SRW: FROM DISCRETE WATERMARK IDENTIFICATION TO SEMANTIC PROVENANCE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarking embeds recoverable source information in AI-generated images to support attribution and auditing after dissemination. Compression, cropping, resizing, and editing can damage this signal, while incomplete provenance records create a second obstacle: recovering a discrete identifier does not by itself retrieve a semantically equivalent source description. We introduce SRW, which reframes watermark identification as semantic provenance retrieval. A provenance payload is source metadata intentionally embedded in the image and may differ from the prompt controlling visible content. SRW selects informative payload tokens, injects their continuous representation into diffusion latents, and learns to recover it from distorted images for cosine-ranked search with calibrated rejection. In the recorded attack sweep, retrieval accuracy is 94% after JPEG quality 50, 95% after $0.5\times$ resizing, and 94% after a 20% crop. In a separate evaluation excluding the exact source text, Recall@1 reaches 93% for paraphrases and 82% for hard semantic neighbors. Payload-shuffling summaries further support recovery of the embedded payload rather than visible content alone. Bank-size diagnostics give 98% Top-1 at 1K candidates and 92% at 10K. These results support a semantic retrieval interface beyond discrete identity lookup, with strong impulse noise, image fidelity, and joint distortion–source-absence evaluation remaining limitations.

1 INTRODUCTION

Watermarking for AI-generated images embeds a recoverable signal into image pixels or the generation process to associate the resulting content with source information (Zhu et al., 2018; Fernandez et al., 2023). It is a key mechanism for supporting copyright attribution, compliance auditing, and accountability in AI-generated content (AIGC). A generation platform embeds a source signal before release; a verifier subsequently extracts or tests that signal from a received image and compares it with registered source records. During dissemination, images commonly undergo compression, cropping, resizing, color adjustment, and further editing. These operations distort the watermark or remove parts of its carrier, while external metadata may also be lost. The central challenge is therefore to recover useful source information reliably from the residual signal in a disseminated image.

Existing approaches address this challenge through three distinct signal types. *First, classical and neural watermarks* recover bit strings or discrete identifiers (Cox et al., 1997; Zhu et al., 2018; Tancik et al., 2020); their native attribution endpoint is a registered message or identity. Error correction can tolerate signal corruption, but codeword proximity need not express that two source descriptions are paraphrases. If decoding or the associated record lookup fails, the identifier alone supplies no semantic fallback. *Second, diffusion-native watermarks* exploit latent carriers for binary detection, signature recovery, or key identification (Fernandez et al., 2023; Wen et al., 2023; Yang et al., 2024; Ci et al., 2024). A continuous carrier does not make the decoded identity space semantically structured: crossing an identity decision boundary can yield another key without a meaningful source relationship. *Third, CLIP-style image–text retrieval* ranks descriptions using visual–language alignment (Radford et al., 2021). Its native objective provides no binding between the image and arbitrary hidden metadata deliberately chosen by its generator. Thus robust signal recovery, discrete attribution, and visible-content retrieval each leave a different part of semantic provenance unresolved.

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.

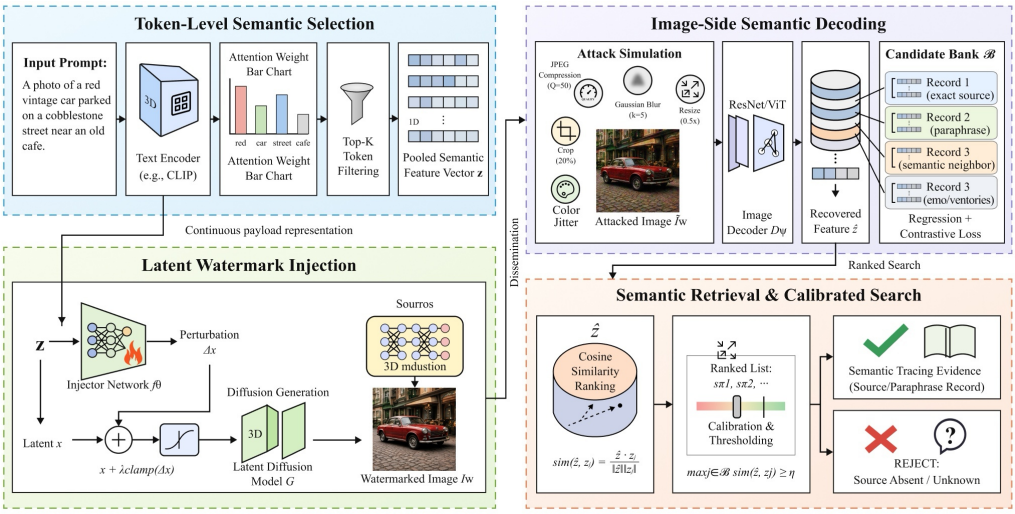

Figure 1: SRW separates the generation prompt p from the embedded provenance payload s. It recovers the payload representation after dissemination and searches a semantic candidate bank. The common case $s = p$ supports prompt attribution; allowing $s \neq p$ makes hidden metadata binding testable. A retrieved semantic neighbor is tracing evidence and does not establish exact identity.

