated re-synthesis in which pixel-level content is largely replaced. Corresponding detection rates at each depth are reported in Table 1.

A.3 ROBUSTNESS AND DETECTION DIAGNOSTICS

This section provides the complete robustness results and additional diagnostics omitted from the main paper for space. We distinguish conventional image perturbations from deliberate generative

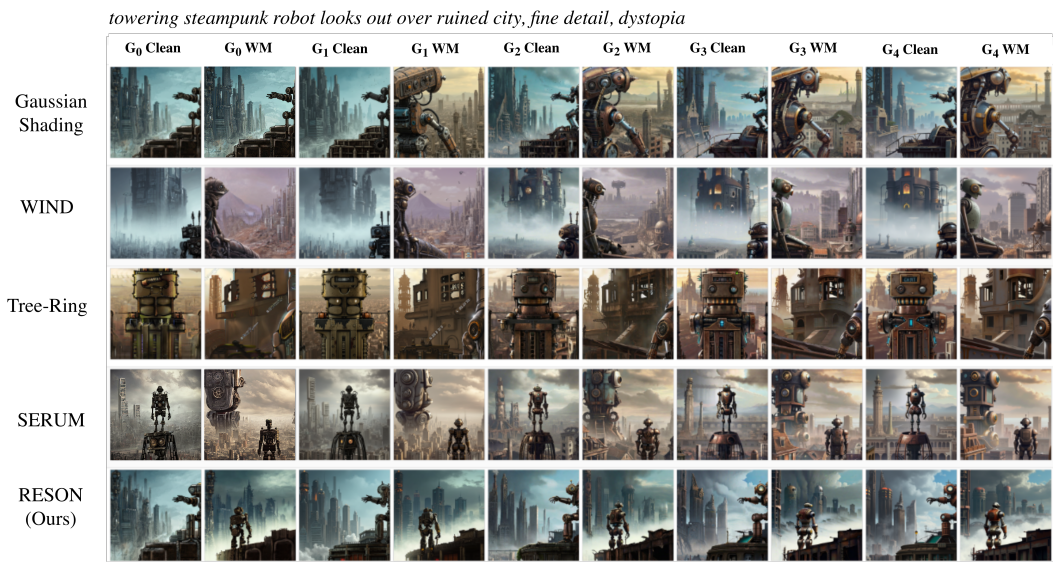

Figure 4: **Qualitative G0–G4 lineages for RESON and baseline watermarks.** Clean and watermarked images for the prompt *"towering steampunk robot looks out over ruined city, fine detail, dystopia"* at each depth of the canonical lineage (SD2.1→FLUX→RealVisXL→DreamShaperXL→FLUX). No watermark is reinserted after G0. Each method uses its own generation pipeline; detection rates are reported in Table 1.

removal: the former modify the rendered image through standard image-processing operations, whereas the latter reconstruct or regenerate image content through a learned generative model.

A.3.1 Complete Conventional-Perturbation Robustness

Figure 5 reports the complete severity sweeps for eight conventional perturbation families. Each condition contains $N = 1{,}000$ matched clean/watermarked pairs. The detector and operating threshold are fixed before perturbation evaluation; in particular, the presence-detection curves use the validation-frozen threshold and do not recalibrate the decision rule for individual perturbations. Payload-bit accuracy is evaluated independently on the corresponding watermarked images.

The severity sweeps reveal a distinction between calibrated watermark presence and persistence of the encoded payload. Under severe geometric transformations, the frozen presence detector can lose sensitivity even when the embedded bit remains recoverable. This behavior is most evident for center cropping and rotation, whereas compression, filtering, resizing, noise, and photometric transformations preserve both signals well over most of the evaluated range. We therefore report detection and payload recovery separately rather than treating them as a single robustness measure.

A.3.2 Deliberate Removal and Generative Reconstruction

In addition to conventional image perturbations, we evaluate deliberate removal through reconstruction and generative regeneration. These experiments are reported separately because generative reconstruction changes image content through a learned generative process rather than through a conventional pixel-space transformation.

