CONTINUOUS SYNDROME FORENSICS FOR PROACTIVE IMAGE MANIPULATION LOCALIZATION AND RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The proliferation of generative AI motivates proactive image forensics capable of accurately localizing manipulations and reconstructing visually plausible semantic content. However, existing proactive defenses primarily rely on empirical deep learning, lacking principled algebraic foundations and remaining fragile under global distortions. While classical coding theories offer rigorous paradigms, their direct adaptation to natural images is non-trivial: finite-field error-correction codes are not directly applicable to continuous-valued pixels without discretization, and tools like classical compressive sensing commonly require strict global sparsity that fails under large-area edits, alongside dense linear projections that degrade visual quality. To bridge this gap, we propose Continuous Syndrome Forensics (CSF), which extends syndrome-based verification to the continuous spatial domain. By imposing block-wise algebraic invariants, CSF introduces an algebraically structured formulation for tamper localization without global sparse-recovery optimization. Although it is effective for small edits, directly applying CSF to natural images suffers from host self-interference and remains underdetermined under large-area manipulations. To address these challenges, we further introduce SyndromeMark, a network-augmented framework. SyndromeMark leverages the continuous syndrome residual proposed in CSF as an algebraically anchored spatial prior to guide both localization and reconstruction. Specifically, this prior enables a localization network to separate tampering from host-texture leakage, while guiding a reconstruction network to synthesize contextually plausible content consistent with the algebraic constraint. Extensive evaluations show that SyndromeMark improves robustness over the evaluated proactive baselines under severe post-processing distortions, supports both localization and semantic reconstruction, and remains resilient in the adaptive surrogate setting.

1 INTRODUCTION

The rapid proliferation of generative AI has lowered the barrier for realistic image manipulation (Nightingale & Farid, 2022; Frank et al., 2024) and led to a severe trust crisis in digital media (Chesney & Citron, 2019; Fallis, 2021). For example, recent sophisticated deepfake attacks have already bypassed corporate visual authentication to cause a $25 million financial loss (Chen & Magramo, 2024). In critical domains where digital images serve as the primary source of evidence, adversaries are increasingly capable of deploying localized semantic tampering to subvert sensitive decision processes, such as injecting false lesions into medical scans or fabricating legal evidence (Mirsky et al., 2019).

To counter this threat, traditional passive forensics primarily focus on detecting and localizing manipulated content by analyzing abnormal visual traces (Kee et al., 2013; Huang et al., 2025; Sheng et al., 2026; Huang et al., 2026), local noise inconsistencies (Lyu et al., 2014; Dong et al., 2022; Guo et al., 2023; Li et al., 2025), and resolution discrepancies (Lin et al., 2009; Tan et al., 2024; ALBarqawi et al., 2026). However, these passive methods rely heavily on prior knowledge and easily fail when forensic traces are obscured by subtle editing or unpredictable post-processing like JPEG compression. Furthermore, even when detection is successful, this paradigm provides an

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.

alarm without a rescue, often leading to the discard of the digital asset and the loss of the original information (Chesney & Citron, 2019).

To address this limitation, proactive defenses have emerged to embed extra redundancy directly into host images via deep watermarking, thereby reducing the reliance on natural inconsistencies (Zhang et al., 2024; 2025; Ying et al., 2021; 2023; An et al., 2026b). While this paradigm enables manipulation localization and, in some cases, content reconstruction, existing methods typically rely on neural networks to empirically detect local degradations within the embedded watermark to identify tampered regions, lacking a principled algebraic binding between the visual semantics and the embedded protection. Consequently, when a protected image is subjected to global distortions like Gaussian noise, the widespread degradation of the watermark confuses the network, causing it to misinterpret the global distortion as malicious tampering, triggering catastrophic false positives.

Fortunately, we find that classical coding theories provide mathematical tools that can theoretically address the vulnerabilities of purely empirical proactive defenses. For instance, error correction codes in digital communications (Lin et al., 2001) provide algebraic mechanisms for exact error correction under code-dependent assumptions and error budgets, yet they are not directly suited to raw continuous-valued image pixels without discretization. While real-valued compressive sensing (Donoho, 2006) provides a continuous theoretical framework, its direct adaptation to images faces two critical bottlenecks. 1) First, satisfying its theoretical guarantees (e.g., the Restricted Isometry Property (RIP)) typically necessitates dense projection matrices, which introduce severe global interference and unacceptably degrade the visual fidelity of the host image. 2) Second, it requires a strict global sparsity assumption that can be violated by practical large-area image manipulations.

To bridge this gap, we propose a comprehensive proactive forensic paradigm that synergizes rigorous algebraic constraints with the capacity of deep neural networks. Specifically, our methodology consists of two core components:

Continuous Syndrome Forensics (CSF): By extending syndrome verification from error correction codes to the continuous spatial domain, CSF naturally accommodates real-valued image pixels. Instead of applying dense global matrices typical of compressive sensing, CSF binds the host image to a watermark message through an ***algebraic invariant*** using a sparse block-diagonal semi-orthogonal matrix. During verification, localized manipulations outside the corresponding projection null spaces produce block-local syndrome responses. These responses provide localization cues when they are sufficiently strong relative to embedding mismatch and external distortions. In addition, CSF directly resolves the bottlenecks of compressive sensing: the localized projection preserves visual fidelity, while the distinct energy anomaly pinpoints manipulated regions, bypassing the need for intractable global optimization (such as ℓ_1-minimization) and broad sparsity assumptions for manipulations.

SyndromeMark: We note that directly applying CSF in the physical world encounters practical limitations such as host self-interference, where high-frequency host textures leak into the continuous syndrome residual and obscure the localized energy anomaly of genuine manipulations. Furthermore, reconstructing the missing semantics of large tampered regions remains mathematically underdetermined without visual context. To resolve these challenges, we introduce SyndromeMark, a network-augmented framework. Notably, SyndromeMark does not operate as an independent empirical method. Instead, it utilizes neural networks as tools guided by the underlying mathematical constraints of CSF. Specifically, the framework leverages the computed continuous syndrome residual as an algebraically-anchored spatial prior. For localization, a localization network uses this high-SNR spatial prior to improve discrimination between tampering responses and host-texture leakage. For semantic reconstruction, a reconstruction network is conditioned on both the localization mask and the unmanipulated context to synthesize visually plausible content, improving consistency with the continuous algebraic constraints.

Extensive experiments confirm the indispensable role of the CSF-derived syndrome residual. Trained exclusively on synthetic images, our framework generalizes to unseen natural datasets (e.g., PASCAL VOC (Everingham et al., 2010) and CelebA (Liu et al., 2015)) without fine-tuning. Quantitatively and qualitatively, SyndromeMark achieves precise tamper localization and visually coherent semantic reconstruction across diverse adversarial edits (e.g., splicing, copy-move, and deepfakes) and severe global distortions, where existing state-of-the-art baselines catastrophically fail. Furthermore, robustness against adaptive surrogate attacks validates the resilience of our algebraic invariant.

Ultimately, by synergizing rigorous algebraic constraints with the semantic capacity of deep neural networks, we believe our proposed methods establish a new paradigm for proactive image forensics.

2 Related Work

We provide a brief overview of related literature here. A comprehensive discussion detailing the background and more related work is available in Appendix A.

Proactive Image Forensics. Proactive defenses embed extra redundancy directly into host images via deep watermarking to enable manipulation localization and content reconstruction. However, these approaches exhibit several vulnerabilities. Frameworks focusing solely on localization, including EditGuard (Zhang et al., 2024) and OmniGuard (Zhang et al., 2025), lack reconstruction capabilities and remain vulnerable to common image processing operations like JPEG compression. Conversely, methods capable of content reconstruction, such as Imuge (Ying et al., 2021) and Imuge+ (Ying et al., 2023), are built upon fragile watermarks, causing the subsequent reconstruction process to fail under similar post-processing. While recent frameworks like RecoverMark (An et al., 2026b) attempt to address robustness, they rely on a strong assumption that the image background remains nearly unperturbed, restricting their practical applicability. More importantly, these approaches lack a principled algebraic binding, leaving unconstrained neural networks highly vulnerable to global distortions where they struggle to differentiate malicious structural tampering from global energy-bounded noise.

Error Correction and Compressive Sensing. Our theoretical formulation is inspired by Systematic Linear Block Codes (SLBC) (Lin et al., 2001), a primary category of error correction codes. SLBC embeds a parity-check constraint to compute a syndrome, which effectively isolates sparse errors from the systematic message. However, discrete SLBC treats any continuous pixel variation as a hard bit-flip, making it highly fragile to energy-bounded distortions like JPEG compression. While real-valued compressive sensing (Donoho, 2006) extends error correction to the continuous domain, its generic encoding mechanisms typically require dense projection matrices that severely degrade the visual fidelity of host images. Furthermore, it relies on a strict global sparsity assumption that practical large-area semantic manipulations routinely violate. These theoretical and physical gaps directly motivate our development of Continuous Syndrome Forensics.

3 Problem Formulation and Operational Setting

To establish a mathematical foundation for proactive image forensics, we formulate the manipulation localization and content reconstruction tasks as a continuous real-valued error correction problem. The foundation of this framework relies on embedding a continuous algebraic invariant into the host image, which serves as a deterministic algebraic anchor for subsequent forensic analysis.

3.1 Mathematical Foundation

Let the original host image be flattened into a vector $I_{ori} \in \mathbb{R}^n$. We introduce a continuous syndrome projection matrix $H \in \mathbb{R}^{d\times n}$, where $d < n$. Concurrently, a syndrome generator G defines a deterministic mapping from a watermark message $m \in \{0,1\}^L$ to a continuous constraint vector $G(m) \in \mathbb{R}^d$, where L is independent of d and $G(\cdot)$ remains fixed throughout the entire pipeline.

The proactive protection process modulates the host image to generate a protected image that satisfies a linear algebraic constraint. Without loss of generality, and to maintain consistency with the subsequent watermarking-based ownership verification, we denote the watermarked image (the redundancy-injected version of I_{ori}) as $I_{wm} \in \mathbb{R}^n$, which satisfies:

$$HI_{wm} = G(m). \tag{1}$$

Under adversarial manipulation in the continuous spatial domain, the attacked image is denoted as $I_{atk} = I_{wm} + e$. We explicitly decompose the composite error vector as $e = e_{tamper} + e_{noise}$, where $e_{tamper} \in \mathbb{R}^n$ represents the malicious image tampering and $e_{noise} \in \mathbb{R}^n$ denotes the global energy-bounded signal distortions (e.g., JPEG compression).