A *provenance payload* s is source metadata intentionally encoded by the watermark, such as a source description, task, or workflow label. It is distinct from the *generation prompt* p, which controls visible image content. For example, p may describe a red car while s records an advertising workflow; that workflow need not be visually inferable. Prompt attribution is the special case $s = p$. If the exact payload text is missing from a candidate bank but a paraphrase or related record remains, exact lookup cannot express that relationship. Detection or an intact ID-to-record lookup may still succeed, but the native ID representation does not rank alternative descriptions by source meaning.

This gap motivates *semantic provenance retrieval*: preserving enough of the embedded payload's meaning to search for relevant source records, even when exact text matching is unavailable. In a continuous semantic space, related descriptions can occupy nearby directions, allowing an imperfectly recovered vector to retain a useful neighborhood. This provides a basis for tolerance to decoding error and textual reformulation, provided semantic separation exceeds recovery error; it does not guarantee robustness to every image edit. When the source description is present, retrieval targets that record. When it is absent, retrieval returns semantically relevant alternatives, with a calibrated threshold permitting rejection. Such alternatives provide semantic tracing evidence, not proof of an exact author or generation event. Watermark detection remains useful, but it is insufficient to answer this richer provenance question.

We introduce **SRW**, a framework that implements this shift from discrete identification to semantic provenance retrieval. SRW selects informative payload tokens, pools their representations, injects the resulting vector into diffusion latents, and trains an image-side decoder under simulated dissemination. The recovered vector queries a bank encoded by the same semantic selector, using cosine ranking and optional calibrated rejection (Figure 1). The representation carries the payload's meaning, while the trained image channel aims to preserve its recoverability under distortion.

Our main result measures retrieval after dissemination: accuracy is 94% for JPEG quality 50, 95% for $0.5\times$ resizing, and 94% for a 20% crop. A separate source-absence test reaches 93% Recall@1 for paraphrases and 82% for hard semantic neighbors. Payload-shuffling summaries support hidden metadata binding; bank-size and threshold diagnostics characterize operating limits. The joint distortion–source-absence setting remains untested.

Our contributions are threefold:

- We formalize semantic provenance retrieval beyond discrete identification, separating exact attribution, semantic-neighbor tracing, and rejection.

- We develop a semantic watermark channel for payload selection, latent injection, recovery under distortion, and semantic search.
- We evaluate dissemination robustness and source-absent retrieval, supported by payload-binding controls and diagnostics of scaling, calibration, and failure modes.

2 Related Work

Robust image watermarking. Transform-domain watermarking embeds low-capacity signals that can be detected after moderate signal processing (Cox et al., 1997; Barni et al., 1998; Petitcolas et al., 1999). Learned systems instead optimize the encoder, corruption layer, and decoder jointly. HiDDeN (Zhu et al., 2018) established this end-to-end pattern for bit recovery, while StegaStamp (Tancik et al., 2020) targeted physical and geometric transformations. These methods provide the robustness machinery that motivates our attack layer, but their native payloads remain bit strings. SRW changes the payload space and the downstream objective: recovering a representation whose neighborhood structure is meaningful for provenance search.

Diffusion-native watermarking and identification. Watermarking the generation process can avoid a detachable post-hoc step. Stable Signature fine-tunes the latent decoder to emit a recoverable signature (Fernandez et al., 2023); Tree-Ring modifies the initial noise in Fourier space (Wen et al., 2023); Gaussian Shading maps watermarks to distribution-preserving latent samples (Yang et al., 2024); and RingID improves multi-key identification (Ci et al., 2024). These systems establish that diffusion latents are effective carriers, yet their standard decisions are detection or identity matching. Robust carrier design can protect those decisions without assigning semantic relationships to different identities. SRW instead associates each carrier signal with a continuous semantic payload and evaluates retrieval when the exact text is not available.

Vision–language retrieval versus semantic provenance. CLIP aligns images and text in a shared space and is a natural baseline for retrieving descriptions of visible content (Radford et al., 2021). Semantic provenance is different: it asks which text representation was intentionally embedded, which may encode a prompt, task, user class, or workflow and need not equal the visible description. Our payload-shuffling control makes this distinction testable by generating content with prompt p_A while injecting payload s_B. A valid provenance channel should recover s_B; an ordinary image–text retriever is expected to prefer p_A. An ID system augmented with an external semantic database is a relevant extension beyond native lookup; Section 5 distinguishes that untested comparator from the recorded controls.

3 Problem Formulation: Identification versus Retrieval

Let p control visible generation, s denote the embedded provenance payload, and A denote a dissemination transformation. The verifier observes only $\widetilde{I}^w = A(I^w)$, along with its decoder and reference bank; the original prompt and undistorted image are not required as query inputs. Separating p and s permits the binding test in Section 5.4.

