International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 1, Issue 2 December 2024 ISSN: 3048-975X
https://ijrasht.com/

AI Watermark Robustness: Evaluating Methods to Identify AI-Generated Content After Editing or Compression

Rohit Kumar, Department of CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, Haryana, India. rohitbhukal01@gmail.com

Swarna Verma, Department of CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, Haryana, India. swarnaverma373@gmail.com

Deepanshu Mittal, Department CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, Haryana, India. deepanshumittal1006@gmail.com

Armaan Singh, Department of CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, Haryana, India. armaansingh8325@gmail.com

Abstract – The explosion of Generative Artificial Intelligence (AI) systems that produce photorealistic images, natural-sounding audio, coherent video and fluent text has intensified the demand for dependable mechanisms to trace data back to its source to establish a provenance of synthetic content. One of the most used technical solutions to this requirement is digital watermarking, which consists of adding a slight and imperceptible signal to media produced. In the real world, however, very little content can be presented to a verifier without some alterations: it is often cropped, re-sized, re-encoded, filtered, paraphrased, or deliberately attacked to remove identifying signals. In this paper, we provide a systematic survey of image, video, text, and audio AI watermarking methods, focusing on their resilience to post-generation editing and lossy compression. We categorise approaches into two types of embedding: post-hoc and in-generation and review the threat models and evaluation metrics employed to assess robustness and then compare representative techniques such as hybrids in the frequency domain, learned encoder-decoder-based approaches, diffusion process watermarks, and token-level language model watermarks. We also explore attacks, such as adversarial removal, regeneration and forgery attacks, which are specifically targeting the watermark persistence, and finally collate benchmark results that show a persistent trade-off between robustness and quality/capacity. It ends with open problems including the lack of standard robustness measures, vulnerability to surrogate generative attacks, and the absence of cross-modal or multilingual robustness guarantees, as well as some further directions that are promising, such as semantically based watermarking and certified robustness frameworks.

Keywords: AI-generated content, digital watermarking, robustness evaluation, content provenance, deepfake detection, diffusion models, large language models, adversarial attacks, image compression, and watermark removal.

I. INTRODUCTION

The past several years has seen generative artificial intelligence models — diffusion models for images and video and large language models (LLMs) for text — reach a level of fidelity at which synthetic media is often indistinguishable from human-created content. This capability has driven legitimate creative and productivity gains, but it has also raised concerns about misinformation, non-consensual media, academic dishonesty, and erosion of trust in digital content [1], [4]. In response, technology providers and researchers have converged on digital watermarking as a practical, scalable mechanism for provenance tracking: an imperceptible signal is embedded into generated content at, or shortly after, creation time, and a corresponding detector later recovers that signal to confirm the content's synthetic origin [2], [7].

Embedding a watermark that survives generation is only half the problem. In practice, content is rarely consumed exactly as generated. Images are re-compressed for web delivery, cropped and resized for social media, filtered, or partially in painted; video is transcoded across multiple codecs and bitrates; text is copy-pasted, paraphrased, translated, and mixed with human-written prose; audio is resampled, denoised, or re-recorded. Any of these transformations can degrade or destroy an embedded watermark, whether accidentally through routine processing or deliberately through an adversary's removal attack [4], [18].

Robustness — the capacity of a watermark to remain detectable after such transformations — is therefore the central technical challenge that determines whether watermarking can serve as a dependable content-provenance tool in adversarial, real-world settings.

This paper reviews the current landscape of AI watermarking research through the specific lens of robustness to editing and compression. Section II introduces a taxonomy of watermarking approaches. Section III formalizes the threat models and metrics used to evaluate robustness. Sections IV through VI examine robustness findings for image/video, text, and audio watermarks, respectively. Section VII discusses cross-cutting adversarial removal and forgery attacks. Section VIII presents a comparative synthesis, and Section IX outlines open challenges before Section X concludes.

II. TAXONOMY OF WATERMARKING METHODS

Watermarking techniques for AI-generated content can be organised along two orthogonal dimensions: the stage at which the watermark is embedded and the modality of the content being protected.