As illustrated in Phase 1 of Figure 1, the defender first extracts a predicted message $\hat{m}$ through a watermark decoder to verify ownership. Upon success, the error-free ground-truth m is retrieved

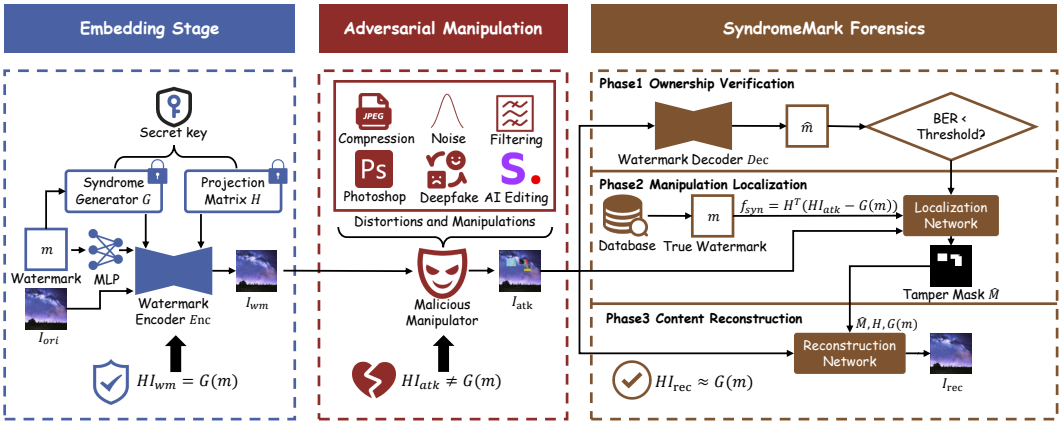


Figure 1: Workflow of SyndromeMark. The embedding stage binds the host image to a continuous algebraic invariant. Upon adversarial manipulation, the framework sequentially executes ownership verification, syndrome-driven localization, and syndrome-guided semantic reconstruction.

from a secure database to construct the reference target $G(m)$. This forensic protocol mitigates the risk of the avalanche effect, where minor bit errors in $\hat{m}$ would cause massive divergence in the target constraint $G(\hat{m})$. By applying the secret projection matrix H to the attacked image alongside the ground-truth m, the defender isolates the continuous syndrome residual $s \in \mathbb{R}^d$:

$$\begin{aligned} s &= HI_{atk} - G(m) \\ &= H(I_{wm} + e_{tamper} + e_{noise}) - G(m) \\ &= He_{tamper} + He_{noise} \approx He_{tamper}, \end{aligned} \tag{2}$$

where the approximation holds due to the physical reality that malicious image manipulations inherently induce massive pixel-level variations, whereas external distortions e_{noise} must remain strictly energy-bounded to preserve image fidelity. Furthermore, as will be shown in Section 4, we structurally design the matrix H as a sparse block-diagonal semi-orthogonal projection, which mathematically guarantees that the absolute energy of any arbitrary signal cannot be amplified during projection. Consequently, the continuous syndrome residual s serves as a robust prior in syndrome space to indicate the tampered regions.

Guided by the foundation, our continuous syndrome forensics framework is formulated to achieve four practical system goals: **1) Invisibility:** The proactive embedding process must not noticeably reduce the visual quality of I_{ori}. **2) Ownership Verification:** The system must robustly extract the message m to claim ownership, even under malicious manipulation or distortion. **3) Robust Localization:** The system must accurately determine the spatial mask corresponding to e_{tamper} by utilizing algebraic syndrome residuals, mitigating false positives caused by global energy-bounded noise. **4) Content Reconstruction:** The system must synthesize visually plausible original content within the manipulated regions such that the reconstructed image I_{rec} aligns with the target algebraic invariant (i.e., $HI_{rec} \approx G(m)$).

3.2 Threat Model

Threat Model. We assume a powerful adversary who possesses full knowledge of the system architecture and training strategies. However, the attacker does not know the exact parameters (weights) of the projection matrix H and the generator G, which act as the system's secret keys and are kept securely by the owner or the scrutinizer. Based on the knowledge of the system mechanisms, the attacker is capable of conducting sophisticated adaptive attacks, such as training proxy surrogate models to approximate the invariant mapping and crafting white-box adversarial perturbations on these surrogates.

The attacker's goal is to modify the protected image I_{wm} to create a forged image $I_{atk} = I_{wm} + e_{tamper} + e_{noise}$. Specifically, the attacker can perform two types of operations:

Localized Tampering (e_{tamper}): Editing specific regions (e.g., object removal, copy-move, and splicing) to alter the semantic integrity or visual facts.

Global Distortion (e_{noise}): Applying image post-processing operations (e.g., JPEG compression, Gaussian blur) to the whole image to obscure the tampering traces.

We exclude trivial Denial-of-Service (DoS) attacks that completely destroy the image (such as cropping 90% of the image or painting it black). While these extreme attacks can easily erase the protection, they also ruin the image itself, making it unusable. Therefore, we assume the forged image still maintains sufficient visual quality to appear realistic.

4 Purely Linear Solution: Continuous Syndrome Forensics (CSF)

4.1 Architectural Design of CSF

With the established algebraic invariant $HI_{wm} = G(m)$, we introduce Continuous Syndrome Forensics (CSF). CSF represents the purely linear track of our forensic paradigm. It realizes the algebraic constraint satisfaction of classical error correction codes within the continuous spatial domain while circumventing the limitations of traditional compressive sensing (Donoho, 2006) to attain mathematical elegance, providing an algebraic basis for the spatial prior used in our framework. To translate this formulation into a practical defense, the construction of the projection matrix H and the embedding process must address three core design challenges:

Joint Optimization. Directly forcing a natural image into a strict algebraic subspace defined by a predefined, randomly initialized matrix H presents a formidable challenge, typically causing unacceptable visual artifacts. To balance algebraic integrity with visual fidelity, we propose a joint optimization strategy where the watermark encoder, the watermark projector (a multi-layer perceptron (MLP)), the watermark decoder, and the projection matrix H are trainable. The binary watermark message m is first projected into a spatial feature map via the projector. This feature map is concatenated with the original image I_{ori} and fed into a deep encoding network (Enc) to generate the watermarked image: $I_{wm} = Enc(I_{ori} \oplus MLP(m))$, where $\oplus$ denotes channel-wise concatenation. To fulfill the ownership verification goal, the watermarked image passes through a differentiable distortion layer (applying random noise, blurring, etc.) before reaching the decoder network (Dec). The decoder extracts the predicted message $\hat{m} = Dec(Distortion(I_{wm}))$. The embedding stage is optimized using a joint loss function:

$$\mathcal{L}_{embed} = \lambda_{fid}\mathcal{L}_{fid}(I_{wm}, I_{ori}) + \lambda_{wm}\mathcal{L}_{wm}(\hat{m}, m) + \lambda_{syn}\|HI_{wm} - G(m)\|_2^2, \tag{3}$$

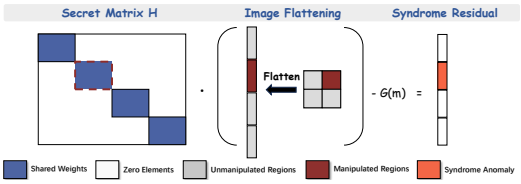
where λ_{fid}, λ_{wm}, and λ_{syn} are corresponding weighting parameters. To make the computational objectives clear, fidelity loss $\mathcal{L}_{fid}$ combines mean squared error (MSE) and LPIPS perceptual loss (Zhang et al., 2018) to encourage visual fidelity; $\mathcal{L}_{wm}$ utilizes binary cross-entropy (BCE) to guarantee robust watermark extraction under distortions; and the syndrome loss directly minimizes the MSE between the projection HI_{wm} and the target $G(m)$ to enforce the continuous algebraic constraint. Once fully trained, the parameters of the optimized semi-orthogonal matrix H and the frozen generator G are securely stored. They are held fixed during subsequent forensic training and inference in both CSF and SyndromeMark.

Spatial Back-Projection via Semi-Orthogonality. When the protected image is attacked, the computed residual $s \approx He_{tamper}$ resides in the lower-dimensional syndrome space $\mathbb{R}^d$. However, our forensic objective is to achieve manipulation localization in the original spatial pixel space $\mathbb{R}^n$. To bridge this dimensional gap, we enforce a semi-orthogonal parameterization constraint on the projection matrix during training, satisfying $HH^\top = I_d$, where I_d is the identity matrix. Rather than relying on soft penalty losses that yield approximation errors, we apply hard orthogonal parameterization to the network weights, operationally implemented via `torch.nn.utils.parametrizations.orthogonal` in PyTorch. Consequently, the transpose $H^\top$ intrinsically acts as the Moore-Penrose pseudo-inverse of H, allowing the defender to compute a back-projected spatial error map $f_{syn} \in \mathbb{R}^n$:

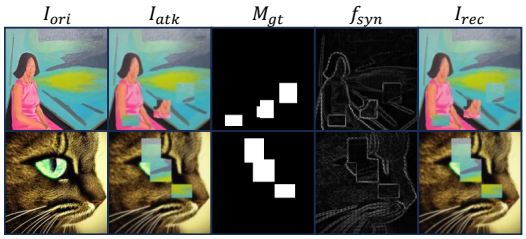
$$f_{syn} = H^\top s \approx H^\top He_{tamper}. \tag{4}$$

The operator $H^\top H$ functions as an orthogonal projector onto the row space of H, yielding the minimum-norm spatial representation of the manipulation errors.

Block-Diagonal Topology for Local Energy Detection. Traditional compressive sensing employs dense global projection matrices and relies on strict global sparsity assumptions to solve for exact



(a) Syndrome Residual Calculation for CSF



(b) Algebraic Curses of CSF

Figure 2: Visualization for theoretical foundation and algebraic curses of CSF: (a) The sparse block-diagonal semi-orthogonal matrix H isolates localized tampering in syndrome residual. (b) High-frequency host textures leak into the spatial syndrome domain (Curse 1), and pure mathematical optimization fails to reconstruct semantic content in underdetermined regions (Curse 2).

pixel-level manipulation. When practical large-area manipulations occur, this global sparsity is violated, causing localized tampering errors to smear across the entire back-projected spatial map. To avoid these issues, CSF adopts a sparse block-diagonal topology. Given the global image dimension n and global syndrome dimension d, we partition the host image into B non-overlapping blocks, where each local block has a dimension of $n_b = n/B$. We employ a shared local linear projection matrix $H_{sub} \in \mathbb{R}^{d_b \times n_b}$ to map each block into a local syndrome of dimension $d_b = d/B$. From a global perspective, the complete projection matrix $H \in \mathbb{R}^{d \times n}$ is constructed as $H = \text{diag}(H_{sub}, H_{sub}, \ldots, H_{sub})$, as visualized in Figure 2a. Operationally implemented as a non-overlapping convolutional layer, this topological structure ensures that computations for different spatial regions remain independent.

Crucially, this design fundamentally shifts the forensic paradigm. Rather than struggling with intractable equations for exact pixel-level localization, CSF leverages this topology for *local residual energy detection*. Because the blocks are isolated, an intact block yields a local syndrome residual with near-zero energy, containing only bounded noise. Conversely, tampering can produce a detectable local syndrome response when its projected contribution is sufficiently large relative to embedding mismatch and external distortions. This block-wise energy mapping not only prevents spatial error smearing but also explains why CSF remains highly effective even under massive, non-sparse tampering.