Discrete detection and identification. Let $\mathcal{M} = \{0,1\}^b$ or a finite registered ID set. A conventional interface tests watermark presence or recovers an identity, possibly using soft scores and error correction internally:

$$h(\widetilde{I}^w) \in \{0,1\}, \qquad \hat{m} = d(\widetilde{I}^w) \in \mathcal{M} \cup \{\perp\}, \qquad \mathrm{match}(\hat{m}, m_j) = \mathbf{1}[\hat{m} = m_j]. \tag{1}$$

Here $\perp$ denotes rejection or decoding failure. An external lookup maps a recovered m to its source record when that mapping is available. The native output is a discrete identity: its Hamming or key distance does not inherently rank unseen source descriptions by semantic relevance. This is a limitation of the attribution interface, not a claim that discrete watermark detection requires the original prompt or lacks distortion tolerance.

Continuous semantic retrieval. SRW encodes payload text using a normalized semantic mapping $e_\theta(s) \in \mathbb{S}^{d-1}$ and stores $\mathcal{B} = \{(s_j, \bar{z}_j)\}_{j=1}^N$, with $\bar{z}_j = e_\theta(s_j)$. It decodes and ranks candidates as

$$\bar{\hat{z}} = \frac{D_\psi(\widetilde{I}^w)}{\|D_\psi(\widetilde{I}^w)\|_2}, \qquad q_j = \bar{\hat{z}}^\top \bar{z}_j, \qquad \pi = \mathrm{argsort}_{j\in\{1,\dots,N\}}(-q_j). \tag{2}$$

Thus search uses cosine geometry in a continuous embedding space, rather than equality in $\mathcal{M}$. We use "semantic geometry" to describe meaningful neighborhoods in that space; no smooth-manifold assumption is needed. The output is the ordered top-K list $(s_{\pi_1}, \dots, s_{\pi_K})$ when $\max_j q_j \geq \eta$, and $\perp$ otherwise.

Exact attribution, semantic tracing, and rejection. Write $\mathcal{S}_\mathcal{B} = \{s_j\}_{j=1}^N$. If $s \in \mathcal{S}_\mathcal{B}$, exact retrieval succeeds when its record appears in the top K. If $s \notin \mathcal{S}_\mathcal{B}$, a relevance set $\mathcal{R}(s) \subseteq \mathcal{S}_\mathcal{B}$ must be defined independently of the predicted ranking. Semantic tracing succeeds when the top-K list intersects $\mathcal{R}(s)$. A paraphrase may preserve source meaning; a broader semantic neighbor need not preserve source identity. For queries with no valid match, the target is rejection. A threshold selected on held-out negatives aims to control

$$\mathrm{FPR}(\eta) = \Pr_{I^- \sim \mathcal{D}_-}\left[\max_j q_j(I^-) \geq \eta\right] \leq \alpha, \tag{3}$$

where $\mathcal{D}_-$ is the calibration distribution of negative queries and α is a target false-positive rate. This is a calibration objective, not a guarantee under bank or distribution shift. Ranking metrics and rejection metrics are reported separately.

Why a neighborhood can tolerate recovery error. For unit vectors, if $\|\bar{\hat{z}} - e_\theta(s)\|_2 \leq \epsilon$, each candidate's cosine score changes by at most ϵ. Consequently, a relevant candidate remains ahead of every irrelevant candidate if its ideal score margin exceeds 2ϵ. This conditional bound explains the opportunity: useful ranking need not require exact vector recovery, and relevance can survive a change of text. It does not establish that any particular attack satisfies the bound; the image channel and the semantic neighborhood require empirical evaluation.

4 METHOD

SRW implements a trainable channel from payload text to diffusion latent and back to semantic space. Figure 1 summarizes the four modules: semantic selection, latent injection, image-side decoding, and retrieval with calibration.

4.1 ATTENTION-GUIDED SEMANTIC PAYLOAD

Payload descriptions contain function words and stylistic modifiers that compete for a finite watermark budget. Given tokens $t = (t_1, \dots, t_L)$, a frozen text encoder E_{text} produces contextual embeddings

$$H = E_{\text{text}}(t) = [h_1; \dots; h_L], \qquad h_i \in \mathbb{R}^d. \tag{4}$$

A learned readout scores each token, $a_i = f_\theta(h_i)$, and normalizes the scores as

$$\alpha_i = \frac{\exp(a_i)}{\sum_{k=1}^L \exp(a_k)}. \tag{5}$$

We retain either the k_{tok} largest weights or all weights above a threshold τ. For selected indices $\mathcal{S}$, the payload representation is

$$z = \frac{\sum_{i\in\mathcal{S}} \alpha_i h_i}{\sum_{i\in\mathcal{S}} \alpha_i}, \qquad \bar{z} = \frac{z}{\|z\|_2}. \tag{6}$$

Dividing by the selected attention mass removes its effect on the pooled scale; the vector still depends on which tokens are retained. This pooled and normalized mapping defines $e_\theta(s) = \bar{z}$ in Section 3. Selection allocates the finite carrier budget to payload content; its benefit over unselected pooling requires a controlled ablation.

