FROM GLOBAL VERIFICATION TO FINE-GRAINED TAMPER LOCALIZATION: EXPLOITING SPATIAL RELIABILITY IN ROBUST WATERMARK DECODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust image watermarking has become an effective tool for content authentication, yet existing systems are primarily designed for global message recovery and provide little information about where image integrity has been compromised. In this work, we show that globally trained watermark decoders already contain a previously overlooked, message-conditioned spatial integrity representation. By probing intermediate features of frozen pretrained watermark decoders, we find that local consistency with the authentic embedded message progressively emerges together with message recovery: intact regions retain strong support for the embedded watermark, whereas manipulated regions exhibit degraded local consistency. Counterfactual experiments with shuffled and random messages further show that this spatial separation is specifically tied to the authentic watermark rather than generic editing artifacts or message-independent feature instability. Building on this observation, we introduce a lightweight decoder-side tamper localization framework that extracts spatially distributed watermark evidence from intermediate decoder features, measures its consistency with the embedded message to construct an unreliability map, and combines this active watermark cue with RGB appearance for pixel-level localization. Importantly, the original watermark encoder and global decoder remain entirely frozen, requiring no modification to the deployed watermarking pipeline. Our findings reveal that robust watermark decoders encode substantially richer spatial information than their global objectives explicitly require, providing a new route for extending existing watermark systems from global verification to fine-grained tamper localization.

1 INTRODUCTION

Recent advances in diffusion-based generation and editing have made it increasingly easy to modify localized image content through object removal, replacement, and inpainting while leaving most surrounding regions visually unchanged (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). This capability creates an important distinction between provenance verification and local content integrity: an image may still carry a valid source-related signal even after only a small part of its content has been manipulated. Reliable image authentication therefore requires not only determining whether a trusted signal remains present, but also identifying whether that signal is still spatially consistent with the received content.

Digital watermarking provides a natural mechanism for provenance verification by embedding an imperceptible message into an image before distribution. Classical spread-spectrum methods and modern neural encoder–decoder systems have progressively improved message recovery under compression, noise, geometric distortion, and semantic editing (Cox et al., 1997; Zhu et al., 2018; Tancik et al., 2020; Bui et al., 2025; Hu et al., 2024). Most robust watermarking systems, however, are optimized for a global objective: recovering the embedded bit sequence from the received image. Such an objective determines which message should be decoded, but does not explicitly describe how different image regions support that prediction. A local manipulation may weaken watermark evidence within the edited region while the remaining intact content still provides sufficient evidence for correct global recovery. Consequently, successful message decoding does not necessarily imply that the entire image remains intact.

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.

This observation raises a fundamental question: does global message recovery discard local integrity evidence, or does it merely aggregate and hide it? Existing watermark-based localization methods mainly address spatial authentication by explicitly redesigning watermark training objectives or embedding mechanisms. EditGuard, OmniGuard, StableGuard, Watermark Anything, and MaskWM introduce localization supervision, spatially distributed messages, masked decoding, or dedicated manipulation-aware components to make watermark information locally accessible (Zhang et al., 2024; 2025; Yang et al., 2025; Sander et al., 2025; Hu et al., 2025). These approaches demonstrate the value of spatial watermark information, but leave open a complementary question: whether localization-relevant spatial structure is already present inside watermark decoders that were trained only for global message recovery.

In this work, we show that globally trained robust watermark decoders retain a previously underexplored form of message-conditioned spatial reliability. Although the final decoder output is a single bit sequence, intermediate representations preserve spatially varying support for that message. To study this phenomenon without injecting manipulation supervision into the decoder, we attach lightweight local message probes to frozen pretrained representations and train them only on clean watermarked images and their embedded messages. Across decoder depth, we observe that local message decodability and the separation between manipulated and intact regions emerge together: early representations provide little local message information and little spatial separation, whereas later representations increasingly support the authentic message in intact regions while edited regions exhibit higher disagreement.