An earlier controlled severity study used $N = 500$ matched clean/watermarked pairs per condition. VAE reconstruction achieved ROC-AUC 0.999808 and 100.0% TPR at 1% FPR. Prompt-conditioned regeneration produced TPRs of 98.8%, 98.6%, 98.0%, and 94.4% at strengths 0.20, 0.40, 0.60, and 0.80, respectively. Table 6 reports the controlled severity study.

This diagnostic indicates gradual degradation rather than immediate watermark collapse as regeneration strength increases. These settings represent internal regeneration-strength conditions and should not be interpreted as standardized Regen-12 or Regen-30 attack protocols.

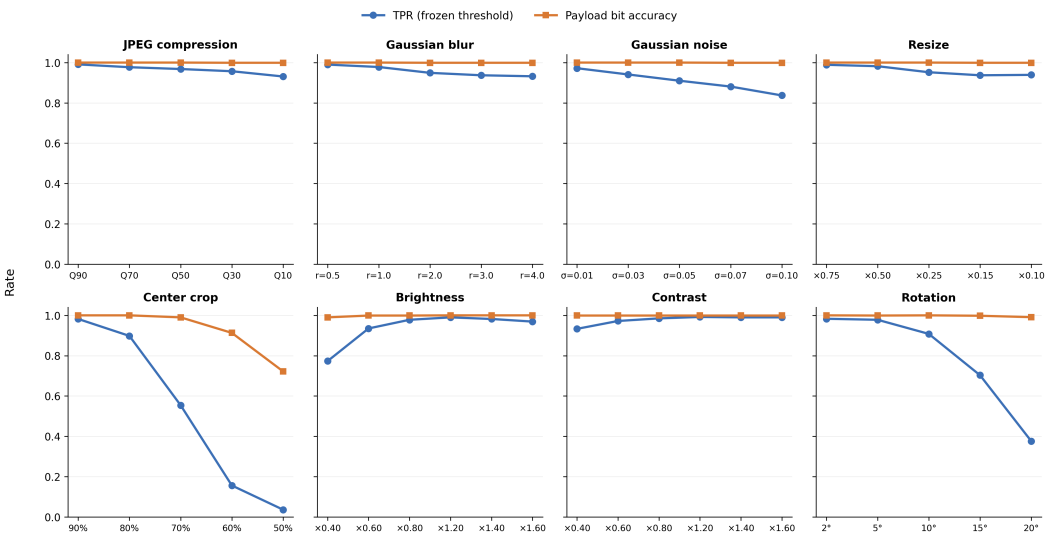

Figure 5: **Robustness across conventional image perturbations.** Detection TPR at the validation-frozen operating threshold and payload-bit accuracy across increasing perturbation severity. Error bars indicate 95% confidence intervals.

Table 6: **Controlled regeneration-severity diagnostic.** Operating points in this diagnostic are attack-specific and therefore should not be interpreted as evaluation under one universal frozen deployment threshold.

Attack	ROC-AUC	TPR@1% FPR	Realized FPR
VAE reconstruction	0.999808	100.0%	1.0%
Regeneration, $s = 0.20$	0.999640	98.8%	0.8%
Regeneration, $s = 0.40$	0.999528	98.6%	1.0%
Regeneration, $s = 0.60$	0.999396	98.0%	1.0%
Regeneration, $s = 0.80$	0.994924	94.4%	1.0%

Table 7 complements the main-paper removal results (Table 4) with oracle-threshold diagnostics. For each attack, the diagnostic threshold is chosen on the test set to maximize TPR subject to FPR $\leq$1%. It is reported only to assess how close the validation-frozen threshold is to the best achievable operating point, not as a deployment setting. The oracle threshold improves TPR by at most 1.5 percentage points over the frozen threshold.

Table 7: **Frozen versus oracle operating points under deliberate generative removal** ($N = 1{,}000$ pairs per attack). Frozen values use the validation-calibrated threshold; diagnostic values use the best test-set threshold at $\leq$1% FPR.