4.2 BOUNDED LATENT INJECTION

Pixel-space perturbations are only loosely coupled to the diffusion trajectory. SRW instead maps $\bar{z}$ to a perturbation with the shape of the generator latent:

$$\Delta x = g_\phi(\bar{z}), \qquad x^w = \mathrm{clip}(x + \lambda \Delta x, -\gamma, \gamma), \tag{7}$$

where λ controls strength and γ bounds the latent range. The generator then produces

$$I^w = G(x^w; p). \tag{8}$$

Using separate arguments for p and s is important: p determines visible content, whereas s determines the injected provenance direction. Their equality is an application choice, not a model assumption.

4.3 RECOVERY UNDER DISSEMINATION

Training exposes the image channel to a sampled transformation A that represents resizing, JPEG compression, cropping, blur, brightness variation, and additive or impulse noise. An image-side decoder predicts

$$\hat{z} = D_\psi(A(I^w)), \qquad \bar{\hat{z}} = \frac{\hat{z}}{\|\hat{z}\|_2}. \tag{9}$$

Attack simulation trains the decoder to recover the payload from the published image. At deployment, decoding uses that image; retrieval additionally requires the trained semantic encoder and candidate bank.

4.4 RETRIEVAL, CALIBRATION, AND LEARNING

Reference payloads are encoded with the same selector used for injection. The top-ranked output of Equation 2 is

$$q_j = \bar{\hat{z}}^\top \bar{z}_j, \qquad \hat{s} = \arg\max_{s_j \in \mathcal{B}} q_j. \tag{10}$$

The acceptance rule in Section 3 couples ranking to calibrated rejection. Brute-force search costs $O(Nd)$ time and $O(Nd)$ bank memory; approximate nearest-neighbor indexing is compatible but is not needed for the reported 10K bank.

The watermark path is optimized by regression and contrastive retrieval. With positive z^+, negatives $\{z_m^-\}_{m=1}^M$, and temperature T,

$$\mathcal{L}_{\mathrm{reg}} = \|\hat{z} - z\|_2^2, \tag{11}$$

$$\mathcal{L}_{\mathrm{nce}} = -\log \frac{\exp(\mathrm{sim}(\hat{z}, z^+)/T)}{\exp(\mathrm{sim}(\hat{z}, z^+)/T) + \sum_{m=1}^M \exp(\mathrm{sim}(\hat{z}, z_m^-)/T)}, \tag{12}$$

$$\mathcal{L} = \beta \mathcal{L}_{\mathrm{reg}} + (1 - \beta)\mathcal{L}_{\mathrm{nce}}. \tag{13}$$

The text encoder and image generator can remain frozen; gradients update the semantic readout, injector, and image decoder. Strength λ and clipping γ expose the fidelity–retrievability operating point rather than concealing it inside the loss.

5 EXPERIMENTS

The primary question is whether a distorted image retains a recoverable semantic watermark. We first test retrieval after dissemination, then test the additional functionality of semantic search when the source text is absent, and finally examine whether the recovered signal follows the hidden payload. Module ablations and operating-point diagnostics are treated separately from these effectiveness questions.

5.1 SETUP AND PROTOCOL

Backbone and records. The main operating point uses a Stable Diffusion 1.5 latent backbone (Rombach et al., 2022). The recorded payload sources are COCO captions (Chen et al., 2015), DiffusionDB (Wang et al., 2022), PartiPrompts (Yu et al., 2022), DrawBench (Saharia et al., 2022),

LAION-Aesthetics (Schuhmann et al., 2022), Midjourney-style prompts, an art-domain set, and a photo-realistic set. Payload texts also populate the candidate bank. Results are preserved aggregate point estimates; query-level predictions, counts, and seed-level uncertainty are unavailable. The attack sweep, source-absence tables, and bank-size sweep are distinct records. We do not infer a common bank size or a joint distortion–source-absence protocol where these are undocumented.

Tasks and metrics. Exact retrieval uses Top-1, Top-5, and mean reciprocal rank (MRR). Semantic-neighbor retrieval excludes the exact source payload; a successful result must belong to its declared relevance set. The original source-absence tables use "prompt" because they evaluate prompt payloads ($s = p$). We preserve their Recall@K, MRR, AUROC, and FPR@95TPR (FPR95), while noting that the precise relevance-label and negative-pool construction needs recovery from the logs. A source-absent query can have a relevant semantic neighbor and is not automatically a rejection negative. PSNR and SSIM compare clean and watermarked generations. Thresholded acceptance is analyzed separately from unthresholded ranking.

Baselines. CLIP image–text retrieval uses visible image semantics (Radford et al., 2021); HiDDeN and RingID expose bit and identity signals (Zhu et al., 2018; Ci et al., 2024). Their recorded task-adapted outputs evaluate downstream provenance utility, not equivalent native tasks. Exact decoding and ID-to-record lookup can remain useful under distortion. The present comparison does not include an ID watermark augmented with semantic search over a retained metadata database; therefore it isolates native lookup limitations rather than all possible extensions of ID systems. Baseline adapters and matched-image/bank protocols require verification from the raw records.