We further examine whether this spatial difference is associated with the embedded message rather than merely with a generic manipulation response. Keeping the received image and decoder features fixed, we replace the authentic message with shuffled or independently sampled reference messages. The inside–outside reliability gap largely disappears under these counterfactual references, whereas it remains pronounced for the authentic embedded message. This result is consistent with the observed cue being conditioned on the watermark message, although it does not exclude the presence of visual editing artifacts in the decoder representation. Taken together, the layer-wise evolution and reference-message controls suggest that robust watermark decoding preserves spatially varying support for the authentic message that is concealed by the final global prediction.

Based on this observation, we introduce a decoder-side tamper localization framework that exposes and exploits this hidden spatial reliability without redesigning the original watermark embedding mechanism. A local message head converts an intermediate decoder feature into spatial watermark predictions, and their disagreement with the trusted embedded message forms an explicit unreliability map. Because this watermark-derived map is spatially coarse, a refinement network combines it with RGB appearance to recover object boundaries and predict a dense tamper mask. The original watermark encoder and global decoder remain frozen throughout, so the existing message-recovery pathway is preserved. Importantly, the clean-only probes described above are used only to diagnose what is already present in the frozen decoder; the final localization branch is separately trained with manipulated images and tamper annotations to exploit this cue for dense prediction.

Experiments on RegionBase editing, Stable Diffusion inpainting, and image Splicing show that the resulting framework achieves F1 scores of 0.955, 0.962, and 0.896, respectively, while retaining the original watermark pathway. Component ablations further show that explicit watermark reliability provides complementary information beyond RGB appearance and raw decoder features, and experiments with an additional watermark backbone indicate that the proposed decoder-side construction is not tied to a single pretrained architecture.

Our contributions are summarized as follows:

- We uncover message-conditioned spatial reliability in intermediate representations of robust watermark decoders trained for global message recovery, revealing that global decoding can hide substantial spatial variation in local watermark support.
- We systematically characterize this phenomenon through clean-only layer-wise probing and counterfactual reference-message analysis, showing that local message decodability and manipulated–intact spatial separation emerge together across decoder depth.
- We translate this observation into a decoder-side tamper localization framework that explicitly converts local message disagreement into a reliability map and combines it with RGB refinement, while keeping the original watermark encoder and global decoder frozen.

2 Related Work

Robust watermarking and global message recovery. Digital watermarking embeds imperceptible signals for ownership protection, provenance verification, and content authentication. Classical spread-spectrum methods established the foundations of robust watermarking, while neural encoder–decoder systems improved recovery under diverse distortions (Cox et al., 1997; Zhu et al., 2018). StegaStamp, PIMoG, MBRS, CIN, and RivaGAN improve robustness to physical capture, compression, and learned transformations (Tancik et al., 2020; Fang et al., 2022; Jia et al., 2021; Ma et al., 2022; Zhang et al., 2019), while TrustMark and Robust-Wide further address diverse transformations and semantic editing (Bui et al., 2025; Hu et al., 2024). Stable Signature integrates watermarking into a diffusion decoder (Fernandez et al., 2023). Despite different designs, these methods mainly assess watermark success through image-level recovery or verification. In contrast, we study whether globally trained decoders retain spatially varying support for the embedded message that is hidden by the final global prediction.

Tamper localization and spatial watermarking. Manipulation localization can be addressed passively from image artifacts or proactively using pre-embedded signals. Passive methods such as CAT-Net, PSCC-Net, MVSS-Net, IML-ViT, and HiFi-Net infer tampered regions from content, noise, compression, or representation inconsistencies (Dong et al., 2022; Liu et al., 2022; Kwon et al., 2021; Ma et al., 2023; Guo et al., 2023). Fragile watermarking and SepMark instead use embedded signals for authentication, localization, recovery, or source tracing (Cheng et al., 2012; NR & Shreelekshmi, 2022; Lin et al., 2023; Wu et al., 2023). More recent methods, including EditGuard, OmniGuard, StableGuard, Watermark Anything, and MaskWM, explicitly introduce spatial authentication through localization objectives, localized messages, or masked decoding (Zhang et al., 2024; 2025; Yang et al., 2025; Sander et al., 2025; Hu et al., 2025). Our setting is complementary: rather than redesigning the watermark model for localization, we investigate whether localization-relevant spatial structure already exists in decoders trained only for global message recovery.

