LATENT WATERMARKS UNDER GENERATIVE EDITING: A BENCHMARK AND ANALYSIS OF DETECTION SURVIVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordinary prompt-based editing can cause latent watermark detection to fail without explicitly targeting the watermark. We benchmark eight watermark methods against five editors across four generative backbones, four editing strengths, and five semantic categories, with edit-validity and threshold checks. Separating editing from seven subsequent distortions reveals that editing alone primarily distinguishes Tree-Ring, while added distortions expose a broader spectrum of detection survival. Sequential edits reveal a second hidden difference: score separation can decline while detection rates remain near their ceiling. Across methods, standardized clean score separation (d') organizes composite-survival tiers, whereas spatial overlap adds little to predicting edit-only survival beyond clean detectability. Embedding-strength interventions in two methods link higher clean separation to higher post-edit separation. In HSTR, the margin contrast is positive, while the angular layout contrast at matched clean separation remains unresolved. Together, outcome decomposition and continuous separation expose differences hidden by aggregate TPR. Method tiers are stable under threshold recalibration at the main operating points and alternative composite weights. Clean d' is thus a useful empirical diagnostic within this benchmark, with mixed transfer to unseen methods. Code will be released upon acceptance.

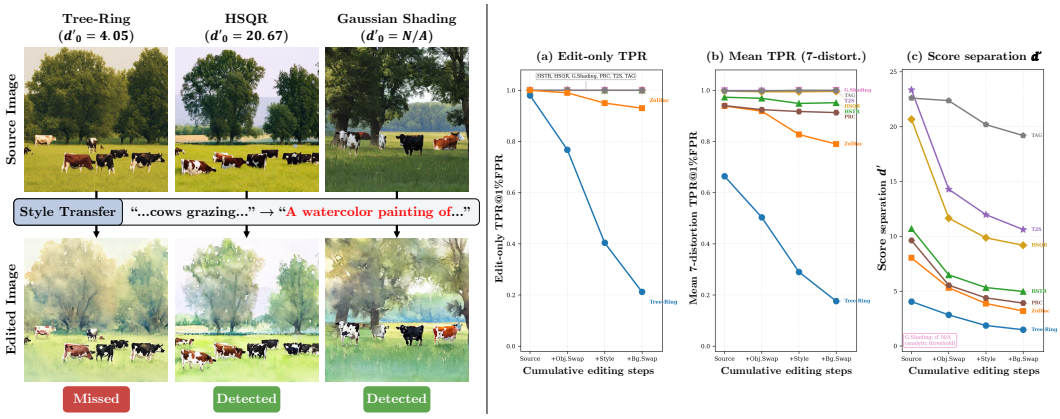


Figure 1: **Watermark survival under generative editing.** Left: InfEdit Style Transfer examples; d'_0 summarizes each method's source bank. Right: a separate cumulative InfEdit cascade (Source, Object Swap, Style Transfer, Background Swap), showing TPR before further distortion, mean TPR over seven separately applied distortions, and score separation. The TPR plots include all eight methods; d' is available for seven, with GS shown as N/A. Source counts and the five-step extension are reported in Section D.

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.

1 INTRODUCTION

Image watermarking provides a persistent signal that can complement provenance records such as C2PA (Coalition for Content Provenance and Authenticity, 2026). Generative approaches modify the decoder (Fernandez et al., 2023) or initial sampling noise z_T (Wen et al., 2023; Yang et al., 2024). Post-hoc approaches also embed messages in VAE-encoded image latents z_0, using learned perturbations (Bui et al., 2023), frequency-domain optimization (Guo et al., 2024), or direct phase modulation (Lee & Cho, 2026b).

Text-guided editors can resynthesize image content without explicitly targeting its watermark (Brooks et al., 2023; Kulikov et al., 2025). This has motivated editing-aware watermark designs, including Robust-Wide (Hu et al., 2024), JIGMARK (Pan et al., 2024), and VINE (Lu et al., 2025). Their development makes systematic evaluation across watermark designs and editing regimes increasingly important.

Existing evaluations span distortion and regeneration stress tests (An et al., 2024), broad editing benchmarks (Lu et al., 2025), and semantic perturbations of in-processing watermarks (Nakra & Wu, 2026). SafeMark additionally trains editors to preserve watermarks and evaluates subsequent distortions (Wu et al., 2026). A prior study of ordinary prompt-based editing proposed Semantic Imprinting to interpret differences in survival (Lee & Cho, 2026a). Building on these studies, we ask: under a common protocol, how much of the robustness gap arises from editing versus subsequent distortions, and which measured properties consistently organize that gap?