A. Post-hoc (Post-processing) Watermarking

Post-hoc methods embed a signal into content after it has been generated, treating the generator as a black box. Classical transform-domain techniques such as DWT-DCT-SVD hybrids modify coefficients in the wavelet, discrete cosine, or singular-value domains to hide a payload while limiting perceptual distortion [3], [6]. More recent learned approaches, including encoder-decoder architectures such as HiDDeN and CNN-based steganographic schemes such as StegaStamp, train and embedding network and an extraction network jointly against a simulated distribution of distortions, which tends to improve robustness to the specific perturbations seen during training [19]. Post-hoc watermarking is attractive because it is model-agnostic and can be layered onto any generation pipeline, but it can introduce visible artefacts, and its robustness is bounded by the diversity of distortions used during training [7].

B. In-Generation (In-Processing) Watermarking

In-processing methods instead alter the generative process itself so that every output is intrinsically watermarked. For diffusion models, Tree-Ring watermarking embeds a structured pattern into the initial noise vector in Fourier space, which is later recovered by inverting the diffusion sampling process; because the signal lives in the latent noise rather than the pixel space, it is largely invariant to cropping, rotation, and other geometric edits [17]. The Stable Signature approach instead fine-tunes the latent decoder of a diffusion model to imprint a binary signature that a pretrained extractor can recover even after common post-processing [18]. Successors such as RingID and ShapeMark extend these ideas to support multiple keys and better preserve output diversity while retaining robustness [21], [23]. For text, in-processing watermarking is dominant: the sampling distribution over the next token is biased — for example by partitioning the vocabulary into pseudo-random "green" and "red" lists — so that watermarked text is statistically detectable without needing model access at detection time [24], [25].

C. Modalities: Image, Video, Text, and Audio

Although the underlying goal is the same, robustness characteristics differ sharply by modality. Image watermarks must survive compression codecs (JPEG, WebP), resizing, and cropping [7]. Video watermarks face the additional challenge of temporal consistency across frames and robustness to lossy video codecs and frame-rate conversion [5]. Text watermarks must survive semantic-preserving edits such as paraphrasing, synonym substitution, and translation, which have no direct analogue in the pixel domain [11], [16]. Audio watermarks must tolerate resampling, format conversion, noise addition, and re-recording (the so-called "over-the-air" channel) [8].

III. THREAT MODEL AND ROBUSTNESS METRICS

A. Non-Malicious Distortions

The mildest and most common threat is ordinary content processing that a legitimate user performs for unrelated reasons: JPEG or video re-compression at reduced bitrate, resizing for a target display, cropping to an aspect ratio, color or brightness adjustment, and, for text, minor copy-editing [3], [6]. A watermark that cannot survive this class of transformation is unlikely to be useful in practice, since almost all AI-generated content passes through at least one such operation before reaching an end viewer or reader.

B. Malicious Removal, Regeneration, and Forgery Attacks

A more robust threat model is for an adversary actively attempting to remove or spoof a watermark. The removal attacks involve malicious perturbations designed for a known or surrogate detector, for example WEvade which perturbs an image generated by the detector, such that the returned message is different from the original watermark [4]. Regeneration attacks (also known as "diffusion purification" [4]) inject noise into a watermarked image and then reconstruct the image using a different generative model, but this yields a result which is also a resampling of the content, but one that obscures the watermarking latent-space signal. In the case of text, attacks that involve paraphrasing the text, such as automated paraphrasers like DIPPER, back-translation, and synonym substitution, are the most common removal strategies, as they change the statistics of the tokens used in the text that many LLM watermarks rely on [11, 14, 16, 24]. The forgery attacks, on the other hand, aim to cause misclassification of unwatermarked or human-written content as being AI-generated, which can be similarly damaging to the trust in a detection system [4], [19].

C. Evaluation Metrics

Robustness is typically quantified using bit accuracy or message recovery rate after a given distortion, together with detection metrics such as true-positive rate at a fixed false-positive rate (TPR@FPR) and the area under the ROC curve (AUROC) [7]. Benchmarks such as WAVES standardize this evaluation by applying a broad suite of non-malicious and adversarial perturbations across multiple watermarking schemes under matched imperceptibility budgets, enabling like-for-like comparison rather than relying on each paper's self-selected distortion set [7].