5.2 Main result I: Recovering provenance after dissemination

SRW recovers provenance from the residual watermark after common image transformations. In the severity-controlled sweep, retrieval accuracy is 94% after JPEG quality 50, 95% after $0.5\times$ resizing, 94% after removing 20% of the image, and 94% for Gaussian blur with kernel 5. Gaussian noise at $\sigma = 0.03$ and the recorded brightness range 0.7–1.3 both retain 92%. These measurements directly address the dissemination challenge: extraction need not reconstruct every part of the original carrier to recover a useful retrieval vector. Each attack family is evaluated separately; the records do not demonstrate a sequential composition of all edits.

Table 1: Primary downstream provenance comparison. Values are recorded Top-1/retrieval rates (higher is better). JPEG, resize, and editing denote the conditions in the baseline table; their exact settings are not specified there and must not be inferred from matching sweep values. Native signal types differ, and adapter details remain a reporting gap.

Method	Recovered signal	Clean	JPEG	Resize	Editing
CLIP	Visible semantics	0.86	0.72	0.75	0.55
HiDDeN	Bits	0.90	0.80	0.82	0.60
RingID	Identity	0.98	0.90	0.92	0.65
SRW	**Semantic payload**	**1.00**	**0.94**	**0.95**	**0.72**

Table 1 supplies a separate downstream comparison. SRW reaches 94%, 95%, and 72% after its recorded JPEG, resize, and editing conditions, compared with RingID's 90%, 92%, and 65%. The bit/ID baselines also retain substantial retrieval, so these results do not imply that discrete watermarking universally fails under distortion. They establish a functioning semantic carrier; the additional advantage of its representation must be assessed when exact source lookup is unavailable (Section 5.3).

Robustness has a clear boundary. In the full sweep (Appendix C), accuracy falls to 76% at JPEG quality 10, 75% after a 50% crop, 65% for Gaussian noise $\sigma = 0.10$, and 12% for salt-and-pepper density 0.10. The clean operating point has PSNR 25.64 dB and SSIM 0.8152, indicating a nontrivial image-quality cost. The evidence therefore supports recovery under the recorded moderate distortions, rather than arbitrary editing or imperceptibility.

5.3 MAIN RESULT II: SEMANTIC TRACING WITHOUT THE EXACT SOURCE TEXT

The source-absence evaluation tests the capability that native discrete lookup lacks: ordering alternative provenance descriptions by meaning. Table 2 excludes the exact source text. SRW reaches 93% Recall@1 for paraphrases and 82% for hard semantic-near queries, with Recall@5 of 99% and 94%, respectively. Cross-domain queries retain 88% Recall@1. These results support semantic tracing when an equivalent or relevant description remains in the bank; they do not reconstruct deleted text or establish the identity of a missing record.

Table 2: Open-set semantic-neighbor retrieval. The exact source prompt is excluded; higher is better except FPR@95TPR.

Unknown type	R@1	R@5	MRR	AUROC	FPR95
Random	0.90	0.98	0.94	**0.97**	**0.04**
Semantic-near	0.82	0.94	0.87	0.90	0.09
Paraphrase	**0.93**	**0.99**	**0.96**	0.93	0.06
Cross-domain	0.88	0.97	0.92	0.95	0.05

Table 3: Semantic provenance retrieval without the exact source text. Recall@1 (higher is better) compares the recorded task-adapted bit/ID lookup controls with SRW. These values measure semantic tracing, not native watermark detection.

Source-absence category	HiDDeN	RingID	SRW
Random	0.012	0.021	**0.900**
Semantic-near	0.018	0.031	**0.820**
Paraphrase	0.014	0.025	**0.930**
Cross-domain	0.010	0.019	**0.880**
Macro-average	0.0135	0.0240	**0.8825**

Averaged over the four source-absence categories, the recorded Recall@1 is 1.35% for HiDDeN, 2.40% for RingID, and 88.25% for SRW (Table 3; full values in Appendix B). The difference matches the interface distinction: bit/ID lookup supplies no native ordering of unseen descriptions, while SRW ranks the semantic payload it was trained to recover. These are task-adapted lookup controls, not evidence of poor native watermark detection. A stronger ID-plus-semantic-database baseline remains necessary to isolate the benefit of carrying semantics inside the watermark.

The table also preserves category-specific AUROC values of 0.90–0.97 and FPR95 values of 0.04–0.09. These recognition summaries depend on the negative-pool definition and must not be treated as the false-positive rate of a deployed bank. Most importantly, source absence and image distortion are separate evaluation axes in the available records. Their conjunction remains untested, so the two main results support complementary capabilities without proving their joint success.

