Algebraic and Lineage-Preserving Watermarking in Generative AI

Pre-requisites. Probability theory, algebra, information theory (mutual information, capacity), and basic Machine Learning.

Figure 1: Left: ordinary image. Right: same image with an invisible watermark.

Background. Every day, more of what we see, read, and share is touched by AI. Articles, images, and even videos are now being generated and remixed by powerful models. With this flood of synthetic content, how can anyone know what is "real"? Watermarking was supposed to be the answer, where tiny, hidden signals were placed in AI outputs to prove their origin. In mathematical style, a watermark is an imperceptible signal embedded in content x (text, image, audio, video) that remains statistically/algorithmically detectable later. In its simplest form,

$$\tilde{x} \;=\; E_k(x, W) \quad \text{and} \quad \text{Detect}(\tilde{x}; k) \in \{0, 1\},$$

where E_k is a key-driven embedder and detection is a hypothesis test (Neyman–Pearson style). But most current watermarks can be stripped away with a single paraphrase, a translation, or by running the text through another model. An image can lose its watermark with a simple crop, filter, or re-rendering. And as content passes through hundreds of unknown models, the watermark traces vanish completely. Recent reviews explain what a good stamp should do (stay hidden, survive ordinary editing, and be checkable) and how people try to break it [1, 2]. Another result further shows that small character-level perturbations can defeat state-of-the-art LLM watermarks under realistic detectors, underscoring the need for principled robustness [4]. In everyday use, AI tools and apps constantly rewrite and touch up content, such as people shortening or rephrasing text, cropping or lightly editing pictures, or passing them through other programs. These common tweaks can already wash out many current stamps, which is why we need more straightforward rules and stronger designs [3].

Problem (gap). Today, we lack a clean *rule of composition* for watermarks under sequences of edits. In practice, content often goes through many steps (summarize → translate → stylize

→ crop). Existing schemes rarely *carry lineage*: they do not let us say "who touched this and in what order" with an explicit detection guarantee. Most importantly, we do *not* have a general theorem that tells us for which transformations a watermark can be made to "travel" predictably.

The Challenge. Design a *cryptographically verifiable, lineage-preserving watermark* that:

1. **Composes algebraically** under transformations $T \in \mathcal{T}$ (translation, paraphrase, filtering, crop/resize, in-/out-painting, diffusion resampling), i.e.

$$W(T(x)) \;=\; f_T\big(W(x)\big), \qquad f_{T_2 \circ T_1} \;=\; f_{T_2} \circ f_{T_1}.$$

2. **Propagates probabilistically** through stochastic generators with quantitative detection guarantees (e.g., lower bounds on $I(W; Y_k)$ after k edits).
3. **Admits verification proofs** (black-/white-box) consistent with reliability/integrity requirements [1].

Important open point. We currently *do not have a general formula/theorem* characterizing the class of content/model transformations that preserve lineage-consistent watermark invariants (for text and images). Closing this gap is the central theoretical aim.

Research questions (explicit).

Q1. *What class of transformations preserves watermark invariants?*
Characterize a semigroup $\mathcal{T}$ of edits T and endomorphisms $f_T : \mathcal{W} \to \mathcal{W}$ so lineage is conserved ($f_{T_2 \circ T_1} = f_{T_2} \circ f_{T_1}$). Develop invariants/subspaces and impossibility boundaries.

Q2. *Can we model paraphrasing or image filtering as linear operators over a watermark subspace?*
For text: linear maps on token-frequency/embedding histograms; for images: near-isometries bands. Prove preservation up to bounded distortion and identify counterexamples.

Q3. *How much entropy (in bits) can a watermark retain after k stochastic edits?*
Derive information-theoretic lower bounds $I(W; Y_k) \geq \underline{I}(\sigma, \mathrm{SNR}, k)$ under paraphrase noise/diffusion schedules; obtain sample-complexity and false-alarm guarantees for optimal tests.

Q4. *Can a lineage chain be represented as a homomorphic-encryption proof that is verifiable yet privacy-preserving? (Over-ambitious)*
Define a *Watermark Lineage Proof*: commitments $C_i = \mathrm{Com}(w_i; r_i)$ and succinct proofs π_i that $w_{i+1} = f_{T_i}(w_i) \oplus \phi_i$ (bounded slack), aggregatable across the chain.

Context. This work will run within the Cyber Threat Intelligence Lab of the School of Computer Science and Engineering (CSE) and the School of Mathematics and Statistics. The project CIs are Prof Scott Sisson, A/Prof Arash Shaghaghi, and Miss Meghali Nandi [PhD Candidate]. The math student will own theorems/bounds (algebra of transforms; MI/detection exponents; verification soundness) while the CS/AI team provides implementations and attack harnesses, aligned with surveyed requirements and modern LLM/diffusion settings [1, 2]. The project is expected to lead to a research paper in top journals. The project is of interest to CTI Lab's industry partners and may lead to follow-up paid opportunities.

References

[1] F. Boenisch, *A Systematic Review on Model Watermarking for Neural Networks*, arXiv:2009.12153v2 (2021).

[2] H. K. Singh and A. K. Singh, *Comprehensive review of watermarking techniques in deep-learning environments*, Journal of Electronic Imaging **32**(3):031804 (2023).

[3] Y. Wen, J. Kirchenbauer, J. Geiping, T. Goldstein, *Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust*, arXiv:2305.20030 (2023).

[4] Z. Zhang, X. Zhang, Y. Zhang, H. Zhang, S. Pan, B. Liu, A. Q. Gill, L. Y. Zhang, *Character-Level Perturbations Disrupt LLM Watermarks*, arXiv:2509.09112 (accepted NDSS 2026).