Representation probing and spatial evidence. Intermediate representations can contain information not exposed by a model's final output. CAM and Grad-CAM reveal spatial evidence in recognition networks (Zhou et al., 2016; Selvaraju et al., 2017), while probing methods examine what can be decoded from hidden representations (Alain & Bengio, 2016). Because probe capacity and supervision can affect interpretation (Hewitt & Liang, 2019; Belinkov, 2022), our diagnostic probes are trained only on clean watermarked images and embedded messages, without manipulated images or tamper masks. We further compare disagreement with the authentic message against shuffled or random references while keeping image features fixed. This allows us to test whether the observed spatial variation reflects local support for the embedded watermark message rather than only a generic response to manipulation.

3 Method

3.1 Overview

Given an original image I_{ori} and an L-bit message $\mathbf{w} \in \{0,1\}^L$, a pretrained watermark encoder produces a watermarked image I_{wm}. After local manipulation and optional benign degradation, the verifier receives I_{rec}.

As illustrated in Fig. 1, our framework preserves the original watermark pathway while adding a decoder-side localization branch. The pretrained encoder and global decoder remain frozen. From an intermediate decoder feature $F^{(l)}$, we first examine whether local support for the embedded message is already spatially structured, and then convert this support into an explicit unreliability map for tamper localization.

3.2 Probing Spatial Message Support

To determine whether globally trained watermark decoders already retain local message information, we attach lightweight probes to selected frozen decoder stages. Each probe predicts the same embedded message at every spatial location and is trained only on clean watermarked images and

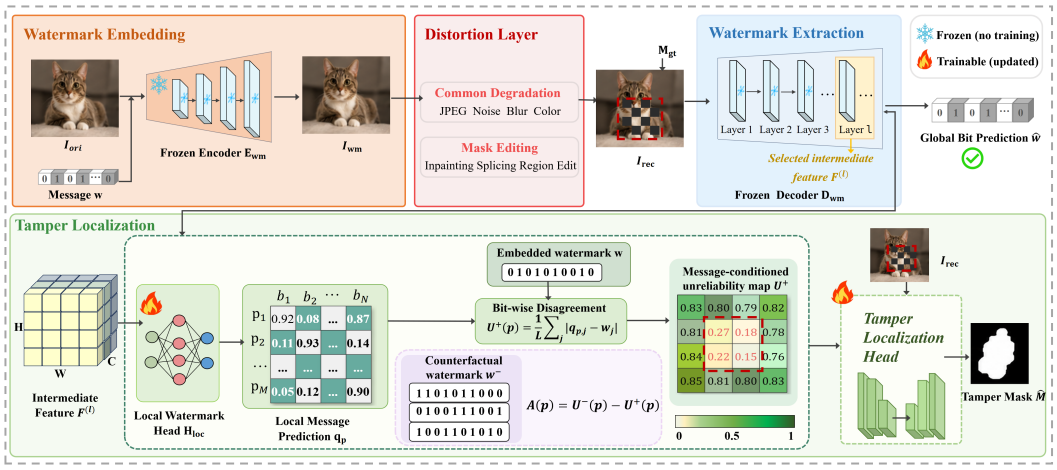

Figure 1: The overall framework of our proposed method.

their messages, without manipulated images or tamper masks. The probes therefore test what is already readable from the pretrained representation rather than learning manipulation-specific cues.

For a local message prediction $\mathbf{q}_p$ and a reference message $\mathbf{v}$, we define the local unreliability as

$$U(p, \mathbf{v}) = \frac{1}{L} \sum_{j=1}^{L} |q_{p,j} - v_j|. \tag{1}$$

For the authentic message $\mathbf{w}$, a smaller $U(p, \mathbf{w})$ indicates stronger local watermark support. We measure spatial separation by the difference between mean unreliability inside and outside the manipulated region.

To examine message dependence, we keep the image and decoder predictions fixed while replacing $\mathbf{w}$ with shuffled or random reference messages. The resulting comparisons test whether the observed spatial variation is associated with support for the authentic embedded message rather than only with a generic response to manipulation.