Formal Theoretical Analysis. Appendix B formalizes four properties of the linear formulation: the absence of cross-block propagation in the projection operator (Proposition 1), a sufficient condition for sparse-error uniqueness and a condition for a nonzero block response (Theorem 1), noise non-amplification and an angle-dependent residual bound (Theorem 2), and the rank condition for unique reconstruction with known unmodified elements under exact constraints (Theorem 3). These results characterize the idealized linear model and do not by themselves establish the localization or reconstruction performance of the learned system discussed in Section 5.

4.2 The Algebraic Curses: Bridging Theory and Reality

While the idealized CSF model provides useful algebraic properties, applying it to natural images introduces two primary algebraic curses, directly motivating our network-augmented SyndromeMark introduced in the next section.

Curse 1: Internal Self-Interference in Localization. In practice, due to the strict visual fidelity requirements, the initial embedding constraint $HI_{wm} = G(m)$ is relaxed into an approximation. This internal residual mismatch contains high-frequency host textures that are leaked into the syndrome domain. Consequently, the f_{syn} map suffers from internal self-interference, making the pure linear filters struggle to effectively distinguish host textures from true tampering (evidenced by the structural host outlines visible in the fourth column of Figure 2b). This physical reality necessitates a non-linear discriminator capable of learning structural differences.

Curse 2: Underdetermined Ambiguity in Reconstruction. Algebraic reconstruction requires the number of tampered pixels within a block to not exceed the local syndrome dimension d_b. Large-

area manipulations easily violate this condition, rendering the local linear system severely underdetermined. Since the resulting infinite null space lacks semantic constraints, purely mathematical solvers (e.g., FISTA (Beck & Teboulle, 2009)) converge to meaningless high-frequency noise rather than visually plausible content (see the I_{rec} column in Figure 2b, which fails to reconstruct).

5 NETWORK-AUGMENTED CSF: SYNDROMEMARK

To resolve the physical curses of CSF, SyndromeMark integrates the algebraic foundation with the empirical semantic capacity of deep neural networks. Rather than operating as a purely data-driven black box that blindly detects visual inconsistencies, SyndromeMark utilizes neural networks as intelligent semantic solvers explicitly guided by the continuous algebraic constraints.

5.1 SYNDROME-DRIVEN LOCALIZATION

As illustrated in Figure 1, to resolve Curse 1, SyndromeMark employs a Localization Network (LNet) to predict the manipulation mask $\hat{M}$. As discussed in Section 3.1 and Theorem 2 in the Appendix B.3, the semi-orthogonal projection prevents noise amplification. Under the assumed magnitude gap between tampering and external distortions, the projected tampering response dominates the projected noise provided that the tampering retains a sufficiently large component in the projection row space. Since the back-projected syndrome f_{syn} inherently isolates this tampering signal, LNet does not need to learn robust features from scratch. Instead, its sole focus is to resolve the remaining internal interference: distinguishing true tampering from host texture leakage.

To achieve this, LNet leverages the algebraically-anchored prior by concatenating the attacked RGB image with the magnitude and energy of the residual:

$$Input = I_{atk} \oplus |f_{syn}| \oplus |f_{syn}|^2. \tag{5}$$

The squared term $|f_{syn}|^2$ highlights severe semantic manipulations while suppressing the low-energy texture leakage. Guided by this augmented input, LNet functions as a spatial discriminator that effectively distinguishes true tampering from the internal self-interference of the host image.

To address the class imbalance typical of sparse manipulations, LNet is supervised by a combined loss:

$$\mathcal{L}_{phase2} = \lambda_{dice}\mathcal{L}_{dice}(\hat{M}, M_{gt}) + \lambda_{focal}\mathcal{L}_{focal}(\hat{M}, M_{gt}), \tag{6}$$

where M_{gt} is the ground-truth mask. $\mathcal{L}_{dice}(\cdot)$ and $\mathcal{L}_{focal}(\cdot)$ denote the commonly used dice loss (Milletari et al., 2016) and focal loss (Lin et al., 2017), weighted by the trade-off hyper-parameters λ_{dice} and λ_{focal}, respectively.

5.2 SYNDROME-GUIDED SEMANTIC RECONSTRUCTION

To resolve the underdetermined ambiguity of Curse 2, SyndromeMark employs a Reconstruction Network (RNet). While pure algebraic solvers yield meaningless numerical noise within this space, RNet navigates the ambiguity by leveraging data-driven visual priors learned from natural images for semantic inpainting. To preserve the unmanipulated authentic background, the reconstructed image I_{rec} is formulated as a conditional composite mapping:

$$I_{rec} = I_{atk} \odot (1 - \hat{M}) + \text{RNet}\Big(I_{atk} \odot (1 - \hat{M}) \oplus \hat{M}\Big) \odot \hat{M}, \tag{7}$$

where $\oplus$ denotes channel-wise concatenation and $\odot$ represents Hadamard product. This design restricts network modifications entirely within the tampered regions bounded by $\hat{M}$.

To ensure the synthesized content is visually plausible and algebraically verifiable, RNet is optimized via a dual-objective loss:

$$\mathcal{L}_{rec} = \lambda_{vis}\mathcal{L}_{vis}(I_{rec}, I_{wm}) + \lambda_{syn}\|HI_{rec} - G(m)\|_2^2. \tag{8}$$

Here, $\mathcal{L}_{vis}$ combines mean squared error and LPIPS to maintain fidelity, while the syndrome term acts as an implicit algebraic regularization. By penalizing deviations from the continuous invariant during training, RNet learns to generate semantic content that naturally aligns with the valid algebraic subspace, bypassing the limitations of pure linear solvers discussed in Curse 2.

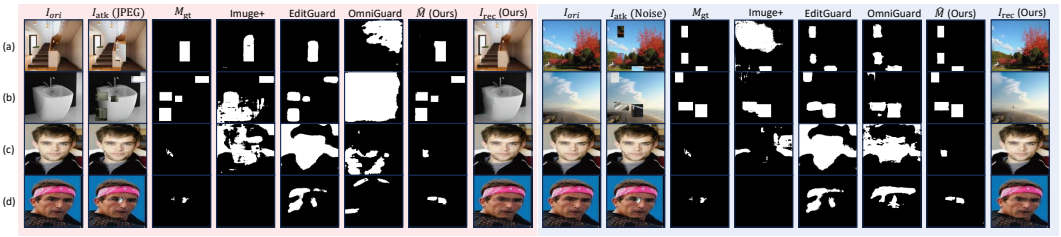


Figure 3: Visual comparison of robustness against JPEG compression (left, red) and Gaussian noise (right, blue) across various manipulations where (a) splicing and (b) copy-move attacks on the PASCAL VOC dataset (Everingham et al., 2010), and semantic edits including (c) adding scars and (d) adding glasses on the CelebA dataset (Liu et al., 2015).

Table 1: Quantitative comparison on the CelebA (Liu et al., 2015) under deepfake manipulation.

	Method	Localization									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Passive	MVSS-Net (Dong et al., 2022)	0.2085	0.8948	0.1618	0.8936	0.1453	0.8020	0.1357	0.8651	0.5101	0.9770
Passive	HiFiNet (Guo et al., 2023)	0.0636	0.8545	0.0384	0.8575	0.0447	0.7676	0.0186	0.6881	0.1601	0.9572
Passive	RelayFormer (Huang et al., 2026)	0.3001	0.9714	0.3221	0.9637	0.2676	0.9612	0.2828	0.9563	0.4049	0.9733
Proactive	Imuge+ (Ying et al., 2023)	0.0260	0.6567	0.0481	0.8523	0.0088	0.3843	0.0339	0.7993	0.0469	0.8084
Proactive	EditGuard (Zhang et al., 2024)	0.0426	0.7895	0.0158	0.5792	0.1680	0.8671	0.1078	0.8746	0.0967	0.9520
Proactive	OmniGuard (Zhang et al., 2025)	0.4099	0.9807	0.0508	0.8252	0.0174	0.5619	0.0243	0.6329	0.5486	0.9978
Proactive	RecoverMark (An et al., 2026b)	0.5489	0.9641	0.5177	0.9685	0.7519	0.9942	0.7870	0.9933	0.9746	0.9963
Proactive	SyndromeMark (Ours)	0.5538	0.9644	0.5220	0.9899	0.4377	0.9028	0.4350	0.9901	0.5057	0.9874

Method	Reconstruction									
	Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Imuge+ (Ying et al., 2023)	14.04	0.0140	12.11	0.0150	10.22	0.0191	10.87	0.0164	10.95	0.0177
RecoverMark (An et al., 2026b)	18.86	0.1448	22.48	0.1428	22.86	0.1413	22.38	0.1417	29.10	0.0169
SyndromeMark (Ours)	23.76	0.1479	24.78	0.1923	25.67	0.1483	29.58	0.0536	32.85	0.0252

6 Experimental Evaluation

We evaluate our proposed methods against state-of-the-art forensic methods to validate the synergy between the continuous algebraic foundation and neural networks. Due to space constraints, the main text focuses on the core capabilities: manipulation localization, semantic reconstruction under global distortions, and the necessity of the algebraic prior. Comprehensive implementation details and complete evaluations are provided in Appendix C and D.

6.1 Experimental Setup

Datasets. To ensure the learned invariants are domain-agnostic rather than overfitted to specific sensor noises or low-level textures inherent in natural datasets, the entire framework is trained on $5,000$ images randomly generated via Stable Diffusion (Rombach et al., 2022). We demonstrate generalization through zero-shot testing on unseen natural datasets, specifically PASCAL VOC (Everingham et al., 2010) for general manipulations and CelebA (Liu et al., 2015) for facial deepfakes, without any fine-tuning.

Baselines and Metrics. We benchmark against passive forensics (MVSS-Net (Dong et al., 2022), HiFiNet (Guo et al., 2023), RelayFormer (Huang et al., 2026)) and proactive defenses (EditGuard (Zhang et al., 2024), OmniGuard (Zhang et al., 2025), Imuge+ (Ying et al., 2023), RecoverMark (An et al., 2026b)). Localization performance is evaluated using pixel-level F1-Score and Area Under the Curve (AUC). Semantic reconstruction fidelity is measured by Peak Signal-to-Noise Ratio (PSNR) in dB and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018).

6.2 Main Results: Localization and Semantic Reconstruction

We evaluate SyndromeMark against state-of-the-art baselines. As illustrated in Figure 3, existing proactive defenses typically rely on fragile empirical embeddings that degrade globally under post-processing. These baselines misinterpret the global distortions as massive tampering, yielding

highly inaccurate masks. Guided by the algebraically-anchored prior, SyndromeMark successfully maintains precise tamper boundaries and reconstructs visually coherent content.

Quantitatively, Table 1 presents evaluations on CelebA (Liu et al., 2015) against Deepfakes (extended results on PASCAL VOC (Everingham et al., 2010) are provided in Appendix D). Under global distortions, SyndromeMark reports moderate F1-scores ($\sim 0.43 - 0.55$) but exceptionally high AUC scores. This is because deepfake manipulations typically involve soft-blended boundaries, making hard-thresholding metrics like F1 highly sensitive to global noise. The high AUC confirms that our syndrome prior consistently preserves the accurate probabilistic ranking of tampered versus untampered pixels. While RecoverMark occasionally yields competitive localization scores, it fundamentally relies on the restrictive assumption of a nearly unmodified background. Regarding reconstruction, SyndromeMark consistently outperforms frameworks like RecoverMark (e.g., 29.58 dB vs. 22.38 dB under JPEG compression), which indicates the neural network can successfully reconstruct visually plausible semantic content guided by the continuous algebraic constraints.