5.4 MAIN CONTROL: HIDDEN PAYLOAD VERSUS VISIBLE IMAGE CONTENT

Payload binding distinguishes semantic watermark recovery from ordinary image–text retrieval. The shuffling control generates visible content with p_A while injecting a different payload s_B. Its recorded summary gives

$$\mathrm{sim}(\hat{z}, z_B) > \mathrm{sim}(\hat{z}, z_A), \tag{14}$$

where z_A encodes the visible description and z_B encodes the injected metadata; the CLIP baseline instead prefers the visible description. This direction supports a channel tied to the intended payload even when $s \neq p$. The clean CLIP–SRW accuracy difference alone cannot establish binding. Because per-example shuffling scores were not retained, neither a binding success rate nor its uncertainty is claimed.

5.5 ABLATION STUDY: MODULES AND HYPERPARAMETERS

Module attribution is distinct from end-to-end effectiveness. Attention selection, latent injection, and the regression/contrastive losses require matched remove-or-replace comparisons; token budget

k_{tok}, selection threshold τ, injection strength λ, clipping bound γ, loss weight β, and temperature T require controlled sensitivity sweeps. The supplied package does not contain quantitative tables for these comparisons. Consequently, the reported results establish the full system's recorded behavior but do not establish each module's individual gain or an optimal hyperparameter choice. Appendix E records the missing comparisons explicitly. Candidate-bank scaling below is a diagnostic of retrieval difficulty, not a module ablation.

5.6 Diagnostic analysis: Bank size, ranking, and calibration

Candidate-bank density and Top-K degradation. The bank-size diagnostic in Appendix F (Table 6; Figure 2a) varies retrieval difficulty without changing the task. Top-1 decreases from 98% at 1K to 92% at 10K, while Top-5 decreases from 99.5% to 97.5%. Thus the Top-1/Top-5 gap grows from 1.5 to 5.5 percentage points, consistent with more rank ambiguity in a denser bank. The recorded average positive–negative margin declines from 0.61 to 0.46 over the same range; this aggregate trend is consistent with, but does not prove, the recovery-error argument in Section 3. Latency grows from 8 ms at 100 candidates to 90 ms at 10K. Missing hardware metadata limits this to within-record scaling rather than a transferable runtime benchmark.

Threshold calibration and fidelity. At the recorded threshold $\eta = 0.35$, recall is 0.6857, FPR is 0.0015, and balanced accuracy is 0.8421. This operating point trades recall for fewer false acceptances; it is distinct from the unthresholded retrieval rates above. A single point is not a threshold-sensitivity curve, and its transfer to another bank or distribution is unestablished. Likewise, the clean fidelity values in Section 5.2 document one operating point, not a measured fidelity–retrieval frontier. Both diagnostics constrain the practical interpretation of the semantic retrieval formulation.

6 Limitations and Broader Impact

The evidence supports a bounded operating regime. Common moderate transformations preserve retrieval, but strong impulse noise sharply degrades it; the inherited supplementary summaries also identify regeneration and adaptive removal as challenges. Fidelity at the recorded clean operating point (25.64 dB PSNR; 0.8152 SSIM) has a nontrivial cost. Source-absent retrieval under simultaneous dissemination distortion, quantitative module ablations, and an ID watermark with external semantic search remain untested in the supplied records. The 10K bank is below platform scale, and the aggregate tables lack query counts, raw predictions, seed-level variance, baseline-adapter details, and latency hardware. These gaps limit causal and statistical conclusions and should be resolved from experiments or logs.

Semantic similarity cannot recover arbitrary missing identity information. Paraphrases can preserve meaning, but semantically close workflow or source descriptions need not refer to the same origin; relevance labels and rejection thresholds therefore require care. Payloads may also contain sensitive information, motivating data minimization, controlled retrieval access, and auditing of false acceptances. SRW is not a cryptographic proof of authorship: removal, overwriting, and spoofing remain possible. Enforcement decisions require corroborating provenance evidence.

7 Conclusion

SRW moves the watermark interface from discrete identity lookup to semantic provenance retrieval. A trained latent-to-image channel enables recovery after the recorded moderate dissemination transformations; a continuous payload space additionally enables tracing to relevant descriptions when the exact source text is absent. The corresponding evaluations retain 94–95% retrieval under representative JPEG, resize, and crop settings and 82–93% Recall@1 across source-absence categories. Their joint setting remains to be tested. The central contribution is a recoverable representation whose neighborhoods carry source meaning, extending the questions watermarking can answer beyond presence and registered identity. Stronger robustness, lower fidelity cost, and fully specified comparative evaluation remain necessary next steps.

AI USE STATEMENT

Generative AI tools assisted manuscript organization and language editing, plotting and LaTeX code, and consistency checks. They did not generate the reported raw measurements. This revision checked values against the supplied manuscript and CSV tables; original log and protocol gaps remain disclosed. The authors are responsible for verifying the final manuscript and artifacts.

ETHICS STATEMENT

Semantic provenance may support accountability but can reveal sensitive metadata, and false attribution can cause harm. Deployment requires data minimization, controlled access, and error auditing. Watermarks should not serve as sole proof of authorship or misconduct.