3.3 Message-Conditioned Tamper Localization

The probing analysis motivates explicitly extracting local watermark reliability from an intermediate decoder feature. A lightweight local message head predicts a soft L-bit message at each spatial location. Comparing these predictions with the trusted embedded message using Eq. 1 produces a single-channel unreliability map U^+.

Unlike the global watermark prediction, U^+ preserves where support for the authentic message becomes weak. Since decoder features are spatially coarse, we upsample this map and combine it with the received RGB image:

$$\hat{M} = S_\phi \left([I_{\text{rec}}; \text{sg}(\text{Up}(U^+))] \right), \tag{2}$$

where $\text{Up}(\cdot)$ denotes spatial upsampling and $\text{sg}(\cdot)$ stops gradients. The reliability map provides a watermark-derived integrity cue, while RGB appearance supplies fine-grained texture and boundary information.

The stop-gradient operation prevents mask supervision from directly turning the reliability extractor into an unconstrained segmentation feature. The local message head is instead optimized through its dedicated message-and reliability-related objectives.

Table 1: Comparison of tamper localization performance under different local editing operations. F1 and AUC are reported at the pixel level. All methods are evaluated on the same 500-image test sets. **Bold** and underlined values denote the best and second-best results, respectively.

	RegionBase		SD Inpaint		Splicing	
Method	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Passive manipulation localization						
ManTra-Net(CVPR 2019)	0.089	0.639	0.207	0.719	0.396	0.679
CAT-Net(WACV 2021)	0.143	0.623	0.243	0.709	0.345	0.674
MVSS-Net(ACM MM 2022)	0.315	0.663	0.404	0.649	0.387	0.550
PSCC-Net(TCSVT 2022)	0.401	0.655	0.523	0.732	0.626	0.736
HiFi-Net(CVPR 2023)	0.332	0.625	0.485	0.701	0.356	0.720
Active watermark-based localization						
EditGuard(CVPR 2024)	<u>0.954</u>	**0.996**	<u>0.953</u>	0.992	**0.952**	<u>0.981</u>
OmniGuard(CVPR 2025)	0.734	0.942	0.802	<u>0.955</u>	0.751	0.930
TAG-WM(ICCV 2025)	0.821	0.913	0.817	0.912	0.812	0.901
LIDMark(CVPR 2026)	0.860	0.895	0.765	0.856	0.736	0.846
Ours	**0.955**	<u>0.975</u>	**0.962**	**0.992**	<u>0.896</u>	**0.990**

3.4 Training and Inference

During localization training, the pretrained watermark encoder and global decoder remain frozen, while only the local message head and localization network are optimized. Unlike the clean-only probing stage, this stage uses manipulated images and their ground-truth masks.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \lambda_1 \mathcal{L}_{\text{msg}} + \lambda_2 \mathcal{L}_{\text{gap}} + \lambda_3 \mathcal{L}_{\text{align}} + \lambda_4 \mathcal{L}_{\text{neg}}. \tag{3}$$

Here, $\mathcal{L}_{\text{msg}}$ preserves local agreement with the authentic message on clean and intact regions, $\mathcal{L}_{\text{gap}}$ and $\mathcal{L}_{\text{align}}$ encourage the unreliability response to separate and align with manipulated regions, and $\mathcal{L}_{\text{loc}}$ supervises the final mask. $\mathcal{L}_{\text{neg}}$ suppresses false responses under clean and benignly degraded inputs.

At inference, only the received image I_{rec} and the trusted embedded message $\mathbf{w}$ are required. The frozen decoder provides the global watermark prediction and intermediate feature, from which the added branch constructs U^+ and predicts $\hat{M}$. Counterfactual references, ground-truth masks, and diagnostic probes at other decoder stages are not required.

4 Experiments

Our experiments investigate three questions: whether frozen watermark decoders retain spatially varying support for the embedded message, whether this information can support tamper localization while preserving global message recovery, and how individual components contribute to localization performance. We first analyze spatial reliability across decoder stages, then evaluate localization and watermark performance, and finally present component ablations and an additional-backbone experiment.

4.1 Experimental Setup

Data and manipulations. We use the RegionBase subset of UltraEdit (Zhao et al., 2024), with 22,104 candidate samples for training and validation. We reserve 250 samples for validation and use the remaining 21,854 for training. Final evaluation uses 500 images from disjoint shards for each manipulation setting. Images and masks are resized to 512×512.