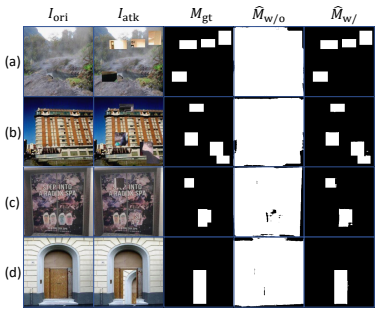


Figure 4: Localization ablation with vs. without the syndrome prior under (a) Splicing + Noise, (b) Splicing + JPEG, (c) Copy-move + Noise, and (d) Copy-move + JPEG.

6.3 Ablation Study: The Synergy of Algebra and Neural Networks

To validate the architectural design, we conduct input-modality ablation studies on the Localization Network (LNet) to isolate how the algebraically-anchored prior mitigates reality gaps.

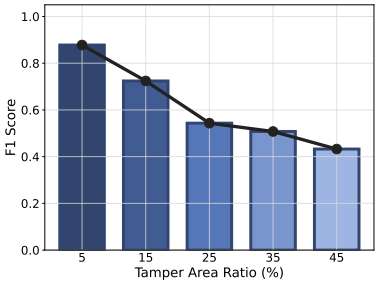


Figure 5: Ablation Study of Phase 2 Localization under Syndrome-Only Input.

Necessity of the Syndrome Prior. We compare the LNet incorporating the syndrome prior ($\hat{M}_{w/}$) against a baseline trained without the prior ($\hat{M}_{w/o}$). As illustrated in Figure 4, the RGB-only baseline fails catastrophically under external distortions. Relying purely on empirical artifacts, it cannot differentiate localized manipulations from global post-processing, yielding massive false positives. In contrast, incorporating f_{syn} precisely isolates tampered regions, confirming that the syndrome prior is indispensable for robust localization.

Necessity of the Host Image. Conversely, feeding LNet only the syndrome prior (f_{syn}) without RGB context (I_{atk}) yields strong F1-scores (~ 0.85) at lower tamper ratios (see Figure 5). However, as the tamper area increases to 45%, performance degrades significantly (F1 drops to ~ 0.43). This aligns with Curse 1: large manipulations trigger severe internal self-interference, causing host textures to leak into the syndrome map. Without the RGB input to provide semantic and spatial context, LNet struggles to distinguish true tampering from texture leakage. Thus, robust localization fundamentally requires fusing the algebraic prior with the semantic context of the host image.

7 Conclusion

In this paper, we have proposed a paradigm shift from empirical fitting to verifiable proactive image forensics. We have introduced Continuous Syndrome Forensics (CSF), which extends classical discrete error-correction codes to the continuous spatial domain to establish a rigorous algebraic foundation. To bridge this idealized theory and the physical reality of natural images, we have developed SyndromeMark as a network-augmented solver. By leveraging continuous syndrome residuals as algebraically-anchored spatial priors, SyndromeMark resolves internal texture leakage and navigates the underdetermined ambiguity of large-area manipulations via semantic reconstruction. Extensive evaluations demonstrate that this neuro-algebraic synergy achieves precise localization and visually coherent reconstruction, outperforming state-of-the-art baselines while remaining robust against severe global distortions.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and generate synthetic datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to create or edit software code, source/search for information, and edit the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, all AI-edited mathematical proofs and texts were rigorously verified manually to ensure logical correctness, and AI-generated code and training datasets were thoroughly tested and inspected. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates proactive image forensics for academic research using synthetic images and existing image datasets. The proposed method is intended to support image integrity assessment. Its localization outputs may contain errors, and reconstructed content is not guaranteed to reproduce the original image and should not be treated as authentic evidence. We acknowledge potential misuse of manipulation and attack techniques and encourage responsible use that respects privacy and the usage conditions of the underlying datasets and models.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3–5 describe the problem formulation, model design, and training objectives. Appendix B presents the assumptions and proofs of the theoretical results. Appendix C provides implementation details, hyperparameter settings, manipulation and distortion configurations, and training pseudocode. We will publicly release the source code upon acceptance.

REFERENCES

Ahmad ALBarqawi, Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah, and NhatHai Phan. Vigtext: Deepfake image detection with vision-language model explanations and graph neural networks. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*. Internet Society, 2026.

Haonan An, Guang Hua, Zhengru Fang, Guowen Xu, Susanto Rahardja, and Yuguang Fang. Decoder gradient shield: Provable and high-fidelity prevention of gradient-based box-free watermark removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13424–13433, 2025.

Haonan An, Guang Hua, Wei Du, Hangcheng Cao, Yihang Tao, Guowen Xu, Susanto Rahardja, and Yuguang Fang. Decoder gradient shields: A family of provable and high-fidelity methods against gradient-based box-free watermark removal. *IEEE Transactions on Dependable and Secure Computing*, 2026a.

Haonan An, Xiaohui Ye, Guang Hua, Yihang Tao, Hangcheng Cao, Xiangyu Yu, and Yuguang Fang. Recovermark: Robust watermarking for localization and recovery of manipulated faces. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18629–18639, 2025.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Heather Chen and Kathleen Magramo. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. CNN, February 2024. URL https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk.

Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Don Fallis. The epistemic threat of deepfakes. *Philosophy & Technology*, 34(4):623–643, 2021.

Han Fang, Zhaoyang Jia, Zehua Ma, Ee-Chien Chang, and Weiming Zhang. Pimog: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 2267–2275, 2022.

Tomáš Filler, Jan Judas, and Jessica Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3):920–935, 2011.

Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 55–73. IEEE, 2024.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3155–3165, 2023.

Wen Huang, Jiarui Yang, Tao Dai, Jiawei Li, Shaoxiong Zhan, Bin Wang, and Shu-Tao Xia. Relayformer: A unified local-global attention framework for scalable image and video manipulation localization. In *International Conference on Learning Representations*, 2026.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Andre Kassis and Urs Hengartner. Unmarker: a universal attack on defensive image watermarking. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2602–2620. IEEE, 2025.

Eric Kee, James F O'Brien, and Hany Farid. Exposing photo manipulation with inconsistent shadows. *ACM Transactions on Graphics (ToG)*, 32(3):1–12, 2013.

Jiafeng Li, Ying Wen, and Lianghua He. M^2rl-net: Multi-view and multi-level relation learning network for weakly-supervised image forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4743–4751, 2025.

Shu Lin, Daniel J Costello, et al. *Error control coding*, volume 2. Prentice hall Scarborough, 2001.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. *Pattern Recognition*, 42(11):2492–2501, 2009.

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 *International Conference on Learning Representations*, volume 2025, pp. 87310–87327, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Siwei Lyu, Xunyu Pan, and Xing Zhang. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110(2):202–221, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 461–478, 2019.

Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Ziqi Sheng, Junyan Wu, Wei Lu, and Jiantao Zhou. Weakly-supervised image forgery localization via vision-language collaborative reasoning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 988–996, 2026.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Eric Wengrowski and Kristin Dana. Light field messaging with deep photographic steganography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1515–1524, 2019.

John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2): 210–227, 2008.

Qichao Ying, Zhenxing Qian, Hang Zhou, Haisheng Xu, Xinpeng Zhang, and Siyi Li. From image to imuge: Immunized image generation. In *Proceedings of the 29th ACM international conference on Multimedia*, pp. 3565–3573, 2021.

Qichao Ying, Hang Zhou, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Learning to immunize images for tamper localization and self-recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13814–13830, 2023.

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

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11964–11974, 2024.

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3008–3018, 2025.

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.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2621–2639. IEEE, 2025.

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 (ECCV)*, pp. 657–672, 2018.

Appendix for: Continuous Syndrome Forensics for Proactive Image Manipulation Localization and Reconstruction

CONTENTS

A Comprehensive Background and Related Works **15**

A.1 Deep Image Watermark & Proactive Forensics 15

A.2 From Block Codes to Continuous Syndromes 15

B Theoretical Analysis for CSF **16**

B.1 Proposition 1 (Block-wise Spatial Isolation) 16

B.2 Theorem 1 (Algebraic Resolution Boundary) 17

B.3 Theorem 2 (Deterministic Subspace Projection Bounds) 18

B.4 Theorem 3 (Critical Rank Condition) 18

C Implementation Details and Pseudocode **19**

C.1 Network Architecture and Hyperparameters 19

C.2 Configuration of Manipulations and Distortions 19

C.3 A Note on RecoverMark 19

C.4 Algorithmic Workflow 20

D Extended Experimental Results **20**

D.1 Additional Qualitative Visualizations 20

D.2 Full Quantitative Evaluations 20

D.3 Robustness Against Adaptive Surrogate Attacks 23

E Limitations & Further Discussion **24**

A Comprehensive Background and Related Works

A.1 Deep Image Watermark & Proactive Forensics

Deep image watermarking primarily serves as a tool for ownership verification and misuse tracing, safeguarding the intellectual property of image assets (An et al., 2025; Kassis & Hengartner, 2025; Zhao et al., 2025; An et al., 2026a). Previously, the primary research has focused on enhancing the robustness and generalization of watermarks against complex, real-world distortions. For instance, HiDDeN (Zhu et al., 2018) enhances watermark robustness by incorporating a distortion layer with differentiable JPEG compression. LFM (Wengrowski & Dana, 2019) achieves robust watermarking in screen-camera scenarios through large-scale image training, while StegaStamp (Tancik et al., 2020) addresses the physical distortions and generalization challenges inherent in print-camera settings. Furthermore, PIMoG (Fang et al., 2022) resolves issues related to Moiré patterns and screen reflections specifically encountered in screen-camera captures. More recently, TrustMark (Bui et al., 2025) addresses the scalability issue by introducing a universal framework compatible with arbitrary-resolution images.

Meanwhile, some research has noticed that watermarking can act as a proactive defense in image forensics, addressing the fundamental limitations of passive forensics (Dong et al., 2022; Guo et al., 2023; Huang et al., 2026), which heavily rely on prior knowledge of manipulation traces and often fail under post-processing. By preemptively embedding redundant signals, proactive frameworks are designed to enable precise manipulation localization and, in some cases, content reconstruction. However, existing methods exhibit critical vulnerabilities. Frameworks focusing primarily on localization, such as EditGuard (Zhang et al., 2024) and OmniGuard (Zhang et al., 2025), employ a dual-watermark strategy. However, their reliance on fragile watermarks for localization leaves them highly susceptible to global distortions and watermark removal attacks, a vulnerability further exacerbated by the severe mutual interference caused by co-embedding the robust signal for ownership verification. Methods that extend to image content reconstruction, such as Imuge (Ying et al., 2021) and Imuge+ (Ying et al., 2023), are similarly built upon fragile embedding mechanisms, causing the entire process to collapse under common post-processing. While recent frameworks like RecoverMark (An et al., 2026b) attempt to achieve robust manipulation localization and semantic reconstruction by embedding salient facial features into the surrounding background, they rely on a strong assumption that the background remains nearly unperturbed and are strictly limited to domain-specific scenarios (e.g., facial deepfakes). Most importantly, all these existing proactive defenses rely on empirical fitting without offering theoretical analysis for verifiable localization or reconstruction, directly motivating our proposed paradigm shift.