We evaluate eight initial-noise or inversion-based latent watermark designs under five editors, four generative backbones, four strengths, and five semantic categories, with validity and threshold checks. Editing alone primarily separates Tree-Ring from the other methods; seven subsequent distortions reveal a broader spectrum. Figure 1 illustrates these differences through examples and a cumulative editing sequence. Across the evaluated methods, clean score separation, d', organizes composite survival, while spatial overlap adds little when predicting edit-only survival. Within-method interventions and held-out analyses test the usefulness and limits of this empirical diagnostic.

Our contributions are **(1) a controlled benchmark** with explicit operating points and validity checks; **(2) an outcome decomposition** separating edit-only survival from subsequent-distortion robustness; and **(3) an empirical diagnosis** showing that clean score separation organizes composite survival and tracks within-method changes, while spatial tests and held-out evaluation delimit its explanatory and predictive scope.

2 RELATED WORK

Watermarks can be integrated through decoder fine-tuning (Fernandez et al., 2023), intermediate diffusion states (Huang et al., 2024), or initial-noise encoding. Post-hoc VAE-latent methods include RoSteALS (Bui et al., 2023), FreqMark (Guo et al., 2024), and PhaseMark (Lee & Cho, 2026b); their image latents z_0 differ from sampling noise z_T. Our evaluation focuses on eight initial-noise or inversion-based designs. Tree-Ring (Wen et al., 2023) and HSTR/HSQR (Lee & Cho, 2025) use Fourier patterning, whereas ZoDiac (Zhang et al., 2024) fits a latent post hoc. PRC (Gunn et al., 2025) uses a pseudorandom error-correcting code and T2SMark (Yang et al., 2025), abbreviated T2S, uses tail-truncated sampling; Gaussian Shading (Yang et al., 2024) and TAG-WM (Chen et al., 2025), abbreviated GS and TAG, use cryptographic bit encoding. SEAL binds watermark verification to image semantics to address forgery (Arabi et al., 2025); our outcome is signal survival, distinct from semantic authenticity.

Prompt-based generative editing. Prompt-to-Prompt (P2P) (Hertz et al., 2022) edits images via cross-attention injection; Null-text Inversion (Mokady et al., 2023) extends this to real images (P2P+NTI). InfEdit (Xu et al., 2024) performs inversion-free, attention-controlled editing on a consistency model. FlowEdit (Kulikov et al., 2025) edits flow models such as SD3 and FLUX via source-target velocity-field differences, without inversion or attention manipulation. These paradigms motivate our editor coverage.

Watermarking under editing. Robust-Wide incorporates instruction-driven editing into watermark training (Hu et al., 2024); JIGMARK learns from edited image pairs without backpropagating

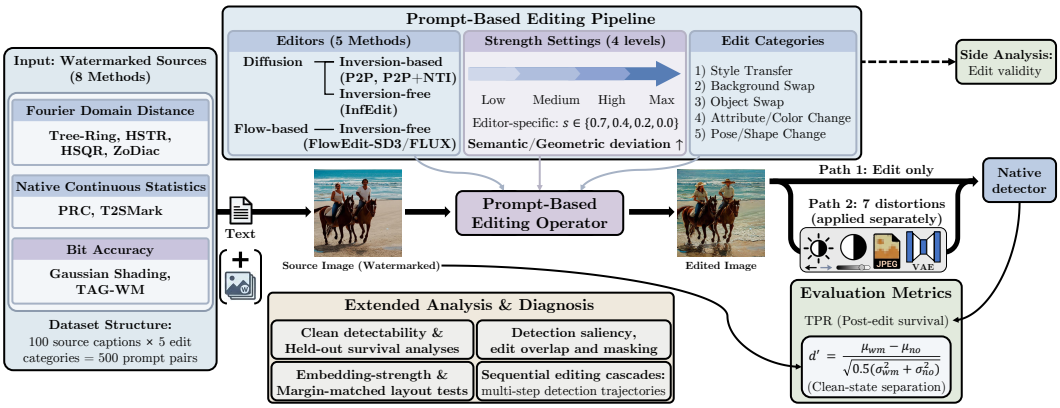


Figure 2: Benchmark and analysis framework. Eight watermark methods are evaluated under five editors, four strength settings, and five edit categories. Detection is measured before and after seven separately applied distortions. Low/Medium/High/Max and the deviation arrow indicate the intended progression of each editor's control settings; Max is the strongest setting tested.

through the editor (Pan et al., 2024); VINE uses generative priors and surrogate degradations (Lu et al., 2025). SafeMark instead optimizes the editor to preserve watermark decoding (Wu et al., 2026). We evaluate ordinary editors without such watermark-preservation objectives.

Benchmarks and empirical diagnostics. WAVES evaluates distortions, regeneration, and adversarial attacks (An et al., 2024), while W-Bench covers global/local editing, regeneration, and image-to-video generation (Lu et al., 2025). Nakra & Wu (2026) study semantic perturbations of Stable Signature, Tree-Ring, and GS; Zhao et al. (2024) show that generative regeneration provably removes watermarks whose image perturbation is ℓ_2-bounded. The Semantic Imprinting hypothesis (Lee & Cho, 2026a) offers a candidate interpretation of survival differences. We extend this empirical question through a shared multi-editor protocol, separate editing and subsequent-distortion outcomes, and examine spatial overlap and clean separation through complementary diagnostics, within-method controls, and held-out analyses.