	Frozen threshold		Diagnostic threshold	
Attack	TPR (%)	FPR (%)	TPR (%)	FPR (%)
Regen-VAE q4	96.4	0.8	96.6	0.9
Regen-Diff t60	98.6	0.6	99.2	0.9
Regen-Diff t100	98.4	0.9	99.0	1.0
RINSE-2× t60	98.3	0.6	99.1	1.0
RINSE-4× t60	96.4	0.3	97.9	1.0

Attack configurations. Regen-VAE uses WatermarkAttacker's `bmshj2018-factorized` VAE attack with $q = 4$; Regen-Diff uses its SD1.4 diffusion attacker with noise step $t \in \{60, 100\}$; and

RINSE-$k\times$ repeats Regen-Diff ($t = 60$) for k rounds. All attacks use 1,000 held-out pairs with the detector and threshold frozen.

A.4 Message Capacity and Persistence

We evaluate how much information RESON can encode and how reliably that information survives repeated regeneration. We first measure source-level capacity under fixed total watermark energy, then evaluate recovery across the canonical G0–G4 lineage. Finally, we examine whether increasing message length affects generation quality.

A.4.1 Source-Level Capacity Sweep

Message lengths $K \in \{1, 2, 4, 8, 16, 32\}$ are evaluated under a common fixed-total-energy construction. All settings use 800 watermarked training images, 100 validation images, and 100 held-out watermarked test images. Table 8 reports mean bit accuracy, bit-error rate (BER), exact-message recovery, and mean Hamming error.

Table 8: **Source-level capacity under fixed total watermark energy.** Exact denotes recovery of the complete K-bit message.

K	Bit Acc. (%)	BER (%)	Exact (%)	Mean Hamming
1	**100.000**	0.000	**100.0**	0.00
2	97.500	2.500	95.0	0.05
4	94.500	5.500	80.0	0.22
8	86.625	13.375	33.0	1.07
16	77.188	22.812	0.0	3.65
32	68.656	31.344	0.0	10.03

Recovery exhibits a clear capacity–reliability trade-off. One- and two-bit messages are recovered almost perfectly, while four bits retain 94.5% mean bit accuracy and 80% exact-message recovery. At eight bits, bit accuracy remains 86.625%, but exact recovery falls to 33%. For 16 and 32 bits, individual bits remain recoverable above chance while complete-message recovery becomes unreliable. Because total watermark energy is fixed, this decline reflects increasing information density rather than stronger embedding.

A.4.2 Persistence Across Lineage Depth

We next test whether multi-bit information survives heterogeneous re-synthesis rather than only source-level decoding. Messages of 1, 2, 4, and 8 bits are propagated through the canonical G0–G4 lineage without watermark reinsertion. Figure 6 summarizes recovery as a function of lineage depth and message length.

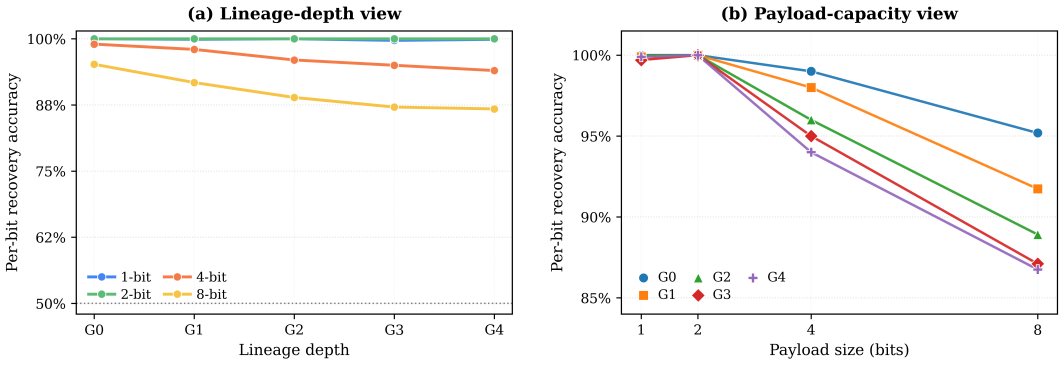

Figure 6: **Multi-bit recovery across the G0–G4 lineage.** Recovery decreases with both lineage depth and message length, revealing the capacity–persistence trade-off under repeated regeneration.

Multi-bit information remains recoverable after repeated heterogeneous re-synthesis, although robustness decreases with message length and lineage depth. The four-bit configuration, for example, retains 94.03% mean bit accuracy and 78.4% exact-message recovery at G4. This demonstrates that the surviving signal contains recoverable message information rather than merely supporting binary watermark-presence detection. Detailed four- and eight-bit results follow below.