REPRODUCIBILITY STATEMENT

The package includes the method, aggregate CSV tables, full attack sweep, and figure scripts. Independent reproduction additionally requires the missing per-query predictions, query counts, seeds, baseline-adapter details, training configuration, and latency hardware described in the appendix and author checklist.

REFERENCES

Mauro Barni, Franco Bartolini, and Alessandro Piva. A dct-domain system for robust image watermarking. *Signal Processing*, 66(3):357–372, 1998.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. *arXiv preprint arXiv:2404.14055*, 2024.

Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding—a survey. *Proceedings of the IEEE*, 87(7):1062–1078, 1999.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 35, 2022.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117–2126, 2020.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision*, pp. 657–672, 2018.

A Evaluation record and reporting scope

The supplied LaTeX package records an earlier reconciliation of a main paper, a supplementary-analysis PDF, and an expanded experiment workbook, with the workbook prioritized for conflicting values. This revision has access to that package and its compiled manuscript; it preserves the packaged CSV measurements and does not independently re-verify the unprovided workbook or raw experiment logs. The workbook provides aggregate tables, not per-query outputs; therefore sample counts, confidence intervals, and significance tests are intentionally not fabricated. The candidate-bank latency values also lack hardware metadata. These fields should be restored from the original experiment logs for a final archival version.

B Task-adapted open-set comparison

Table 4 extends the main semantic comparison in Table 3 with Recall@5 and MRR. The exact source prompt is absent. The category label "Random" is inherited from the aggregate table and is not evidence that these queries have empty semantic relevance sets; the detailed category and negative-pool definitions remain to be recovered. A result is correct only when the returned candidate belongs to the predefined semantically relevant set. HiDDeN and RingID expose discrete signals and have no native operation that orders unseen prompts semantically; their task-adapted numbers should therefore be interpreted as lookup controls, not as a replacement for their native detection or identification evaluation. The adapter implementation and matched candidate-bank construction are not specified in the available record. Low recall alone cannot be called "chance-level" without the bank size and relevance-set cardinalities. No ID-plus-semantic-database comparator is available.

C Full robustness sweep

Table 5 gives every severity point in the expanded workbook. The monotonic curves support the main-text claim for moderate dissemination transformations and expose strong impulse noise as the dominant reported failure.

D Additional supplied analyses

The inherited package summarizes five additional diagnostic families from a supplementary PDF that is not included in the present inputs. We retain these as reported qualitative directions, without claiming fresh verification or unavailable numerical values.

Controls and calibration. Non-watermarked images occupy lower maximum-similarity regions than correctly watermarked images; shuffled payloads lie between them. This motivates threshold calibration but does not eliminate the need to recalibrate under new generators or data distributions.

Table 4: Full open-set comparison under the task-adapted semantic relevance protocol.

Unknown type	Method	Recall@1 ↑	Recall@5 ↑	MRR ↑
Random	HiDDeN	0.012	0.041	0.024
	RingID	0.021	0.063	0.036
	SRW	**0.900**	**0.980**	**0.940**
Semantic-near	HiDDeN	0.018	0.052	0.030
	RingID	0.031	0.081	0.047
	SRW	**0.820**	**0.940**	**0.870**
Paraphrase	HiDDeN	0.014	0.045	0.026
	RingID	0.025	0.072	0.041
	SRW	**0.930**	**0.990**	**0.960**
Cross-domain	HiDDeN	0.010	0.037	0.022
	RingID	0.019	0.058	0.034
	SRW	**0.880**	**0.970**	**0.920**

Table 5: Retrieval accuracy under the full recorded post-processing sweep.

Attack	Severity	Accuracy	Attack	Severity	Accuracy
Clean	None	1.00	Crop	30% removed	0.90
JPEG	Quality 90	0.99	Crop	40% removed	0.84
JPEG	Quality 70	0.97	Crop	50% removed	0.75
JPEG	Quality 50	0.94	Gaussian blur	Kernel 3	0.97
JPEG	Quality 30	0.89	Gaussian blur	Kernel 5	0.94
JPEG	Quality 10	0.76	Gaussian blur	Kernel 7	0.90
Resize	$0.75\times$	0.98	Gaussian blur	Kernel 9	0.84
Resize	$0.50\times$	0.95	Gaussian blur	Kernel 11	0.77
Resize	$0.35\times$	0.90	Gaussian noise	$\sigma = 0.01$	0.97
Resize	$0.25\times$	0.83	Gaussian noise	$\sigma = 0.03$	0.92
Crop	10% removed	0.97	Gaussian noise	$\sigma = 0.05$	0.85
Crop	20% removed	0.94	Gaussian noise	$\sigma = 0.08$	0.74
Gaussian noise	$\sigma = 0.10$	0.65	Salt–pepper	Density 0.01	0.88
Salt–pepper	Density 0.03	0.58	Salt–pepper	Density 0.05	0.35
Salt–pepper	Density 0.08	0.20	Salt–pepper	Density 0.10	0.12
Brightness	0.7–1.3	0.92			

Payload binding and injector diversity. Under p_A/s_B shuffling, the recovered feature is closer to the injected payload s_B than to visible prompt p_A. A cross-prompt perturbation-similarity matrix has low off-diagonal similarity, arguing against collapse to one universal latent perturbation.