3 Benchmark and Analysis Framework

Figure 2 summarizes the benchmark and analysis framework.

3.1 Threat Model and Benchmark Matrix

A watermark is introduced at the source image stage, either in the initial sampling latent at generation time or by fitting a post-hoc inverted latent to an existing image. The published image then undergoes ordinary prompt-based generative editing, without optimization against the watermark. Detection is blind: the detector operates on the edited output without the original image. Our operational outcome is watermark detection survival; edit validity, perceptual quality, and cryptographic security are separate evaluation dimensions.

Eight watermark methods and five editors spanning four generative backbones are evaluated at four settings, yielding $8 \times 5 \times 4 = 160$ method–editor–strength cells. Each cell uses 100 MS-COCO (Lin et al., 2014) source captions, with one target edit per caption in each of five categories. These repeated edits share the same source-caption identities.

Separating editing from subsequent distortions. *Edit only* measures TPR immediately after editing. *Post-distortion mean* averages TPR over seven distortions applied separately to that edited output. Avg_5 averages either measure over the five categories. The *8-condition composite* includes Edit only and all seven distortion TPRs, averaged within category and then across categories. The main table and strength-sweep figure separate the first two outcomes; the appendix also tabulates

the composite sweep used in the clean-d' analyses. The seven distortions are brightness, contrast, JPEG, Gaussian blur, BM3D, and two learned re-encodings: a fixed stress suite (Section E.1).

Thresholds target a nominal 1% false-positive rate (FPR); Sections A.1 and E.1 specify the operating rules and measured FPRs.

3.2 Evaluation Metrics and Edit Validity

Detection survival and score regimes. All methods are evaluated through their native statistic and a declared decision rule. Six methods use continuous statistics, while TAG uses bit accuracy; these seven use strict score $>$ empirical null q_{99}. GS uses the analytic bit-accuracy rule $\geq 147/256$, and TPR reports the fraction of surviving detections under each rule. Continuous-separation analyses use the six informative native statistics; TAG's finite d' is limited by clean-score saturation, and GS has no compatible empirical null estimate in this benchmark.

Standardized score separation. The sensitivity index d' from signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005) compares the watermarked and null score distributions. Let $\mu_{\mathrm{wm}}, \sigma^2_{\mathrm{wm}}$ and $\mu_{\mathrm{no}}, \sigma^2_{\mathrm{no}}$ denote their respective means and variances:

$$d' = \frac{\mu_{\mathrm{wm}} - \mu_{\mathrm{no}}}{\sqrt{0.5\left(\sigma^2_{\mathrm{wm}} + \sigma^2_{\mathrm{no}}\right)}}, \tag{1}$$

with scores oriented so that larger values indicate stronger watermark evidence. *Clean* d', also called clean detection margin, uses pre-edit source scores. Post-edit d' applies the same definition to edited scores with the corresponding null reference. This standardized separation is distinct from distance to an operating threshold. Section F gives score conventions; the controlled analyses specify their aggregation units.

Edit validity and severity. Relative to the watermarked editing input, LPIPS (Zhang et al., 2018) measures perceptual distance and DINOv2 (Oquab et al., 2023) measures structural/content similarity. ΔCLIP (Radford et al., 2021) and ΔBLIP2 ITM (Li et al., 2023) measure changes in target-text alignment, using cosine similarity and positive matching probability, respectively. LPIPS also serves as a per-image edit-magnitude covariate in the severity analysis. Sections E.2 and E.5 detail these metrics and their use.

3.3 Candidate Explanations and Controlled Tests

We examine two candidate explanations for survival differences: the spatial relationship between detector sensitivity and editing changes, and clean score separation. We then vary embedding strength within HSTR and Tree-Ring, vary HSTR angular layout at matched clean separation, and test whether the observed associations transfer to held-out methods after accounting for edit magnitude.

Detection saliency. For a detector score $S(x)$ evaluated on the RGB query x, saliency measures image-space sensitivity:

$$M_{\mathrm{sal}}(x) = \mathrm{Norm}\left(\frac{1}{C}\sum_{c=1}^{C}\left|\frac{\partial S(x)}{\partial x_c}\right|\right), \tag{2}$$

where C is the channel count and Norm is min-max normalization to $[0, 1]$. Section F describes the gradient chains and spatial surrogates.

Maps of editing changes. Cross-attention differences summarize source-to-target attention changes during editing; pixel changes measure source-to-edited RGB differences. Qualitative comparisons use attention differences for P2P, P2P+NTI, and InfEdit, and pixel changes for FlowEdit. Quantitative overlap uses pixel changes for every editor, providing a common image-space comparison. Section F.1 defines the maps.