A.4.3 Generation Quality Across Message Lengths

We also examine whether the capacity–robustness trade-off is accompanied by degradation in generation quality. Table 9 reports FID and CLIP similarity for clean generations and representative message lengths.

Table 9: **Generation quality across message lengths.** FID is computed against real COCO images; CLIP measures prompt–image similarity.

Clean		1-bit		4-bit		8-bit	
FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
17.829	0.30965	19.545	0.30890	19.286	0.30715	19.203	0.30704

FID and CLIP similarity remain comparable across the evaluated message lengths, with no monotonic degradation as capacity increases. Increasing capacity therefore primarily reduces recovery robustness rather than generation quality.

A.4.4 Four-Bit Depth-Wise Analysis

We evaluate four-bit recovery across the canonical G0–G4 lineage using $N = 1{,}000$ matched clean/watermarked pairs. The message is embedded once at G0 and is never reinserted downstream. Detection uses the same threshold calibrated on validation data to target 1% FPR at every lineage depth; because this threshold is frozen, the realized held-out FPR may differ from 1%. We therefore report the realized FPR alongside frozen-threshold TPR, and separately provide diagnostic TPR at an ROC-derived operating point satisfying FPR $\leq 1\%$. Table 10 reports the complete results.

Table 10: **Four-bit recovery across G0–G4.** Frozen TPR/FPR use the validation-frozen threshold; diagnostic TPR is measured at $\leq 1\%$ FPR. Bit Acc. is mean per-bit accuracy and Exact denotes complete-message recovery.

Depth	AUC ↑	Frozen TPR (%) ↑	Frozen FPR (%) ↓	TPR@ ≤1% FPR (%) ↑	Bit Acc. (%) ↑	Exact (%) ↑
G0	0.9965	94.7	1.5	93.1	99.13	96.5
G1	0.9831	82.8	1.1	82.0	97.85	91.8
G2	0.9594	65.7	1.1	65.5	95.78	83.8
G3	0.9362	52.7	0.9	53.5	94.48	80.2
G4	0.9383	54.0	0.8	58.0	94.03	78.4

After four heterogeneous transitions, mean bit accuracy remains 94.03%, with 78.4% exact four-bit recovery. Presence-detection sensitivity declines more sharply: at G4, frozen-threshold TPR is 54.0%, while diagnostic TPR at $\leq 1\%$ FPR is 58.0%. Thus, encoded information remains recoverable in many samples even after calibrated presence evidence has weakened.

A.4.5 Eight-Bit Depth-Wise Analysis

We repeat the lineage evaluation with an eight-bit message on $N = 1{,}000$ matched clean/watermarked pairs. As above, information is inserted only at G0 and propagated through the complete lineage without reinsertion or detector retraining. Table 11 reports the results.

The eight-bit setting exhibits the expected capacity–persistence trade-off. Mean bit accuracy decreases from 95.19% at G0 to 86.75% at G4, while exact recovery decreases from 68.0% to 34.2%. Presence detection declines more sharply, reaching 36.4% frozen-threshold TPR and 38.8% diagnostic TPR at

Table 11: **Eight-bit recovery across G0–G4.** Frozen TPR/FPR use the validation-frozen threshold; diagnostic TPR is measured at $\leq 1\%$ FPR. Bit Acc. is mean per-bit accuracy and Exact denotes complete-message recovery.

Depth	AUC ↑	Frozen TPR (%) ↑	Frozen FPR (%) ↓	TPR@ $\leq$1% FPR (%) ↑	Bit Acc. (%) ↑	Exact (%) ↑
G0	0.9815	82.4	1.8	76.5	95.19	68.0
G1	0.9528	64.7	1.6	60.0	91.74	51.5
G2	0.9227	46.5	0.8	50.8	88.91	40.6
G3	0.8908	36.8	1.1	35.4	87.10	35.8
G4	0.8828	36.4	1.0	38.8	86.75	34.2

$\leq 1\%$ FPR at G4. This again shows that recoverable encoded information can persist after calibrated watermark-presence evidence has substantially weakened.