A.2 From Block Codes to Continuous Syndromes

The core theoretical inspiration of our paradigm stems from classical Systematic Linear Block Codes (SLBC), a primary category of error-correction codes in digital communications (Lin et al., 2001; Filler et al., 2011). In a discrete finite field (e.g., Galois Field $\mathbb{F}_2$), an SLBC encodes a K-bit message x into an N-bit codeword $c \in \mathbb{F}_2^N$. A critical feature of a systematic code is that the original message is preserved unaltered, with $(N-K)$ parity bits p simply appended, such that $c = [x \mid p]$. Mathematically, this establishes an algebraic constraint using a parity-check matrix $\Phi \in \mathbb{F}_2^{(N-K)\times N}$, ensuring that any valid codeword satisfies $\Phi c^\top = 0$. When transmitting through a noisy channel, the received signal becomes $y = c + \epsilon$, where ϵ represents sparse bit-flip errors. The receiver can locate and recover these errors by calculating the syndrome vector $z = \Phi y^\top = \Phi(c + \epsilon)^\top = \Phi\epsilon^\top$. This elegant operation annihilates the massive message to strictly isolate the sparse error. The correction capacity is bounded by the minimum Hamming distance d_{min}, guaranteeing the exact reconstruction of up to $t = \lfloor (d_{min} - 1)/2 \rfloor$ bit flips (Lin et al., 2001).

We draw a fundamental analogy between SLBC and proactive image forensics. The original host image acts as the systematic message that must remain visually similar, the proactive injected redundancy serves as the parity injection, and malicious image manipulations and distortions correspond to the bit-flip errors. Consequently, by validating the received image against a continuous secret check matrix, we can extract an error map that mirrors the concept of a syndrome. This analogy provides a highly desirable property: rather than relying on empirical feature extraction to guess

tampering locations, we can theoretically compute a syndrome that effectively isolates the manipulation from the complex image background.

Despite its mathematical elegance, directly applying discrete coding theory to natural images is impractical. Natural images are continuous real-valued signals in a high-dimensional manifold. Discrete SLBC operates over finite fields and treats any pixel value deviation as a hard bit-flip. Consequently, it is highly vulnerable to energy-bounded distortions like JPEG compression or random noise, which are ubiquitous in multimedia transmission but do not alter the semantic truth (Wang et al., 2004). This physical reality necessitates a paradigm shift from finite fields to the continuous real field. Previously, the concept of real-number error correction was explored by Candes and Tao (Candes & Tao, 2005), whose pioneering work laid the foundation for compressive sensing (Donoho, 2006) by demonstrating that a continuous real-valued signal can be exactly recovered from sparse arbitrary errors using convex ℓ_1-minimization, which is a principle later successfully adapted to robust computer vision (Wright et al., 2008).

While Candes and Tao established the elegant theoretical foundation for real-field error correction, their methodology was designed for abstract signal transmission over noisy channels rather than multimedia security. Specifically, satisfying theoretical guarantees such as the Restricted Isometry Property (RIP) typically necessitates dense encoding matrices. However, directly applying such dense matrices would severely distort the visual content of an image, violating the fundamental invisibility requirement of proactive forensics. Additionally, classical compressive sensing relies on a strict global sparsity assumption that practical large-area semantic manipulations typically violate. Furthermore, their communication-centric model is built for reliability against blind channel errors, and thus fundamentally lacks the cryptographic key-binding mechanisms required to resist intelligent adversarial forgery. This exact gap between abstract coding theory and adversarial multimedia protection motivates our theoretical formulation. By mathematically modeling the host image as a continuous systematic message subject to an immutable, secret-keyed spatial constraint, we adapt the powerful insight of real-field syndromes into a dedicated paradigm for verifiable image forensics.

B Theoretical Analysis for CSF

In this section, we provide the formal theoretical analysis that underpins Continuous Syndrome Forensics (CSF). Let the host image be flattened into a vector of n elements, and the global continuous syndrome constraint be defined as $HI_{wm} = G(m)$, where $H \in \mathbb{R}^{d\times n}$ $(d < n)$. To avoid the intractable global optimization and error smearing associated with dense projection matrices discussed in classic compressive sensing, CSF structurally restricts H to a sparse block-diagonal topology.

Specifically, aligning with the formulation in Section 4.1, the image is partitioned into B non-overlapping local blocks, where each local block contains $n_b = n/B$ elements. The global matrix H is constructed as a block-diagonal matrix $H = \mathrm{diag}(H_{sub}, H_{sub}, \ldots, H_{sub})$, where the shared local projection block $H_{sub} \in \mathbb{R}^{d_b\times n_b}$ maps a local image patch to a local syndrome vector of dimension $d_b = d/B$. We note that, unless explicitly stated otherwise, all discussions in this section assume that the exact embedding constraint $HI_{wm} = G(m)$ strictly holds for the pure algebraic analysis.

B.1 Proposition 1 (Block-wise Spatial Isolation)

A fundamental vulnerability of compressive sensing is error smearing, where localized manipulations mathematically contaminate the entire back-projected spatial map. Proposition 1 formally demonstrates that our block-diagonal topology inherently circumvents this issue, providing an isolated spatial prior for manipulation localization.

Proposition 1. *(Block-wise Spatial Isolation). Let the composite error across the entire image be $e \in \mathbb{R}^n$. By spatially partitioning the image into B non-overlapping blocks, we define $e_b \in \mathbb{R}^{n_b}$ as the local error vector within the b-th block for all $b \in \{1, 2, \ldots, B\}$. Consequently, the global error can be represented as the concatenated vector $e = [e_1^\top, e_2^\top, \ldots, e_B^\top]^\top$. Under the shared block-diagonal topology $H = \mathrm{diag}(H_{sub}, \ldots, H_{sub})$, the back-projected spatial syndrome map $f_{syn} = H^\top(He)$ can be correspondingly partitioned as $f_{syn} = [f_{syn,1}^\top, \ldots, f_{syn,B}^\top]^\top$. This global operation decomposes into mutually independent local mappings:*

$$f_{syn,b} = H_{sub}^\top H_{sub} e_b, \quad \forall b \in \{1, 2, \ldots, B\}. \tag{9}$$

Therefore, the cross-block mathematical interference is exactly zero: $\frac{\partial f_{syn,i}}{\partial e_j} = \mathbf{0}$ *for any* $i \neq j$.

Proof. By the definition of the block-diagonal matrix H, its transpose $H^\top$ is identically block-diagonal with $H^\top = \mathrm{diag}(H_{sub}^\top, \ldots, H_{sub}^\top)$. The global back-projection can be expanded via block matrix multiplication:

$$f_{syn} = H^\top He = \begin{bmatrix} H_{sub}^\top H_{sub} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & H_{sub}^\top H_{sub} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & H_{sub}^\top H_{sub} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_B \end{bmatrix} = \begin{bmatrix} H_{sub}^\top H_{sub} e_1 \\ H_{sub}^\top H_{sub} e_2 \\ \vdots \\ H_{sub}^\top H_{sub} e_B \end{bmatrix}. \quad (10)$$

By matching the b-th block component, we obtain $f_{syn,b} = H_{sub}^\top H_{sub} e_b$. Taking the partial derivative of $f_{syn,i}$ with respect to e_j ($i \neq j$) yields a zero matrix. This confirms that any manipulation in block j has no mathematical influence on the spatial residual of block i. □

Remark: Rather than relying on strict global sparsity assumptions and intractable global optimization (such as ℓ_1-minimization) for exact element-level reconstruction, CSF uses Proposition 1 to securely isolate abnormal energy within independent blocks. If an anomaly is detected in $f_{syn,b}$, the tampering is deterministically bounded within the spatial footprint of block b. This establishes a robust and noise-isolated spatial prior for LNet to perform subsequent processing.

B.2 Theorem 1 (Algebraic Resolution Boundary)

Having established the spatial independence of local blocks, we now analyze the theoretical localization capability within a single block. Theorem 1 demonstrates that the algebraic syndrome exhibits a dual-resolution nature, providing a deterministic resolution boundary based on the sparsity of the manipulation.

Theorem 1. *(Algebraic Resolution Boundary). Let* $H_{sub} \in \mathbb{R}^{d_b \times n_b}$ *(where* $d_b < n_b$*) be the local projection matrix. Let* $e_b \in \mathbb{R}^{n_b} \setminus \{\mathbf{0}\}$ *represent a non-zero local manipulation vector within a specific image block* b*. Assuming zero external noise for the pure algebraic analysis, the generated local syndrome residual* $s_b = H_{sub} e_b \in \mathbb{R}^{d_b}$ *provides the following dual-resolution guarantees:*

1. ***Element-level Determinism:*** *Let* $\mathrm{spark}(H_{sub})$ *denote the smallest number of linearly dependent columns in* H_{sub}*. If the number of manipulated elements satisfies* $\|e_b\|_0 < \mathrm{spark}(H_{sub})/2$*, then* e_b *is the mathematically unique sparsest solution to the underdetermined linear system. Consequently, the tampering can be theoretically localized at the exact element level.*

2. ***Block-level Graceful Degradation:*** *If* $\|e_b\|_0 \geq \mathrm{spark}(H_{sub})/2$*, uniqueness of the sparsest solution is no longer guaranteed by the preceding sufficient condition. However, provided* $e_b \notin \mathcal{N}(H_{sub})$*, where* $\mathcal{N}(H_{sub})$ *is the null space of* H_{sub}*, the local syndrome retains a non-zero energy response (i.e.,* $\|s_b\|_2 > 0$*), yielding a nonzero block-level response in the exact, noise-free setting.*

Proof. Part 1 (Element-level Determinism): We prove this by contradiction. Assume there exists an alternative manipulation pattern $e_b' \in \mathbb{R}^{n_b}$ ($e_b' \neq e_b$) that generates the exact same syndrome s_b, with its number of non-zero elements also satisfying $\|e_b'\|_0 < \mathrm{spark}(H_{sub})/2$. Subtracting the two equations yields $H_{sub}(e_b - e_b') = \mathbf{0}$. Let $\delta = e_b - e_b'$ be a non-zero difference vector in the null space of H_{sub}. By the triangle inequality for the ℓ_0 pseudo-norm, we have $\|\delta\|_0 \leq \|e_b\|_0 + \|e_b'\|_0 < \mathrm{spark}(H_{sub})$. However, $H_{sub}\delta = \mathbf{0}$ implies that the non-zero elements of δ linearly combine the corresponding columns of H_{sub} to form a zero vector, implying that these columns are linearly dependent. According to the definition of spark, this requires $\|\delta\|_0 \geq \mathrm{spark}(H_{sub})$. This contradiction confirms that e_b is the unique sparsest solution.