Table 1: Watermark survival at the main editing operating points. Entries are Avg_5 TPR over five edit categories. Edit only precedes added distortion; shaded Post-dist. columns average seven distortions applied separately to that output. P2P, P2P+NTI, InfEdit, and FlowEdit-SD3 use s=0.4; FlowEdit-FLUX uses s=0.2, $n_{\max}$=23. Section E.1 lists native editor settings.

	P2P		P2P+NTI		InfEdit		FlowEdit-SD3		FlowEdit-FLUX	
Method	Edit only	Post-dist.	Edit only	Post-dist.	Edit only	Post-dist.	Edit only	Post-dist.	Edit only	Post-dist.
Tree-Ring	0.632	0.356	0.804	0.471	0.682	0.444	0.692	0.403	0.734	0.427
ZoDiac	0.996	0.818	1.000	0.873	0.992	0.908	0.986	0.877	0.992	0.880
PRC	0.994	0.800	1.000	0.861	1.000	0.928	1.000	0.861	1.000	0.848
HSTR	1.000	0.934	1.000	0.952	1.000	0.965	1.000	0.944	1.000	0.937
HSQR	1.000	0.991	1.000	0.996	1.000	0.997	1.000	0.992	1.000	0.991
TAG	1.000	0.999	1.000	0.997	1.000	0.997	1.000	0.998	1.000	0.997
T2S	1.000	0.998	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999
GS	1.000	0.998	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998

4 Robustness under Generative Editing

4.1 Editing-Only Survival and Subsequent Distortions

Table 1 reports Avg_5 TPR for all eight methods and five editors at the main operating point, separating the edited output before additional distortion (Edit only) from the mean over seven distortions applied separately after editing (Post-distortion mean). Editing alone already distinguishes Tree-Ring: its Edit-only TPR ranges from 0.632 to 0.804, whereas every other method lies between 0.986 and 1.000. The smallest within-editor gap is 0.196. Tree-Ring's lowest Edit-only TPR (0.632) occurs under P2P, which shares the SD2.1 source backbone (Section E.1).

Subsequent distortions widen the Tree-Ring gap and separate ZoDiac, PRC, and HSTR from the near-ceiling methods. This broader spectrum also appears in the 8-condition composite used for the clean-d' analyses. Table 8 reports the composite's mean and worst-category values together with the clean d' values used in those analyses.

Operating-point sensitivity. Realized FPRs vary by method around the nominal 1% (Section A.1). Recalibrating on a 600-image null bank changes composite TPR by at most 0.042 at the five main operating points and preserves every pairwise order among the seven empirically thresholded methods whose original gap is at least 0.02. Its 500-image cross-fitted counterpart yields method-wise condition-averaged FPRs of 0.8–1.2%. Stricter rules also preserve the ordering of the five tested methods (Section A.3); Section D.3 reports repeated-seed intervals.

4.2 Sensitivity to Editor-Specific Settings

Figure 3 separates edit-only survival from post-distortion Mean7 across the full sweep. Tree-Ring remains below the other methods in edit-only survival; ZoDiac, PRC, and HSTR also lose survival at some settings, while HSQR, TAG, T2S, and GS remain at the ceiling. Subsequent distortions expose a broader spread. Both outcomes show greater setting dependence for the flow-based editors than for P2P or P2P+NTI. Setting definitions appear in Section E.1.

Style Transfer is the most frequent worst category under the composite (77 of 160 cells, 48%); category details appear in Section B.2.

4.3 Edit Validity and Sanity Checks

Table 2 reports perceptual and positive average target-alignment changes across the five original methods, with an HSQR example showing object changes. At the selected budget, FlowEdit-FLUX yields smaller changes: LPIPS and ΔBLIP2 ITM are 0.187 and 0.064, versus 0.205 and 0.321 for FlowEdit-SD3. Section B.3 gives the solver-budget sweep.