IV. ROBUSTNESS OF IMAGE AND VIDEO WATERMARKS

Frequency-domain hybrids such as DWT-DCT-SVD remain widely deployed because of their low computational cost and reasonable resilience to compression and noise, but they degrade sharply under geometric attacks such as cropping and rotation and can be defeated by targeted filtering once the embedding domain is known [3], [6]. Learned post-hoc encoder-decoder schemes such as HiDDeN and StegaStamp improve robustness by training against a simulated distortion layer, achieving high accuracy for perturbations represented in training, but they generalize poorly to distortion types not seen during training, including many adversarial and regeneration attacks [19], [20].

In-generation watermarks tend to outperform post-hoc alternatives under common edits. The Stable Signature scheme has been shown to remain detectable after typical image manipulations including cropping and moderate compression, because the signature is embedded through the model's decoder rather than as a post-processing overlay [18]. Tree-Ring watermarking exploits the invariance of Fourier-domain patterns to convolutions, cropping, and rotation, and reports substantially higher robustness than deployed alternatives, though detection generally requires access to the underlying diffusion process for noise inversion, which limits third-party verifiability [17]. Follow-up work such as ONRW and ShapeMark seeks to further optimize the trade-off between image quality, output diversity, and robustness inherent to noise-based watermarking [21], [22]. Independent evaluations caution, however, that even these methods are not immune to regeneration attacks, where an adversary uses a second generative model to reconstruct the image from a noised version, destroying the fine latent structure the watermark depends on [4].

For video, Video Seal jointly trains an embedder and extractor with codec-aware augmentation applied between the two stages during training, which the authors report substantially improves robustness to real-world video re-encoding compared with adapting image watermarks frame-by-frame [5]. This reflects a broader pattern across modalities: robustness improves most reliably when the expected distortion channel is explicitly incorporated into training or design, rather than treated as an afterthought.

V. ROBUSTNESS OF TEXT WATERMARKS

Token-level watermarking schemes, exemplified by green/red-list decoding, bias the LLM's sampling distribution using a pseudo-random function of the preceding token context, then apply a statistical test at detection time to determine whether the proportion of "green" tokens is higher than chance [24]. These schemes are computationally cheap and require no access to model internals at detection time, but their reliance on local n-gram context makes them fragile: because paraphrasing rewrites the surrounding tokens, the green-list assignment for each position is disrupted, and empirical studies report large drops in detection accuracy after even moderate paraphrasing, with the effect worsening as the watermarked context window shrinks [11], [14], [16].

Distortion-free watermarking schemes reformulate the embedding as an unbiased sampling procedure keyed to a pseudo-random sequence, aiming to preserve the exact output distribution of the underlying model while still allowing statistical detection [10], [25]. Google DeepMind's SynthID-Text follows a related tournament-sampling approach and has been adopted at production scale; independent robustness assessments, however, find that it remains vulnerable to meaning-preserving attacks such as paraphrasing, copy-paste rearrangement, and back-translation, with detection F1 scores dropping below acceptable thresholds under these conditions [14], [15]. Comparative benchmarks such as BiMark report that SynthID-family methods nonetheless tend to retain higher post-paraphrase detection accuracy than simple soft red-list or DiPmark baselines across a range of text lengths, though all evaluated schemes lose substantial accuracy as paraphrase strength increases [13].

A more recent line of work targets semantic robustness directly. SemStamp partitions the sentence-embedding space using locality-sensitive hashing so that the watermark signal survives lexical substitution as long as sentence meaning is preserved [16]. SimMark applies a similarity-based, logit-free watermark at the sentence level,

reporting improved robustness to paraphrasing without requiring access to the generating model's internal probabilities [12]. Hybrid schemes such as SynGuard combine SynthID's token-level mechanism with semantic-alignment signals to recover some of the robustness lost to back-translation and synonym substitution [14], [15]. Finally, TextSeal introduces a localized watermark designed to remain detectable even in documents that mix AI-generated and human-written passages, and demonstrates that its signal persists through model distillation, allowing downstream models trained on watermarked outputs to also be identified — a property the authors term "radioactivity" [9].