Part 2 (Block-level Graceful Degradation): When the sparsity condition is violated, the uniqueness of the sparsest solution is no longer guaranteed. However, the vector e_b generates a zero syndrome (i.e., $s_b = \mathbf{0}$) if and only if $e_b \in \mathcal{N}(H_{sub})$. Since the condition explicitly states $e_b \notin \mathcal{N}(H_{sub})$, the matrix-vector multiplication yields a non-zero syndrome vector, resulting in an observable energy response $\|s_b\|_2 > 0$. This confirms the anomaly at the block level regardless of internal sparsity. □

Remark: This theorem provides sufficient conditions for sparse-error identifiability and characterizes nonzero block responses in our local projection setting. As detailed in Appendix C, our spatial block consists of a 2×2 RGB patch, yielding a flattened input vector of $n_b = 12$ elements, mapped to an 8-dimensional local syndrome ($d_b = 8$). If the learned 8×12 matrix H_{sub} is full spark, then $spark(H_{sub}) = d_b + 1 = 9$. Under this condition, Theorem 1 guarantees uniqueness of the sparsest error solution when at most four scalar elements within a block are manipulated. However, realistic semantic manipulations typically exceed the threshold. Beyond this sufficient sparsity condition, Theorem 1 still establishes a nonzero block response for errors outside the local null space in the exact, noise-free setting.

B.3 Theorem 2 (Deterministic Subspace Projection Bounds)

As discussed in Section 3.1, our framework must manage energy-bounded global distortions. Here, we derive a deterministic lower bound for the back-projected syndrome that depends on the tampering magnitude, its alignment with the projection row space, and the noise magnitude.

Theorem 2. *(Deterministic Subspace Projection Bounds). Let $H_{sub} \in \mathbb{R}^{d_b \times n_b}$ (where $d_b < n_b$) be the local projection matrix satisfying the semi-orthogonal constraint $H_{sub}H_{sub}^\top = I_{d_b}$. Within any local block b, let the composite error be $e_b = e_{t,b} + e_{n,b} \in \mathbb{R}^{n_b}$, where $e_{t,b}$ is the localized tampering vector and $e_{n,b}$ represents an arbitrary external noise vector. The back-projected local syndrome residual $f_{syn,b} = H_{sub}^\top H_{sub}(e_{t,b} + e_{n,b})$ satisfies the following energy lower bound:*

$$\|f_{syn,b}\|_2 \geq \|e_{t,b}\|_2 \cos\theta_b - \|e_{n,b}\|_2, \tag{11}$$

where $\theta_b \in [0, \pi/2]$ is the principal angle between $e_{t,b}$ and the row space of H_{sub}.

Proof. Let the projection operator be denoted as $P_{sub} = H_{sub}^\top H_{sub}$. Because the semi-orthogonal parameterization explicitly enforces $H_{sub}H_{sub}^\top = I_{d_b}$, we can expand the squared operator as $P_{sub}^2 = H_{sub}^\top (H_{sub}H_{sub}^\top) H_{sub} = H_{sub}^\top I_{d_b} H_{sub} = P_{sub}$. Furthermore, its transpose evaluates to $P_{sub}^\top = (H_{sub}^\top H_{sub})^\top = H_{sub}^\top (H_{sub}^\top)^\top = H_{sub}^\top H_{sub} = P_{sub}$. These two derived properties ($P_{sub}^2 = P_{sub}$ and $P_{sub}^\top = P_{sub}$) mathematically confirm that P_{sub} acts as a symmetric orthogonal projector onto the row space of H_{sub}.

By applying the reverse triangle inequality to the projected composite error, we obtain:

$$\|f_{syn,b}\|_2 = \|P_{sub}(e_{t,b} + e_{n,b})\|_2 \geq \|P_{sub}e_{t,b}\|_2 - \|P_{sub}e_{n,b}\|_2. \tag{12}$$

For the noise component $e_{n,b}$, since the spectral norm of an orthogonal projector is exactly 1 (i.e., $\|P_{sub}\|_2 = 1$), it acts as a non-expansive mapping: $\|P_{sub}e_{n,b}\|_2 \leq \|e_{n,b}\|_2$. Consequently, we have $-\|P_{sub}e_{n,b}\|_2 \geq -\|e_{n,b}\|_2$.

For the tampering component $e_{t,b}$, its projection onto the row space of H_{sub} geometrically forms a right triangle with its orthogonal complement. The Euclidean length of this projected vector evaluates exactly to $\|P_{sub}e_{t,b}\|_2 = \|e_{t,b}\|_2 \cos\theta_b$.

Substituting these two geometric derivations into the reverse triangle inequality directly yields the stated lower bound: $\|f_{syn,b}\|_2 \geq \|e_{t,b}\|_2 \cos\theta_b - \|e_{n,b}\|_2$. □

Remark: The projection operator is non-expansive for both tampering and noise. Under the assumed magnitude gap between tampering and external distortions, Theorem 2 helps explain when a tampering response can remain distinguishable in the residual domain. Specifically, the lower bound is positive provided that $\|e_{t,b}\|_2 \cos\theta_b > \|e_{n,b}\|_2$. Thus, a large tampering magnitude is favorable, but its alignment with the projection row space also matters. Under exact embedding, this condition ensures a nonzero residual despite the bounded noise. In practice, embedding mismatch also contributes to the residual, which LNet uses together with RGB context for localization.

B.4 Theorem 3 (Critical Rank Condition)

In Theorem 3, we formalize the limitation on reconstructing the original image content solely through algebraic equations.

Theorem 3. *(Critical Rank Condition). Let $H_{sub} \in \mathbb{R}^{d_b \times n_b}$ (where $d_b < n_b$) be the local projection matrix. Within a local block b, let the n_b scalar elements of the watermarked image $I_{wm,b} \in \mathbb{R}^{n_b}$ be partitioned into a manipulated set M (where $|M| = k$) and an untouched set U (where $|U| = n_b - k$). This yields a vector representation $I_{wm,b} = [I_M^\top, I_U^\top]^\top$ and a corresponding column partition of the projection matrix $H_{sub} = [H_M, H_U]$, where $H_M \in \mathbb{R}^{d_b \times k}$. Given the known untouched elements I_U, the original values of the manipulated elements I_M can be uniquely recovered solely via the algebraic constraint $H_{sub} I_{wm,b} = G(m)_b$ if and only if* $\mathrm{rank}(H_M) = k$.

Proof. The fundamental algebraic constraint for the local block dictates $H_{sub} I_{wm,b} = G(m)_b$. Expanding this equation using the spatial block partitions yields $H_M I_M + H_U I_U = G(m)_b$. Isolating the unknown elements I_M, we reduce the task to solving the linear system:

$$H_M I_M = G(m)_b - H_U I_U. \tag{13}$$

Sufficiency: If $\mathrm{rank}(H_M) = k$, which inherently requires $k \leq d_b$, the Gram matrix $H_M^\top H_M$ is full-rank and invertible. This yields a unique analytical solution: $I_M = (H_M^\top H_M)^{-1} H_M^\top (G(m)_b - H_U I_U)$.

Necessity: If $\mathrm{rank}(H_M) < k$, which inevitably occurs when the number of manipulated elements exceeds the local syndrome dimension ($k > d_b$), the Rank-Nullity Theorem indicates that the null space $\mathcal{N}(H_M)$ has a strictly positive dimension. Any alternative configuration $I'_M = I_M + z$ (where $z \in \mathcal{N}(H_M) \setminus \{\mathbf{0}\}$) also satisfies $H_M I'_M = G(m)_b - H_U I_U$, yielding an infinite number of valid mathematical solutions. □

Remark: When large-area image manipulations occur, the number of manipulated elements k easily exceeds the local syndrome dimension d_b. Under this condition, the linear system becomes underdetermined. The linear constraints alone do not select a unique or perceptually plausible reconstruction. This mathematical ambiguity shows that navigating this infinite solution space requires additional priors or constraints, directly motivating the integration of the semantic Reconstruction Network (RNet) in our proposed SyndromeMark framework.

C Implementation Details and Pseudocode

C.1 Network Architecture and Hyperparameters

The shared local projection H_{sub} is implemented via non-overlapping 2×2 convolutions (stride 2). For RGB images, this maps a local spatial patch to a flattened input vector of $n_b = 12$ scalar elements, which is then projected to a local syndrome of dimension $d_b = 8$. The watermark message length is set to $L = 128$. For the network components, we employ commonly used U-Net (Ronneberger et al., 2015) backbones for both LNet and RNet. The entire framework is optimized by the Adam optimizer with a learning rate of 0.0002 and a batch size of 16 on a single NVIDIA RTX 5000 Ada GPU.

C.2 Configuration of Manipulations and Distortions

Malicious manipulations involve randomly located 10%–50% area modifications, including splicing, copy-move, and deepfake generation. The applied external distortions comprise: patch removal covering 10%–20% of the image area, Gaussian blur with a kernel size of 5×5 and a standard deviation of $\sigma = 1.0$, Gaussian noise with a zero mean and a standard deviation of $\sigma = 0.05$, and JPEG compression with a quality factor of 40.

C.3 A Note on RecoverMark

Unlike our paradigm which offers universal protection, RecoverMark (An et al., 2026b) operates under the strong domain-specific assumption that original saliency content can be redundantly embedded into the surrounding background. Its reconstruction capability fundamentally hinges on this strict definition of saliency and background. Consequently, the comparison with RecoverMark is specifically restricted to facial deepfake scenarios (CelebA dataset), as its architecture is intrinsically

not designed to handle the unconstrained object-to-background images found in general tampering datasets like PASCAL VOC (Everingham et al., 2010).

C.4 Algorithmic Workflow

Algorithm 1 summarizes the end-to-end training pipeline of SyndromeMark. Note that the hard orthogonal parameterization of the projection weights guarantees that the semi-orthogonal constraint $H_{sub}H_{sub}^{\top} = I_{d_b}$ remains strictly satisfied throughout all gradient updates in the continuous parameter space.

Algorithm 1 Training Pipeline of SyndromeMark

Require: Original image I_{ori}, Message m, Constraint generator $G(\cdot)$.

// Phase 1: Proactive Embedding (Joint Optimization)
Apply orthogonal parameterization to local projection weights H_{sub} to implicitly construct a global semi-orthogonal projection matrix H (satisfying $HH^{\top} = I_d$).
$I_{wm} \leftarrow Enc(I_{ori} \oplus MLP(m))$.
$I_{dist} \leftarrow$ Distortion(I_{wm}) (e.g., adding JPEG or Noise).
$\hat{m} \leftarrow Dec(I_{dist})$.
Compute embedding loss: $\mathcal{L}_{embed} = \lambda_{fid}\mathcal{L}_{fid} + \lambda_{wm}\mathcal{L}_{wm} + \lambda_{syn}\|HI_{wm} - G(m)\|_2^2$.
Update Enc, MLP, Dec, and H_{sub} via gradient descent.