A.5 Carrier and Strength Ablations

We analyze the watermark-strength parameter α and carrier-blend parameter β to characterize the trade-off among detectability, lineage persistence, and generation fidelity. These experiments complement the canonical operating-point analysis by isolating the contribution of carrier structure.

A.5.1 Depth-Wise Carrier Persistence

The controlled carrier experiment reported in Table 5 is examined here across the complete G0–G4 lineage. Because absolute source detectability differs among carrier constructions, we additionally normalize each condition by its G0 TPR to separate initial detectability from relative persistence with lineage depth.

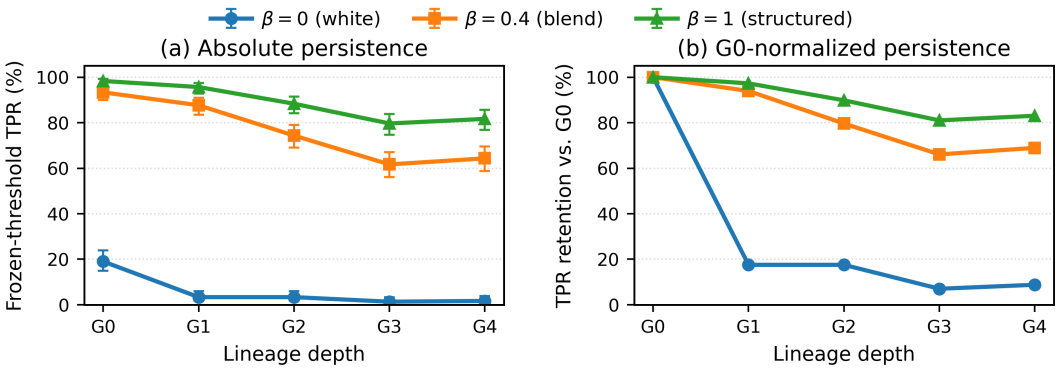

Figure 7: **Depth-wise carrier-structure ablation.** Frozen-threshold TPR (left) and G0-normalized retention (right) for $\beta \in \{0, 0.4, 1\}$ on $N = 300$ matched prompts. Error bars denote 95% confidence intervals.

The white carrier loses most of its detectable signal after the first regeneration, whereas both structured conditions degrade substantially more slowly. At G4, the canonical $\beta = 0.4$ blend retains 68.9% of its G0 detection rate, compared with 83.1% for $\beta = 1$. These trajectories reinforce the main-paper finding that carrier structure substantially contributes to persistence under repeated regeneration beyond injection energy alone. The stronger persistence of $\beta = 1$ does not make it the canonical system configuration: this experiment isolates carrier mechanism, whereas $\beta = 0.40$ is selected by the overall detection–fidelity trade-off.

A.6 Extended Attribution and Interference Experiments

A.6.1 Multiple-Watermark Interference

We evaluate whether adding a second, independently keyed watermark interferes with detection of an existing RESON watermark. Let $\mathbf{c}_A$ denote the original RESON carrier and $\mathbf{c}_B$ an independent smooth Fourier-band carrier, with $\mathbf{c}_A \perp_{\text{approx}} \mathbf{c}_B$. The second watermark is introduced with strength $\varepsilon_B \in \{0.002, 0.004, 0.008\}$, yielding

$$\mathbf{z}_{A+B} = \text{Norm}(\mathbf{z} + \alpha_A s_A \mathbf{c}_A + \varepsilon_B s_B \mathbf{c}_B)\,. \tag{7}$$

The detector for watermark A is kept unchanged; no retraining or adaptation is performed after introducing watermark B. We measure interference as the change in TPR at 1% FPR for detecting watermark A relative to the corresponding A-only condition. Table 12 reports the results.

Table 12: **Multiple-watermark interference.** Detection of the original watermark A is evaluated after adding an independently keyed watermark B. The A-only baseline is recomputed under the same protocol, so absolute TPRs differ slightly from Table 1. Δ denotes the change in TPR@1%FPR relative to the corresponding A-only condition.