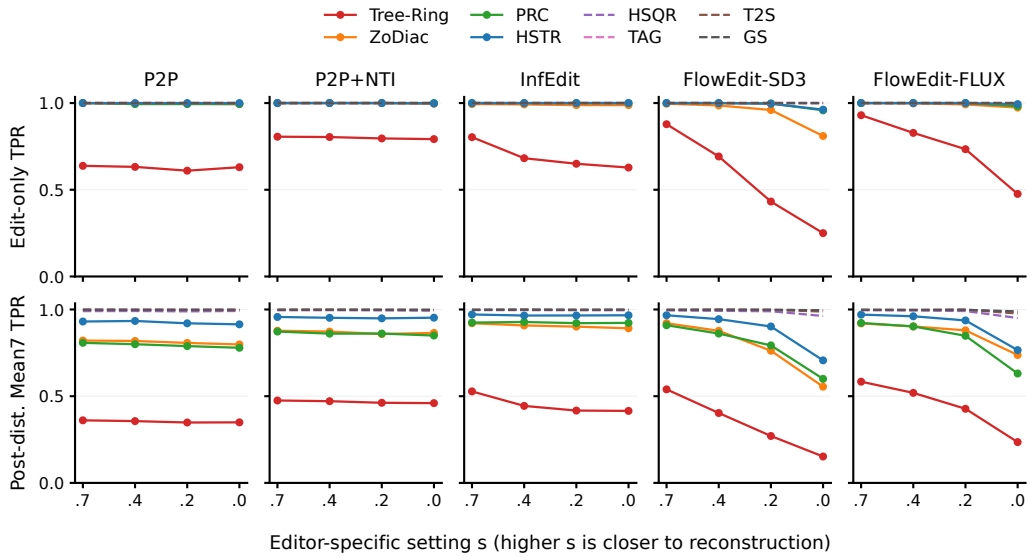


Figure 3: Edit-only TPR (top) and the mean TPR over seven separately applied post-edit distortions (bottom), across eight methods, five editors and four settings. Each point averages five categories equally. Higher s is closer to reconstruction within an editor. Table 9 retains the corresponding 8-condition composite sweep.

Table 2: Edit magnitude and target alignment relative to the watermarked input at the main operating points. Top: each editor aggregates Tree-Ring, ZoDiac, HSTR, HSQR, and GS over five categories and 100 source captions. LPIPS is a distance, DINO a similarity, and the two deltas measure changes in target alignment; Section E.2 defines the metrics. Bottom: a selected HSQR Object Swap example (source 59), shown for illustration rather than success-rate estimation.

Editor	LPIPS	DINO	ΔCLIP	ΔBLIP2 ITM
P2P	0.282	0.793	0.020	0.177
P2P+NTI	0.283	0.747	0.025	0.202
InfEdit	0.250	0.751	0.034	0.308
FlowEdit-SD3	0.205	0.751	0.044	0.321
FlowEdit-FLUX	0.187	0.812	0.018	0.064

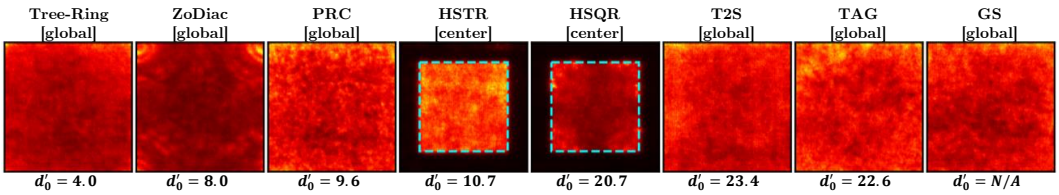


HSQR source | P2P | P2P+NTI | InfEdit | FlowEdit-SD3 | FlowEdit-FLUX

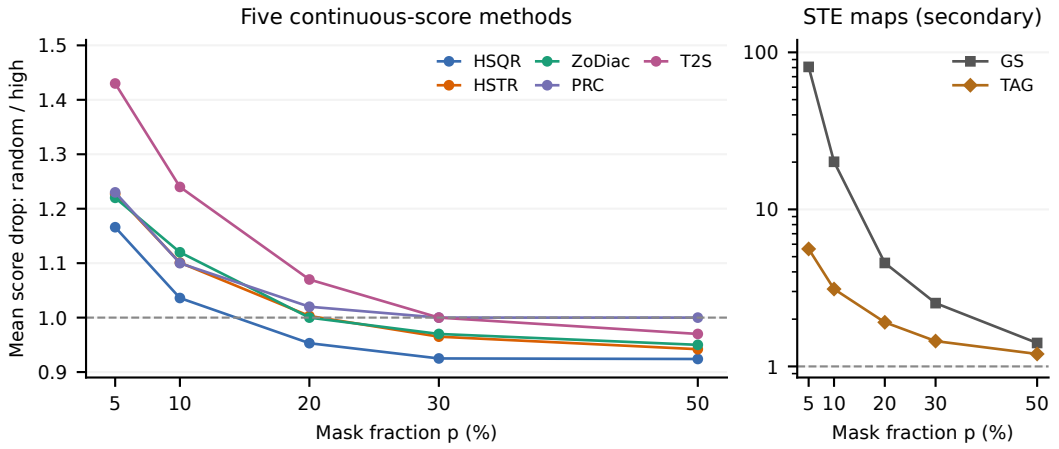
Target: “Two tigers in a jungle area fighting in a dirt area.” All outputs use the same source and target prompt at the main editor-specific settings in Table 1.

5 What Explains the Robustness Gap?

We first test whether spatial overlap adds predictive information for edit-only survival, then relate clean separation to composite survival across methods. Within-method interventions probe changes in score separation, and a final analysis tests severity adjustment and transfer of per-image detection predictions.