Given an edited image I_e, a watermarked image I_w, and a binary manipulation mask M, we construct the manipulated image as

$$I_t = M \odot I_e + (1 - M) \odot I_w, \tag{4}$$

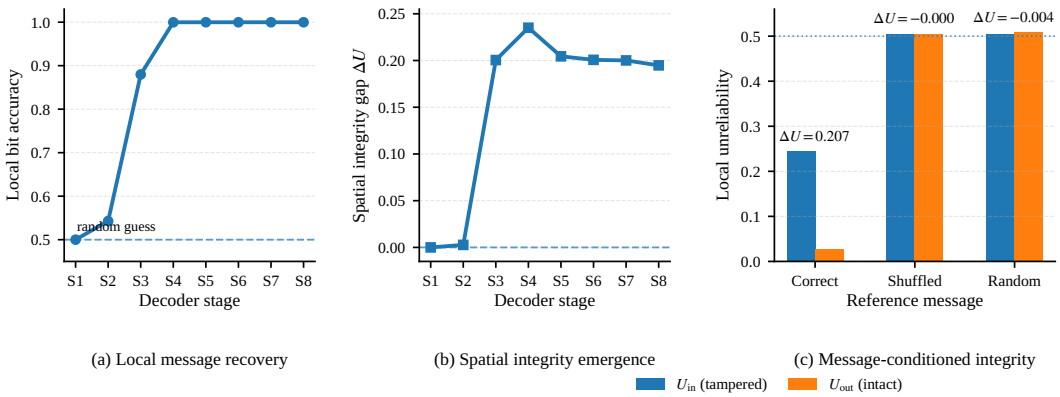

Figure 2: Emergence of message-conditioned spatial reliability in frozen watermark decoder representations. (a) Local message recovery across decoder stages. (b) Manipulated–intact unreliability gap across stages. (c) The gap is pronounced for the authentic message but approaches zero for shuffled and random references.

where $M = 1$ denotes the replaced region. This construction preserves the watermarked content outside the manipulation mask before optional degradation. The localization branch is trained on RegionBase and additionally evaluated on Stable Diffusion v1.5 inpainting and image splicing to examine transfer across manipulation processes under this compositing protocol.

Training and evaluation. The main experiments use a frozen 64-bit Robust-Wide backbone (Hu et al., 2024). Only the local message head and localization network are optimized. Localization training includes JPEG compression, resizing, Gaussian blur, and clean negative samples. We report pixel-level F1 and AUC for tamper localization, PSNR and SSIM for watermarked-image fidelity, and bit accuracy for global message recovery. Model selection and localization thresholds are determined using the validation set.

Implementation details. SD v1.5 inpainting uses 10 inference steps, strength 1.0, and a guidance scale of 7.5. Training degradations include JPEG compression with quality 60–90, resizing with scale factors 0.5–0.9, and Gaussian blur with $\sigma = 0.5$–1.5. The localization network and local message head use learning rates of 5×10^{-5} and 2×10^{-5}, respectively. The loss weights in Eq. 3 are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.5, 0.2, 0.1, 0.25)$.

4.2 Spatial Reliability in Frozen Decoder Features

We first examine whether local message support can reveal spatial integrity differences without tamper supervision. For this diagnostic experiment, lightweight probes are attached to frozen decoder stages and trained only on clean watermarked images and their embedded messages. This setting is distinct from the supervised training of the final localization branch.

Figure 2(a) shows that local bit accuracy is close to random guessing at early stages and increases as decoding progresses. Figure 2(b) reveals a corresponding emergence of spatial separation: the inside–outside unreliability gap is small at early stages and becomes pronounced once the message is locally decodable. The gap peaks around S4 and decreases slightly at subsequent stages, while local bit accuracy remains near saturation. Thus, accurate local message recovery and strong spatial discrimination are related but do not necessarily favor the same decoder stage.