// Phase 2: Forensic Networks (LNet & RNet Training)
Simulate Attack: $I_{atk} \leftarrow I_{wm} + e_{tamper} + e_{noise}$.
Compute global continuous syndrome residual: $s \leftarrow HI_{atk} - G(m)$.
Back-project to continuous spatial prior: $f_{syn} \leftarrow H^{\top}s$.
Predict manipulation mask: $\hat{M} \leftarrow$ LNet$(I_{atk} \oplus |f_{syn}| \oplus |f_{syn}|^2)$.
Reconstruct semantic content: $I_{rec} \leftarrow I_{atk} \odot (1 - \hat{M}) +$ RNet$\Big(I_{atk} \odot (1 - \hat{M}) \oplus \hat{M}\Big) \odot \hat{M}$.
Compute forensic losses: $\mathcal{L}_{phase2}$ (for LNet) and $\mathcal{L}_{rec}$ (for RNet).
Update LNet and RNet via gradient descent.

D Extended Experimental Results

In this section, we provide comprehensive experimental results that supplement the main text, including full quantitative evaluations, additional visual comparisons, and a result of the framework's robustness against adaptive surrogate attacks.

D.1 Additional Qualitative Visualizations

We comprehensively evaluate the manipulation localization and semantic reconstruction capabilities of SyndromeMark against state-of-the-art (SOTA) baselines.

As already illustrated in Figure 3 of the main text, SOTA proactive baselines, which typically rely on fragile embeddings, produce inaccurate localization masks when subjected to adversarial post-processing. In contrast, SyndromeMark successfully maintains precise tamper boundaries and reconstructs visually coherent content. We further demonstrate this capability in Figure 6, which presents SyndromeMark's consistent localization and reconstruction performance across a broader range of severe global distortions. Moreover, we conduct a false-positive test in Figure 7, where protected images undergo pure global distortions without any manipulation. SyndromeMark rarely yields false positives, outputting nearly all-black masks. Conversely, a baseline LNet trained without the syndrome residual prior ($\hat{M}_{w/o}$) incorrectly flags these distortions as malicious tampering, confirming that our algebraic syndrome prior is necessary and effective in preventing false alarms.

D.2 Full Quantitative Evaluations

Robustness of the Algebraic Invariant Embedding. The foundation of our proactive paradigm relies on the continuous algebraic invariant. Operationally, the defender must reliably extract the

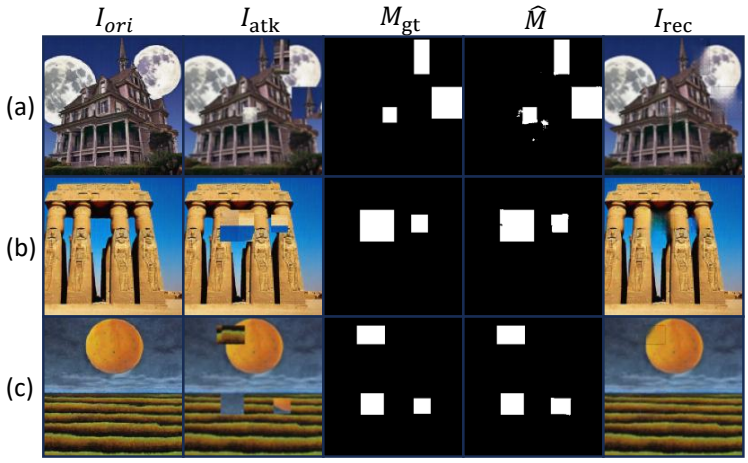


Figure 6: Qualitative results of localization and content reconstruction under copy-move manipulation with various distortions, using random generated images by Stable Diffusion. Rows (a)-(c) illustrate the system's performance against copy-move manipulation under Gaussian blur, Gaussian noise, and JPEG compression, respectively.

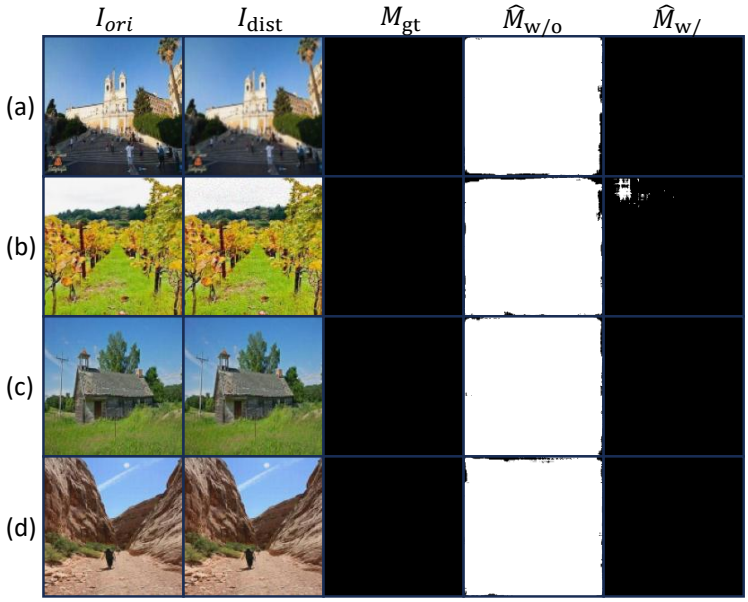


Figure 7: Qualitative results of the false-positive test under various distortions: (a) Gaussian blur, (b) Gaussian noise, (c) JPEG compression, and (d) no distortion.

watermark message m for ownership verification, which serves as the prerequisite for retrieving the exact constraint target $G(m)$ and computing the subsequent continuous syndrome residual. Therefore, evaluating the robustness of this algebraic embedding under various distortions is critical.

As shown in Table 2, the Success Rate of Extraction (SR_{E}) consistently exceeds 95% under most adversarial distortions and manipulations. This high reliability shows that our algebraic invariant is intrinsically robust. Furthermore, the table illustrates a fundamental trade-off between the strength of the forensic spatial prior and the visual fidelity. Increasing the constraint weight λ_{syn} significantly boosts the spatial Signal-to-Noise Ratio (SNR) in the syndrome domain (e.g., surging from ~ 5 dB to > 17 dB), making tampering traces distinct from the background. However, enforcing stronger algebraic constraints naturally incurs a cost in visual fidelity, evidenced by the PSNR dropping from > 32 dB to ~ 26 dB.

Table 2: Demonstration of message extraction robustness ($SR_E \uparrow$) and the trade-off between forensic Signal-to-Noise Ratio (SNR ↑) and visual fidelity (PSNR ↑) under different distortions against splicing and copy-move manipulations on PASCAL VOC (Everingham et al., 2010) dataset.

λ_{syn}	Splicing										
	Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion		
	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	PSNR ↑
1	0.9844	0.7921	1.0000	-1.9116	0.9969	5.1756	0.9969	1.2468	0.9841	1.3334	32.36
10	0.9938	0.8564	0.9969	-1.3488	0.9812	5.5119	0.9719	1.4086	1.0000	1.3241	32.74
100	1.0000	1.9562	1.0000	3.1407	1.0000	9.3517	0.9969	2.8934	0.9781	2.1318	31.65
1000	1.0000	4.3725	1.0000	10.0430	0.9969	15.2999	0.9969	7.7159	0.9844	9.0792	28.60
10000	0.9938	5.8183	0.9969	13.6044	0.9906	17.7134	0.9625	10.1550	0.9531	12.3108	26.98
λ_{syn}	Copy-Move										
	Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion		
	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	$SR_E \uparrow$	SNR ↑	PSNR ↑
1	1.0000	3.1048	1.0000	-1.6017	1.0000	5.0558	1.0000	1.0511	1.0000	1.1791	32.38
10	0.9969	3.4861	0.9875	-1.6201	0.9938	5.4827	0.9500	1.2978	0.9812	1.3087	32.74
100	0.9969	7.3545	1.0000	3.3341	0.9969	10.4121	0.9969	5.2752	0.9969	5.3398	31.65
1000	0.9938	13.5569	1.0000	10.6662	0.9938	17.0463	0.9875	11.9624	0.9906	11.7953	28.60
10000	0.9812	16.6613	0.9938	14.0707	0.9812	19.2736	0.9750	14.8160	0.9750	14.7377	26.99

Table 3: Quantitative comparison on the PASCAL VOC (Everingham et al., 2010) dataset under splicing manipulation.

	Method	Localization									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Passive	MVSS-Net (Dong et al., 2022)	0.5625	0.8748	0.8233	0.9775	0.5404	0.8397	0.5829	0.8617	0.5942	0.8668
Passive	HiFiNet (Guo et al., 2023)	0.3915	0.9024	0.3531	0.7933	0.3364	0.6970	0.3397	0.7232	0.3405	0.7231
Passive	RelayFormer (Huang et al., 2026)	0.9586	0.9978	0.9731	0.9992	0.9549	0.9962	0.9554	0.9963	0.9572	0.9965
Proactive	Imuge+ (Ying et al., 2023)	0.6190	0.9744	0.9777	0.9995	0.5614	0.9405	0.5350	0.9229	0.5248	0.9198
Proactive	EditGuard (Zhang et al., 2024)	0.3588	0.7362	0.8263	0.9966	0.0231	0.3372	0.0301	0.5494	0.0150	0.4896
Proactive	OmniGuard (Zhang et al., 2025)	0.1326	0.5833	0.2069	0.6629	0.1547	0.6033	0.1725	0.6321	0.1934	0.6495
Proactive	SyndromeMark (Ours)	0.8589	0.9821	0.8936	0.9972	0.9320	0.9986	0.8566	0.9891	0.9495	0.9983
	Method	Reconstruction									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
	Imuge+ (Ying et al., 2023)	13.44	0.1633	19.78	0.1363	14.12	0.1607	13.61	0.1637	13.48	0.1649
	SyndromeMark (Ours)	21.32	0.1583	23.05	0.1418	23.45	0.1332	22.59	0.1515	26.98	0.1238

Table 4: Quantitative comparison on the PASCAL VOC (Everingham et al., 2010) dataset under copy-move manipulation.

	Method	Localization									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Passive	MVSS-Net (Dong et al., 2022)	0.5365	0.8635	0.7959	0.9664	0.5495	0.8319	0.5836	0.8556	0.5780	0.8486
Passive	HiFiNet (Guo et al., 2023)	0.4272	0.9080	0.3519	0.7809	0.3398	0.6708	0.3500	0.7452	0.3517	0.7325
Passive	RelayFormer (Huang et al., 2026)	0.9418	0.9957	0.9646	0.9990	0.9436	0.9958	0.9391	0.9941	0.9343	0.9936
Proactive	Imuge+ (Ying et al., 2023)	0.9042	0.9942	0.9787	0.9997	0.8749	0.9893	0.9245	0.9948	0.9059	0.9930
Proactive	EditGuard (Zhang et al., 2024)	0.5855	0.8605	0.8229	0.9969	0.4563	0.7442	0.4447	0.7960	0.4396	0.7449
Proactive	OmniGuard (Zhang et al., 2025)	0.1164	0.5757	0.1883	0.6659	0.1628	0.6181	0.1595	0.6365	0.1739	0.6490
Proactive	SyndromeMark (Ours)	0.9063	0.9878	0.8912	0.9941	0.9401	0.9989	0.8915	0.9913	0.9251	0.9984
	Method	Reconstruction									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
	Imuge+ (Ying et al., 2023)	18.24	0.1414	19.82	0.1357	18.26	0.1409	18.91	0.1391	18.61	0.1398
	SyndromeMark (Ours)	22.01	0.1410	22.93	0.1419	23.64	0.1287	22.83	0.1427	26.99	0.1259

Comprehensive Localization and Semantic Reconstruction.