(a) Source sensitivity: eight methods



(b) Masking: seven methods

Mean top-10% IoU (8 methods)	
Observed	0.022
Uniform reference	0.053

Edit-only prediction (6 methods)	
Pooled Δlog-loss	-0.0004
LOMO Δlog-loss	$+0.0005$

(c) Overlap with RGB editing changes

Figure 4: Spatial diagnostics. **(a)** Mean source sensitivity (n=100 per method); cyan guides mark the HSTR/HSQR latent-crop extent. Normalized maps do not compare absolute sensitivity. **(b)** Random-to-high-saliency mean score-drop ratios; 1 means equal mean damage. GS/TAG use STE maps and a separate log axis. Low-saliency results remain in Figure 8. **(c)** The uniform IoU reference is geometric; Δlog-loss adds overlap to clean d' ($-$ is better).

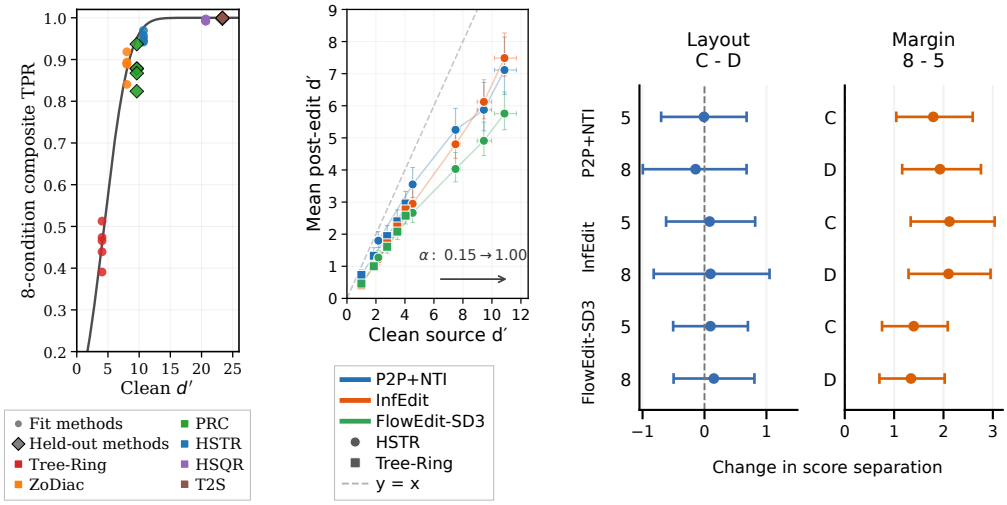
5.1 Testing a Spatial Explanation

A spatial account suggests that editing preserves detection by avoiding detector-sensitive regions. We test source sensitivity, its ranking through masking, and overlap with editing changes.

Source-map distribution. Figure 4(a) shows central RGB sensitivity for HSTR/HSQR and broader maps for other methods. Spatial statistics show no monotone ordering with clean d' or composite TPR (Table 15); these maps measure sensitivity rather than watermark location.

Masking as a ranking check. We mask equally sized pixel sets in source images before editing. Random-to-high-saliency score-drop ratios lie on both sides of 1 for continuous scores at larger fractions; GS/TAG remain above 1 (Figure 4b). Low-saliency masking is not uniformly less damaging: HSTR's low-saliency drop exceeds its high-saliency drop at p=30% and 50%. The tested ranking thus offers no consistent targeted-removal advantage (Figure 8 and tables 13 and 14).

Overlap and edit-only survival. Figure 4(c) uses RGB changes for every editor. IoU is below the independent-uniform reference; adding overlap to clean d' barely reduces pooled log-loss and slightly worsens leave-one-method-out (LOMO) prediction. Tree-Ring supplies most informative held-out variation. Sections C.3 and C.4 give the quantitative and qualitative comparisons.



(a) **(b)** **(c)**

Figure 5: **(a)** Clean d' versus 8-condition composite TPR with the global probit summary; circles denote fitted methods and diamonds held-out methods. **(b)** Strength intervention: clean versus mean category-level pre-distortion post-edit d', with marginal 95% bootstrap intervals. The arrow marks increasing α along each series. **(c)** HSTR effects on separation of source-mean scores, with paired-complete 95% bootstrap intervals. Left rows give matched clean targets 5/8; right rows fix Compact (C) or Distributed (D). The two effect axes have different ranges. All six layout intervals include zero; all six margin intervals are positive. Sections B.4, B.5 and H.2 give the estimators.