We further examine the dependence of this separation on the reference message. As shown in Fig. 2(c), the authentic message produces an inside–outside gap of 0.207, whereas shuffled and random references yield gaps close to zero. Since the decoder predictions remain fixed, this comparison indicates that the measured separation depends on agreement with the authentic message. Random binary references also have an expected disagreement of 0.5 with any fixed soft prediction, so this control should be interpreted together with the clean-only probing results rather than as independent proof that visual editing artifacts play no role.

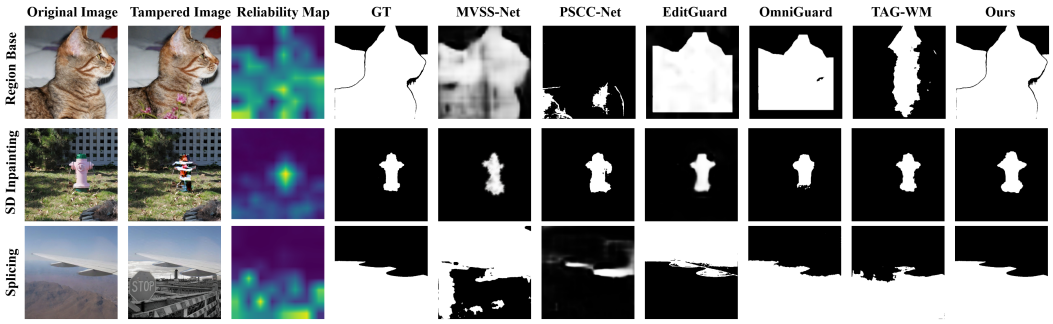

Figure 3: Qualitative comparison under RegionBase editing, SD inpainting, and Splicing (top to bottom). Columns show the original image, tampered image, message-conditioned watermark map, ground-truth mask, and predictions from MVSS-Net, PSCC-Net, EditGuard, OmniGuard, TAG-WM, and our method.

4.3 Tamper Localization

Quantitative comparison. Table 1 compares our framework with passive forensic methods and active watermark-based localization methods on the reported 500-image test sets. Our method achieves F1/AUC scores of $0.955/0.975$, $0.962/0.992$, and $0.896/0.990$ on RegionBase, SD inpainting, and Splicing, respectively. These results show that the added decoder-side branch supports accurate localization across the evaluated manipulation settings.

Compared with the passive methods in the table, our framework obtains higher F1 and AUC in all three settings. Among active methods, it achieves the highest reported F1 on RegionBase and SD inpainting, exceeding EditGuard by 0.001 and 0.009, respectively. The small RegionBase difference indicates comparable performance rather than a substantial advantage. On Splicing, EditGuard achieves a higher F1 of 0.952, compared with our 0.896, while our method obtains the highest AUC of 0.990. This difference highlights that threshold-independent score ranking and the quality of a thresholded mask need not produce the same ordering of methods.

The SD inpainting and Splicing evaluations also examine transfer beyond the training manipulation process. Despite being trained on RegionBase, the localization branch maintains strong performance on SD inpainting and remains effective on Splicing. These results suggest that its utility extends beyond the training editor under the shared compositing protocol, although the lower splicing F1 shows that performance still depends on the manipulation type.

Qualitative comparison. Figure 3 presents representative examples of RegionBase editing, SD inpainting, and Splicing. The watermark-derived maps exhibit spatially coarse responses near manipulated content, whereas the final predictions provide more clearly delineated regions. For example, the SD inpainting case shows a concentrated response around the edited object, followed by a more detailed object-shaped prediction after refinement. In the splicing example, the final mask captures the broad replaced region more continuously than the coarse watermark map.

These examples illustrate the complementary roles of watermark reliability and RGB appearance. Local message disagreement provides a spatial cue for weakened watermark support, while RGB refinement helps recover region structure and boundaries. The coarse maps also contain responses outside the target regions, indicating that they should be treated as localization cues rather than final tamper masks.

4.4 Preserving Global Watermark Functionality

The proposed localization branch keeps the watermark encoder and global decoder frozen and does not feed its predictions back into the original decoding pathway. Table 2 reports watermarked-image fidelity and global bit recovery under local editing and common degradations.