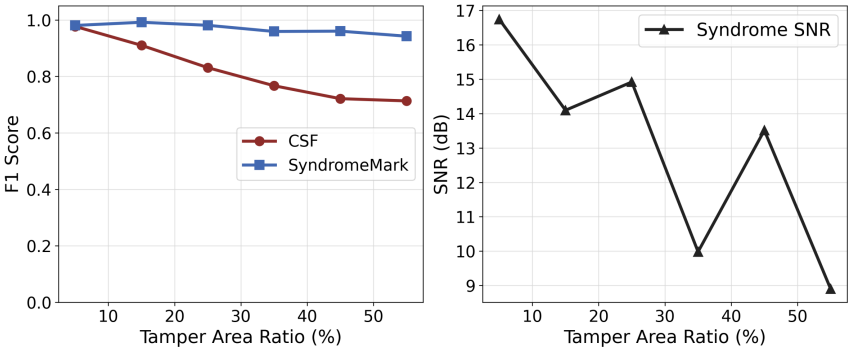
Tables 3 and 4 detail the manipulation localization and content reconstruction performance on the PASCAL VOC dataset under splicing and copy-move manipulations, respectively. To provide a complete and unified overview of all quantitative evaluations in a single section, Table 5 is included to reproduce the CelebA deepfake results previously presented in the main text.

The quantitative results across these three tables align well with our qualitative observations. While most baselines perform well under no distortion, their performance degrades sharply under global distortions. For instance, under JPEG compression in Table 3, the localization F1-score of EditGuard falls to 0.0301, whereas SyndromeMark maintains a robust score of 0.8566. In Table 5, SyndromeMark reports moderate F1-scores ($\sim 0.43 - 0.55$) but exceptionally high AUC scores.

Table 5: Quantitative comparison on the CelebA (Liu et al., 2015) dataset under deepfake manipulation.

	Method	Localization									
		Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
		F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Passive	MVSS-Net (Dong et al., 2022)	0.2085	0.8948	0.1618	0.8936	0.1453	0.8020	0.1357	0.8651	0.5101	0.9770
Passive	HiFiNet (Guo et al., 2023)	0.0636	0.8545	0.0384	0.8575	0.0447	0.7676	0.0186	0.6881	0.1601	0.9572
Passive	RelayFormer (Huang et al., 2026)	0.3001	0.9714	0.3221	0.9637	0.2676	0.9612	0.2828	0.9563	0.4049	0.9733
Proactive	Imuge+ (Ying et al., 2023)	0.0260	0.6567	0.0481	0.8523	0.0088	0.3843	0.0339	0.7993	0.0469	0.8084
Proactive	EditGuard (Zhang et al., 2024)	0.0426	0.7895	0.0158	0.5792	0.1680	0.8671	0.1078	0.8746	0.0967	0.9520
Proactive	OmniGuard (Zhang et al., 2025)	0.4099	0.9807	0.0508	0.8252	0.0174	0.5619	0.0243	0.6329	0.5486	0.9978
Proactive	RecoverMark (An et al., 2026b)	0.5489	0.9641	0.5177	0.9685	0.7519	0.9942	0.7870	0.9933	0.9746	0.9963
Proactive	SyndromeMark (Ours)	0.5538	0.9644	0.5220	0.9899	0.4377	0.9028	0.4350	0.9901	0.5057	0.9874

Method	Reconstruction									
	Patch Remove		Gaussian Blur		Gaussian Noise		JPEG		No Distortion	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Imuge+ (Ying et al., 2023)	14.04	0.0140	12.11	0.0150	10.22	0.0191	10.87	0.0164	10.95	0.0177
RecoverMark (An et al., 2026b)	18.86	0.1448	22.48	0.1428	22.86	0.1413	22.38	0.1417	29.10	0.0169
SyndromeMark (Ours)	23.76	0.1479	24.78	0.1923	25.67	0.1483	29.58	0.0536	32.85	0.0252



(a) F1 Comparison: CSF vs. SyndromeMark (b) Syndrome SNR vs. Tamper Area Ratio

Figure 8: Impact of increasing manipulation area ratios on the localization performance and spatial syndrome SNR under copy-move manipulation.

This is because deepfake manipulations typically involve soft-blended boundaries, making hard-thresholding metrics like F1 highly sensitive to global noise. The high AUC confirms that our syndrome prior consistently preserves the accurate probabilistic ranking of tampered versus untampered pixels. Regarding content reconstruction, SyndromeMark consistently outperforms reconstruction-capable frameworks across different distortions. Under Gaussian noise (Table 4), our method achieves a PSNR of 23.64 dB compared to 18.26 dB for Imuge+. Furthermore, against RecoverMark, which relies on the strict assumption of backgrounds, SyndromeMark achieves superior reconstruction (e.g., under JPEG compression, 29.58 dB vs. 22.38 dB, Table 5).

Empirical Validation of Host Self-Interference and Neural Synergy. To empirically validate our architectural design choices, we examine the localization performance under monotonically increasing manipulation area ratios. As shown in Figure 8, expanding the manipulation area inherently exacerbates internal self-interference, causing the spatial Syndrome SNR to degrade and the pure linear CSF's localization F1-score to decrease significantly. In contrast, SyndromeMark resolves this physical limitation (Curse 1) by employing the neural network as a non-linear solver. By effectively learning the structural differences between host texture leakage and actual malicious tampering, it maintains a robust F1-score (~ 0.93) even at a 50% tamper ratio.

D.3 Robustness Against Adaptive Surrogate Attacks

As mentioned in the threat model in Section 3.2, following Kerckhoffs's principle, we evaluate SyndromeMark against a powerful adversary who knows the entire system architecture and training scheme, lacking only the secret weights of H and G. The adversary mounts a *Surrogate Transfer Attack*, training proxy models ($H_{surrogate}, G_{surrogate}$) to approximate the secret invariant and craft a null-space forgery (i.e., a tailored semantic manipulation e_{forge} strictly satisfying $He_{forge} = 0$) to evade detection. We evaluate this proxy training under two conditions. First, under unconstrained

Table 6: Syndrome Residual Energy under Constrained Adaptive Surrogate Attack (PGD (Madry et al., 2018), $\epsilon = 0.05$).

Image State	Syndrome Residual Energy MSE$(HI - G(m))\downarrow$	
	Attacker $(H_{surrogate}, G_{surrogate})$	Defender (H, G)
Untampered Protected Image (I_{wm})	3.000×10^{-6}	3.000×10^{-6}
Tampered Image (I_{tamper})	2.474×10^{-3}	2.475×10^{-3}
PGD Forged Tampered Image (I_{forged})	1.697×10^{-3}	1.698×10^{-3}

optimization, the neural networks tend to bypass the complex algebraic structure and trivially converge to near-zero weights. They exploit the trivial mathematical shortcut $H_{surrogate} I_{wm} \approx 0$ (where $G_{surrogate}(m) \to 0$), which effectively destroys any valid null space required for embedding realistic forgeries. To mount the strongest possible attack, the adversary then applies the strict orthogonal parameterization ($H_{surrogate} H_{surrogate}^{\top} = I_d$) to train a full-rank, non-collapsing surrogate. Subsequently, they apply a Projected Gradient Descent (PGD) attack (Madry et al., 2018) ($\ell_\infty \leq 0.05$) to optimize an adversarial perturbation δ, aiming to bypass detection locally by minimizing the surrogate residual $\|H_{surrogate}(I_{tamper} + \delta) - G_{surrogate}(m)\|_2^2 \to 0$.

As shown in Table 6, this adaptive attack fails to evade detection. For an untampered protected image, the syndrome residual energy is near-zero (0.000003). Once tampering is applied, this energy drastically spikes to 0.002475. The attacker then uses PGD to hide these tampering traces based on their surrogate models. Although the optimization manages to slightly lower the residual energy on both the surrogate and true models (reducing it to roughly 0.001698), this numerical reduction is practically insignificant. Because the forged residual is still orders of magnitude larger than the safe baseline (0.000003), the defender's system will confidently flag the image as manipulated. In a massive continuous space, the algebraic rules learned by the attacker's surrogate model are entirely different from the true secret keys (i.e., their valid subspaces are disjoint). Therefore, optimizing an image to trick the surrogate model provides no meaningful progress toward closing the massive gap to the true valid subspace, which confirms that without the exact cryptographic H and G, crafting a valid continuous forgery is nearly impossible.

E Limitations & Further Discussion

While CSF and SyndromeMark establish a foundation for proactive image forensics, bridging the gap between rigorous algebraic constraints and the high-dimensional nature of natural images leaves several open challenges for future exploration.

Underdetermined Ambiguity in Large-Area Reconstruction. First, although our framework accurately localizes extensive manipulations, reconstructing these massive tampered areas remains challenging. As formally stated by the Critical Rank Condition (Theorem 3), large-scale edits render the linear system severely underdetermined. While RNet attempts to navigate this infinite null space via a neural network, the synthesized output is fundamentally a data-driven semantic approximation rather than a mathematically unique origin. Although the generated content is visually plausible, it naturally prioritizes semantic coherence over exact algebraic satisfaction. To achieve truly verifiable reconstruction, future research may move beyond feed-forward neural generation and instead integrate continuous syndrome equations as direct guidance scores within the iterative sampling process of diffusion models, ensuring the generated distribution inherently intersects with the target algebraic subspace.

Challenge of Generative Regeneration Attacks. Second, recent advancements in global watermark removal, particularly generative regeneration attacks (Zhao et al., 2024; Liu et al., 2025), pose a formidable challenge to the entire field of proactive image forensics. Because these attacks substantially destroy and resynthesize the pixel array while preserving high-level visual semantics, any defense relying on spatial-domain embedding can be bypassed. Consequently, our current spatial implementation of the algebraic invariant ($HI_{wm} = G(m)$) inherently shares this universal limitation, causing verifiable localization to fail under such semantic-level reconstruction. While

defending against regeneration attacks remains an open problem across the literature, the theoretical framework of CSF provides a clear roadmap for future solutions. By elevating the CSF paradigm from the spatial pixel domain into a robust semantic latent space, future work could establish algebraic invariants that are intrinsically resilient to pixel-level resynthesis.

Trade-off Between Algebraic Rigor and Visual Fidelity. Finally, our empirical results reveal a fundamental trade-off between visual fidelity and the strength of the algebraic prior. Enforcing a more strict continuous syndrome constraint enhances the forensic signal-to-noise ratio against severe global distortions but inevitably introduces noticeable visual artifacts into the host image. Exploring spatially adaptive projection matrices, which dynamically adjust the algebraic constraint weight based on local texture complexity, could optimize this balance and facilitate practical deployment.