5.2 Clean Detectability Organizes Composite Survival

Edit-only survival is near ceiling for most methods, so we examine *8-condition composite TPR*. Seven of its eight components include added distortions. Its tiers—Tree-Ring $<$ {ZoDiac, PRC} $<$ HSTR $<$ {HSQR, TAG, T2S, GS}—are unchanged in all 20 settings when any distortion is dropped, the Edit-only term is double-weighted, or only the seven distortions are averaged (Section B.1). Among methods with informative clean d', these tiers follow its order. Figure 5(a) shows six methods at the main operating points: each point averages five categories, sharing method-level clean d' across editors. This aggregate association is distinct from per-image edit-only prediction.

A global probit curve fitted to four methods summarizes this association, with RMSE 0.086 on 40 held-out PRC/T2S cells. Section B.4 distinguishes this global fit from condition-specific fits and their method-wise LOMO errors.

Sequential edits and score separation. Continuous separation can reveal changes hidden by ceiling TPR. Figure 1 shows the first three cumulative InfEdit edits, and Section D extends the chain to five. By Step 5, Tree-Ring edit-only TPR falls from 0.98 to 0.061, while five methods retain 1.00, including four with decreasing d'. Mean TPR over seven subsequent distortions also separates their responses: PRC falls from 0.940 to 0.870, whereas HSQR, GS, T2S, and TAG remain near the ceiling. After the first edit, the six methods with informative clean separation retain 0.56–0.70 of their source d'; in later steps, higher-separation methods generally retain a larger fraction per edit (Table 16; Figure 14). These are descriptive trajectories, motivating separation alongside survival rates.

5.3 Controlled Tests within Methods

Embedding strength: does separation change within a method? We vary embedding strength α in HSTR and Tree-Ring under P2P+NTI, InfEdit, and FlowEdit-SD3. Figure 5(b) compares clean d' with pre-distortion post-edit d', averaged over five categories. Post-edit separation increases monotonically with α in all six method–editor conditions (Spearman ρ=1.0; bootstrap $P(\text{monotonic}) \geq 0.9993$). Across 30 aggregates, clean margin alone gives R^2=0.9608. Section B.5

provides the aggregation and resampling details, and Table 12 documents the accompanying source-fidelity cost on 100 sources per method and strength.

Angular layout: what changes at matched clean separation? Within HSTR, Compact and Distributed angular supports have equal coefficient budgets and matched radial histograms at clean-d' targets near 5 and 8. Using paired complete sources and their category-mean native scores, Figure 5(c) contrasts layout at matched separation with margin targets at fixed layout. All six layout intervals include zero, with point estimates from -0.144 to $+0.151$; all six margin intervals exclude zero, with effects from $+1.342$ to $+2.122$. Section H reports the controls and full intervals.

These within-method tests target separation; Table 12 and section H quantify their fidelity costs.

5.4 Controlling for Severity and Holding Out Methods

We predict *per-image edit-only detection* from clean d' and input-to-output LPIPS across six methods, with repeated source identities (Section E.5). On the covered common-LPIPS-support cohort, five-bin reweighting changes Edit-only TPR by at most 0.014, preserving the observed separation between Tree-Ring and the near-ceiling methods.

Combining d' and LPIPS improves pooled AUC but worsens held-out log-loss for Tree-Ring and the two-method macro, while improving ZoDiac (Table 3). Thus the pooled gain does not transfer reliably. Editor terms add little to pooled AUC (Section E.5).

Table 3: Per-image edit-only logistic prediction: pooled AUC ↑ (left) and held-out log-loss ↓ (right). Macro averages Tree-Ring and ZoDiac; the other four held-out methods are near ceiling. Section E.5 specifies the models.

Pooled model	AUC ↑
LPIPS only	0.633
Clean d' only	0.916
d' + LPIPS	0.945

Held-out method	d' only	+ LPIPS
Tree-Ring	0.648	0.845
ZoDiac	0.089	0.072
Macro (two methods)	0.369	0.459

6 Discussion, Limitations, and Conclusion

Findings and practical use. Editing primarily separates Tree-Ring; subsequent distortions expose a broader spectrum organized by clean separation. Sequential edits reveal declining separation beneath ceiling TPR, and within-method interventions connect clean and post-edit separation. Spatial overlap adds little predictive information, while the matched-separation HSTR layout contrast remains unresolved.

Recommended evaluation protocol. Report Edit-only survival separately from the mean over subsequent distortions. When TPR approaches its ceiling, include d' where informative score distributions are available. Track separation through at least a short cumulative sequence, such as Object Swap, Style Transfer, and Background Swap (Figure 1). Check realized FPR on held-out null images and assess threshold recalibration, using cross-fitting to evaluate that procedure (Section A.3).

Limitations. The clean-d' relationship is empirical, supported by within-method interventions and first-edit contraction; interventions also change source fidelity (Table 12 and section H). Mixed held-out transfer makes the diagnostic complementary to direct editing evaluation. Realized FPR varies, but recalibration preserves separated pairs at the main points. Editor settings and distortion severities are not perceptually matched, yet tiers survive all tested weight changes; the Edit-only LPIPS check does not match semantics. GS/TAG restrict continuous-margin comparison but retain native-rule TPR evaluation. RGB spatial tests have qualitative attention-map counterparts. Editing-aware learned watermarks and autoregressive editors such as VAREdit (Mao et al., 2026) are outside the evaluated scope; the decomposition applies unchanged, and evaluating these families is a natural next step.