The 64-bit backbone achieves a PSNR of 42.26 dB and an SSIM of 0.993. Global bit accuracy remains 97.9%, 97.5%, and 97.1% under RegionBase editing, SD inpainting, and Splicing, re-

Table 2: Watermarked-image fidelity and global bit accuracy. The parenthesized number denotes payload length. Payloads and watermark designs differ across methods.

	Image Quality		Local Editing			Common Degradation			
Method	PSNR ↑	SSIM ↑	RegionBase	SD Inpaint	Splicing	JPEG	Color Jitter	G. Blur	G. Noise
PIMoG(ACM MM 2022)	38.09	0.981	0.862	0.826	0.849	0.871	0.994	0.997	0.997
CIN(ACM MM 2022)	33.01	0.912	0.884	0.877	0.833	0.865	0.982	0.993	0.995
SepMark(ACM MM 2023)	36.02	0.942	0.928	0.937	0.924	0.994	0.999	0.999	0.998
StegaStamp(CVPR 2020)	29.75	0.910	0.845	0.810	0.796	0.973	0.910	0.975	0.872
TrustMark	40.73	0.991	0.803	0.810	0.796	0.973	0.980	0.998	0.995
StableGuard(NeurIPS 2025)	23.97	0.745	0.978	0.975	0.970	0.504	0.896	0.505	0.845
EditGuard(CVPR 2024)	37.61	0.941	0.952	0.954	0.940	0.997	0.995	0.997	0.995
OmniGuard(CVPR 2025)	40.21	0.992	0.942	0.965	0.952	0.998	0.996	0.997	0.997
TAG-WM(ICCV 2025)	40.21	0.992	0.973	0.971	0.972	0.998	0.983	0.992	0.991
WAM(ICLR 2025)	39.65	0.983	0.978	0.973	0.965	0.995	0.993	0.995	0.997
AdvMark(CVPR 2026)	39.54	0.989	0.975	0.974	**0.975**	0.996	0.999	0.997	0.998
Ours	**42.26**	**0.993**	**0.979**	**0.975**	0.971	**0.999**	**0.999**	**0.999**	**0.999**

spectively. Under the evaluated JPEG compression, color jitter, Gaussian blur, and Gaussian noise settings, bit accuracy is 99.9% at the reported precision.

Together with the localization results, these measurements illustrate that successful global message recovery can coexist with detectable local disruption. Most message bits remain recoverable from the received image, while the added branch identifies regions with weakened local support. This distinction motivates examining spatial reliability in addition to the final image-level prediction.

The fidelity and recovery results characterize the underlying watermark backbone rather than improvements introduced by the localization branch. For the same received image and unchanged inference settings, the frozen global pathway retains its original output. Cross-method comparisons should also account for differences in payload length and embedding design.

4.5 Ablation and Backbone Generalization

Component analysis. Table 3 examines the contribution of the localization inputs and gradient separation. RGB alone achieves an F1 of 0.86, while reliability alone reaches 0.72. Combining the two in the full model increases F1 to 0.95, an absolute gain of 0.09 over RGB alone. The weaker reliability-only result is consistent with the coarse spatial resolution of the watermark cue, while the improvement from fusion supports its complementary value for RGB-based localization.

Directly combining RGB with raw decoder features achieves an F1 of 0.88, compared with 0.95 for explicit message-conditioned reliability. Under the reported configurations, converting decoder features into disagreement with the trusted message therefore provides a more useful input than direct feature fusion. This comparison supports the proposed representation, although its effect should be distinguished from possible differences in branch capacity and optimization.

Removing stop-gradient reduces F1 from 0.95 to 0.87, while AUC remains 0.97 at the displayed precision. The observed benefit is therefore more pronounced for the thresholded localization result than for AUC. This supports separating final mask optimization from reliability extraction. The separation specifically blocks gradients from the final localization loss; the local message head still receives its dedicated message and reliability supervision.

Additional watermark backbone. To examine applicability beyond the main backbone, we additionally adapt the localization branch to the StableGuard configuration reported in Table 4, while keeping its watermark encoder and global decoder frozen. This configuration achieves an F1 of 0.918, an IoU of 0.863, and an AUC of 0.988 on RegionBase.

This experiment evaluates whether the same decoder-side construction can be instantiated with another watermark architecture and message length. The localization branches are adapted separately, so the comparison concerns architectural applicability rather than zero-shot transfer of a single trained branch.