Depth	ε_B	A-only TPR@1%	$A{+}B$ TPR@1%	Δ
G0	0.002	98.80%	98.80%	0.00 pp
G0	0.004	98.80%	98.77%	−0.03 pp
G0	0.008	98.80%	98.77%	−0.03 pp
G4	0.002	92.50%	92.50%	0.00 pp
G4	0.004	92.50%	92.50%	0.00 pp
G4	0.008	92.50%	92.45%	−0.05 pp

Interference with the original watermark is negligible across all tested strengths and depths. At G0, the largest observed reduction is only 0.03 percentage points, from 98.80% to 98.77% TPR@1%FPR. After four generative transitions, the two lower strengths match the A-only condition exactly, and the largest reduction (at $\varepsilon_B = 0.008$) is only 0.05 percentage points. Thus, within the evaluated interference regime, superposing a second independently keyed carrier does not measurably degrade detection of the original RESON watermark, including at the deepest evaluated lineage stage.

A.7 Additional Generation-Quality Analysis

Because RESON modifies the initial diffusion noise, clean and watermarked images need not follow pixel-aligned generation trajectories even under the same prompt. Consequently, semantic fidelity is evaluated primarily using prompt–image alignment and distribution-level quality metrics rather than interpreting paired pixel similarity as the sole measure of generation quality.

Table 13 reports generation quality on the full canonical benchmark of 10,000 matched clean/watermarked pairs. Because RESON perturbs the initial latent, a watermarked image is a different valid realization of the prompt rather than a pixel-level copy of the clean image. We therefore use prompt–image CLIP similarity of the watermarked image as the primary measure of prompt-conditioned semantic fidelity, and report paired clean-versus-watermarked differences as secondary diagnostics.

Across matched clean–watermarked pairs, the mean relative change in prompt–image CLIP similarity is -1.27%. The corresponding aggregate means are 0.2750 for clean generations and 0.2699 for RESON. RESON scores higher than its matched clean counterpart in 42.96% of pairs, while only 0.24% of watermarked images fall below a CLIP similarity of 0.15. Manual inspection of the lowest-scoring watermarked images and the pairs with the largest CLIP reduction found no systematic watermark-induced corruption; difficult cases were generally coherent alternative realizations of the prompt, and several corresponding clean images were similarly difficult.

Table 13: **Prompt-conditioned semantic fidelity on the canonical 10K benchmark.** Prompt–image CLIP similarity is computed with the CLIP ViT-L/14 image and text encoders (Radford et al., 2021). 95% confidence intervals are shown in brackets. The clean–RESON FID measures distribution shift between the clean and watermarked image sets; it is not a real-reference FID and is not comparable to the FID values in Table 9.

Metric	Value
Prompt–image CLIP similarity	
Clean, mean	0.2750 [0.2743, 0.2757]
RESON, mean	0.2699 [0.2692, 0.2707]
RESON, median	0.2707
RESON, 5th percentile	0.2063
RESON images below 0.15	0.24%
Paired clean-versus-watermarked diagnostics	
Mean ΔCLIP (RESON $-$ clean)	-0.0051
Mean relative ΔCLIP	-1.27%
Pairs where RESON scores higher	42.96%
Clean–RESON FID (distribution shift)	7.36

A.8 Lineage Trajectory and Failure Analysis

We further analyze how RESON's detection and payload recovery evolve across successive regeneration depths under the SD2.0 lineage (the secondary source model in Table 1). At the 1% FPR operating point, detection remains consistently high across the lineage, with TPRs of 100.00%, 98.33%, 94.33%, 93.67%, and 94.00% from G0 through G4, respectively. The corresponding AUC values are 1.0000, 0.9988, 0.9951, 0.9830, and 0.9824. Thus, most of the degradation occurs during the earlier regeneration steps, after which detection largely stabilizes: from G2 to G4, TPR varies by less than one percentage point, and G4 shows a slight recovery relative to G3 (94.00% versus 93.67%).

Payload recovery is substantially more stable than thresholded presence detection. Bit accuracy remains 100% at every depth from G0 through G4, including balanced 100% recovery for both bit-0 and bit-1 samples. This separation between presence detection and payload recovery indicates that successive regeneration progressively reduces detector confidence for a subset of samples without destroying the encoded payload. Even after four regeneration hops, RESON therefore retains both strong presence separability (AUC $= 0.9824$; TPR $= 94.00\%$ at 1% FPR) and perfect payload recovery on the evaluated SD2.0 lineage.