Backbone and prompt shift. Transfer is strongest on the training or adapted Stable Diffusion backbone and declines for more distant latent spaces and sampling mechanisms. Prompt retrieval also becomes harder as length and the number of semantic concepts increase. These observations bound, rather than prove, model-agnostic generalization.

Strong attacks and editing. Diffusion regeneration, high-strength image-to-image editing, adaptive white-box perturbations, and spoofing-oriented attacks cause larger degradation than ordinary post-processing. The current paper therefore claims robustness to common dissemination transformations, not to unrestricted adaptive removal.

Local tamper localization. A patch-level inconsistency map was explored as an auxiliary diagnostic. Because local integrity is a distinct task and the supplied record contains no standardized localization metrics, we do not include it among the main contributions.

E ABLATION AND JOINT-EVALUATION SCOPE

The available CSV files cover attacks, baseline comparisons, source absence, and bank-size scaling. They contain no quantitative module-removal or hyperparameter-sensitivity results. The following comparisons define the outstanding ablation scope, rather than completed experiments: attention-guided selection versus full-token and mean pooling; the full loss versus regression-only and contrastive-only training; latent injection versus a matched alternative carrier; and controlled sweeps of $k_{\text{tok}}, \tau, \lambda, \gamma, \beta$, and T. Comparisons should hold data splits and candidate banks fixed and report retrieval alongside fidelity at comparable operating points. No ablation gain is inferred from the full-model aggregates.

The key outstanding effectiveness test crosses candidate availability (exact source present, paraphrase only, semantic neighbor only, and no relevant candidate) with image condition (clean and each specified distortion). It should report retrieval and rejection separately, preserve relevance labels, and include bit/ID lookup, ID plus external semantic search, CLIP, and SRW with documented adapters. Quantitative p_A/s_B binding under the same conditions is needed to exclude visible-content shortcuts. These tests would determine whether the separately observed dissemination robustness and semantic tracing benefits hold jointly.

F CANDIDATE-BANK AND SEVERITY DIAGNOSTICS

These diagnostics supplement the main effectiveness results. Bank size changes ranking difficulty, while the separate attack sweeps expose severity-dependent failure. The two panels are not a joint bank-size-by-attack evaluation.

Table 6: Exact provenance retrieval as the candidate bank grows. All metrics except latency are unitless; higher is better except latency.

Bank size	Top-1 ↑	Top-5 ↑	MRR ↑	Margin ↑	Latency (ms) ↓
100	**1.000**	**1.000**	**1.000**	**0.82**	**8**
240	**1.000**	**1.000**	**1.000**	0.73	12
500	0.990	**1.000**	0.995	0.68	18
1,000	0.980	0.995	0.985	0.61	28
5,000	0.950	0.985	0.960	0.52	55
10,000	0.920	0.975	0.940	0.46	90

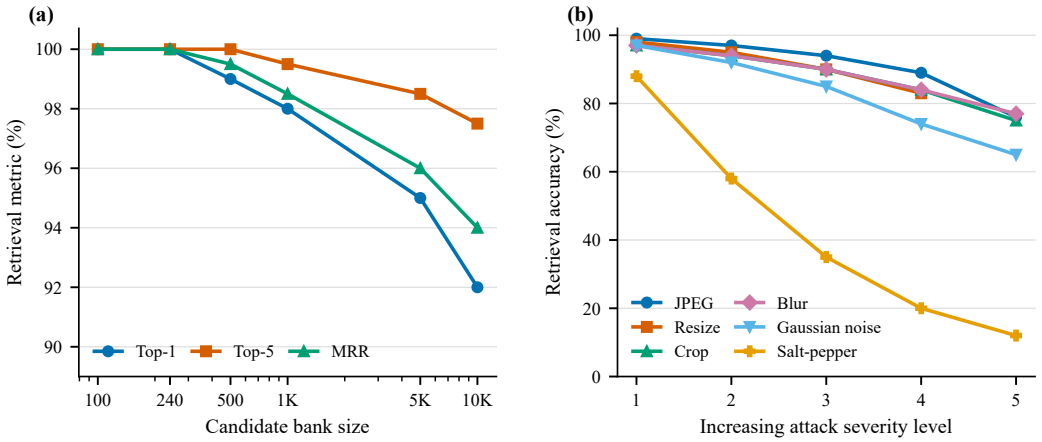

Figure 2: Diagnostics of the operating regime. (a) Candidate-bank scaling and Top-K degradation. (b) Attack-severity curves supplement the primary dissemination result in Section 5.2; strong impulse noise is a failure mode. Severity indices are attack-specific and cannot be compared across families. Appendix C gives exact settings. Brightness has only one recorded setting and is omitted from the curves. Counts and uncertainty were not retained.