Table 3: Ablation study of the key components in our decoder-side tamper localization framework. The full model combines message-conditioned watermark reliability with RGB appearance while preventing the localization gradient from altering the reliability extractor.

Variant	Components				Localization	
	RGB	Reliability	Decoder Feature	Stop-Grad	F1 ↑	AUC ↑
RGB only	✓	–	–	–	0.86	0.95
Reliability only	–	✓	–	✓	0.72	0.90
RGB + Decoder Feature	✓	–	✓	–	0.88	0.97
w/o Stop-Grad	✓	✓	–	–	0.87	0.97
Full	✓	✓	–	✓	**0.95**	**0.97**

Table 4: Decoder-side localization with an additional 100-bit watermark backbone on RegionBase. The localization branch is adapted to this backbone; its original watermark encoder and global decoder remain frozen.

Backbone	Payload	F1 ↑	IoU ↑	AUC ↑
StableGuard	100	0.918	0.863	0.988
Robust-Wide	64	0.940	0.935	0.975

5 Conclusion

This work investigates how robust watermark decoders trained for global message recovery can support local integrity analysis. Clean-only probing reveals that local message decodability and manipulated–intact reliability separation emerge at similar decoder stages. Reference-message controls further show that the measured separation depends on agreement with the authentic embedded message. Motivated by these observations, we develop a decoder-side localization framework that converts local message predictions into an explicit unreliability map and combines this map with RGB appearance for dense tamper prediction. The original watermark encoder and global decoder remain frozen throughout training.

Experiments on RegionBase editing, Stable Diffusion inpainting, and Splicing demonstrate the utility of this construction under the evaluated protocol. Component ablations support the complementary roles of watermark-derived reliability and RGB refinement, as well as the benefit of separating final mask optimization from reliability extraction. These findings suggest that spatial message support provides useful integrity information beyond a global watermark prediction. The current framework assumes access to a trusted embedded message; extending its evaluation to unrestricted editing and more challenging benign transformations remains an important direction for future work.

AI Assistance Disclosure

During the preparation of this manuscript, the authors used large language models (LLMs) for limited assistance in literature exploration and manuscript refinement. Specifically, AI tools were used to improve writing clarity and readability, and to assist in identifying potentially relevant research works. All technical decisions, methodological designs, experimental implementations, and scientific conclusions were independently developed and verified by the authors. The authors take full responsibility for the content of this manuscript.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

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

Baotian Cheng, Rongrong Ni, and Yao Zhao. A refining localization watermarking for image tamper detection and recovery. In *2012 IEEE 11th International Conference on Signal Processing*, volume 2, pp. 984–988, 2012.

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

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.

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.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.15435.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3155–3165, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

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

Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Mask image watermarking. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.12739.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 41–49, 2021.

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *2021 IEEE winter conference on applications of computer vision (WACV)*, pp. 375–384. IEEE, 2021.

Chia-Chen Lin, Ting-Lin Lee, Ya-Fen Chang, Pei-Feng Shiu, and Bohan Zhang. Fragile watermarking for tamper localization and self-recovery based on ambtc and vq. *Electronics*, 12(2):415, 2023.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1532–1542, 2022.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Xia Du, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

Neena Raj NR and R Shreelekshmi. Fragile watermarking scheme for tamper localization in images using logistic map and singular value decomposition. *Journal of Visual Communication and Image Representation*, 85:103500, 2022.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 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.

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

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1610.02391.

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(CVPR)*, pp. 2114–2123, 2020.

Xiaoshuai Wu, Xin Liao, and Bo Ou. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1190–1201, 2023.

Haoxin Yang, Bangzhen Liu, Xuemiao Xu, Cheng Xu, Yuyang Yu, Zikai Huang, Yi Wang, and Shengfeng He. Stableguard: Towards unified copyright protection and tamper localization in latent diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *ArXiv*, 2019.

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. 11964–11973, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024. URL https://arxiv.org/abs/2407.05282.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://arxiv.org/abs/1512.04150.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *European conference on computer vision(ECCV)*, pp. 682–697, 2018.