AI USE STATEMENT

The research questions and core ideas were developed by the authors. Large language models assisted literature retrieval, drafting and polishing, and research execution: experimental and analysis code, debugging, result checking and interpretation, and figure and table preparation. The authors reviewed the AI-assisted work against experimental evidence and take responsibility for the final code, results, citations, and claims.

ETHICS STATEMENT

This work studies the robustness of generative-image watermarks under published diffusion and flow editing methods. Such evaluation can inform both more robust watermark design and attempts at evasion; we report it to clarify existing systems' limits. Experiments use public pretrained models and MS-COCO image-caption data, without recruiting human participants or collecting private personal data.

REPRODUCIBILITY STATEMENT

The benchmark definitions, operating points, and analysis protocols are documented in Sections 3, E.1, E.3 and E.5; Tables 19 and 21 list the editor and watermark settings. Sample counts, filtering, bootstrap procedures, and validation scope accompany the corresponding analyses. Section E.6 identifies the core pipeline and the separate result artifacts required for the later controlled analyses. The prompt pairs, per-image scores, null banks, and configurations will be released with the code.

REFERENCES

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. WAVES: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

Kasra Arabi, R Teal Witter, Chinmay Hegde, and Niv Cohen. SEAL: Semantic aware image watermarking. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16196–16205. IEEE, 2025.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. RoSteALS: Robust steganography using autoencoder latent space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 933–942. IEEE, 2023.

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. IEEE, 2025.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7936–7945. IEEE, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html, 2026. Version 2.4.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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. IEEE, 2023.

David M. Green and John A. Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966.

Samuel Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *International Conference on Learning Representations*, volume 2025, pp. 6612–6637, 2025.

Yiyang Guo, Ruizhe Li, Mude Hui, Hanzhong Guo, Chen Zhang, Chuangjian Cai, Le Wan, and Shangfei Wang. FreqMark: Invisible image watermarking via frequency based optimization in latent space. *Advances in Neural Information Processing Systems*, 37:112237–112261, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Runyi Hu, Jie Zhang, Ting Xu, Jiwei Li, and Tianwei Zhang. Robust-Wide: Robust watermarking against instruction-driven image editing. In *European conference on computer vision*, pp. 20–37. Springer, 2024.

Huayang Huang, Yu Wu, and Qian Wang. ROBIN: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Sung Ju Lee and Nam Ik Cho. Semantic watermarking reinvented: Enhancing robustness and generation quality with Fourier integrity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18759–18769. IEEE, 2025.

Sung Ju Lee and Nam Ik Cho. The Semantic Imprinting Hypothesis: How semantic watermarks survive prompt-based editing. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI-Interpretability, Robustness, and Safety across Modalities*, 2026a.

Sung Ju Lee and Nam Ik Cho. PhaseMark: A post-hoc, optimization-free watermarking of AI-generated images in the latent frequency domain. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13472–13476. IEEE, 2026b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *International Conference on Learning Representations*, volume 2025, pp. 83902–83936, 2025.

Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User's Guide*. Psychology Press, New York, 2nd edition, 2005.

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. In *International Conference on Learning Representations*, volume 2026, pp. 85702–85723, 2026.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047. IEEE, 2023.

Anirudh Nakra and Min Wu. Understanding semantic perturbations on in-processing generative image watermarks. *arXiv preprint arXiv:2603.27513*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Minzhou Pan, Yi Zeng, Xue Lin, Ning Yu, Cho-Jui Hsieh, Peter Henderson, and Ruoxi Jia. JIGMARK: A black-box approach for enhancing image watermarks against diffusion model edits. *arXiv preprint arXiv:2406.03720*, 2024.

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. IEEE, 2022.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063, 2023.

Xiaodong Wu, Qi Li, Xiangman Li, Zelin Zhang, Lingshuang Liu, and Jianbing Ni. Are watermarked images editable? SafeMark for watermark-preserving text-guided image editing. *arXiv preprint arXiv:2605.19511*, 2026.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9454–9461. IEEE, 2024.

Jindong Yang, Han Fang, Weiming Zhang, Nenghai Yu, and Kejiang Chen. T2SMark: balancing robustness and diversity in noise-as-watermark for diffusion models. *Advances in Neural Information Processing Systems*, 38:118642–118668, 2025.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12162–12171. IEEE, 2024.

Lijun Zhang, Xiao Liu, Antoni V Martin, Cindy X Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using Stable Diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

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. *Advances in neural information processing systems*, 37:8643–8672, 2024.