VI. ROBUSTNESS OF AUDIO WATERMARKS

Audio watermarking for synthetic speech spans spectral/temporal-domain methods, which modify phase or subtly alter phoneme timing, and deep-learning-based encoder-decoder methods analogous to those used for images [3]. A systematic robustness assessment across generative audio systems finds that resampling, format conversion (e.g., MP3 re-encoding), additive noise, and, most damagingly, re-recording through a physical playback-and-microphone channel can substantially degrade watermark recoverability, and that robustness and imperceptibility remain in tension in the same way observed for image and text watermarks [8]. Because voice cloning and synthetic speech are increasingly used in fraud and impersonation, robustness to the "over-the-air" channel is considered a particularly high-priority open problem for this modality [8].

VII. ADVERSARIAL REMOVAL AND FORGERY ATTACKS

Beyond ordinary editing, a growing body of work studies watermarking under an adaptive adversary who has partial or full knowledge of the watermarking scheme. Reconstruction attacks, in which an image is noised and then regenerated using an auxiliary diffusion model, have been shown to remove watermarks from multiple deployed schemes with limited perceptual quality loss, effectively resetting the latent structure the watermark relies on [4]. White-box adversarial attacks such as PTW demonstrate that even schemes considered robust under standard benchmarks can fail against an adaptive attacker with access to the detector, and WEvade shows that a learned adversarial perturbation layer can defeat watermark detectors in both white-box and black-box settings [4]. Related work using surrogate detection keys shows that an attacker who does not have the true key can nonetheless approximate it well enough to mount an effective optimal attack [4]. On the text side, retrieval-based defences have been proposed as a complement to watermarking rather than relying solely on statistical detection. A verifier can retrieve the closest match to a suspect passage from a database of known model outputs, which remains effective even after paraphrasing attacks that defeat purely statistical watermark detectors [24].

VIII. COMPARATIVE SYNTHESIS

Table 1 summarises representative methods across modalities, indicating their embedding stage and qualitative robustness to compression and to editing/attack scenarios, synthesised from the sources reviewed above. Two patterns recur across the literature.

First, in-generation watermarking generally outperforms post-hoc watermarking against both routine compression and moderate editing, because the signal is coupled to structures — decoder weights, initial noise, or the sampling distribution itself — that are harder to disturb without materially changing the content [17], [18].

Second, no method reviewed here is robust against a fully adaptive, white-box adversary or against regeneration-style attacks that resample the content through an independent generative model; robustness claims in the literature should therefore be read as relative to a specified, and often limited, threat model rather than as absolute guarantees [4], [7].

Method	Modality	Embedding Stage	Compression Robustness	Editing/Attack Robustness
DWT-DCT-SVD [6]	Image	Post-hoc	Moderate–High	Low against geometric attacks
HiDDeN [20]	Image	Post-hoc (learned)	Moderate	Moderate; trained-in distortions only
StegaStamp [22]	Image	Post-hoc (learned)	High (trained)	Moderate; fails on unseen attacks

Stable Signature [18]	Image (diffusion)	In-generation (decoder)	High	High for common edits; weak vs. regeneration
Tree-Ring / RingID [17,21]	Image (diffusion)	In-generation (init noise)	High	High for crop/rotate; needs model access to verify
Video Seal [5]	Video	In/post-hoc (trained)	High (codec-aware)	High against re-encoding
Green/Red-list LLM WM [23]	Text	In-generation (decoding)	N/A	Low–Moderate under paraphrasing
SynthID-Text [15,16]	Text	In-generation (sampling)	N/A	Moderate; degrades under back-translation
SemStamp / SimMark [12,17]	Text	In-generation (semantic)	N/A	Moderate–High under paraphrasing
TextSeal [9]	Text	In-generation (localized)	N/A	High; survives dilution and distillation

IX. OPEN CHALLENGES AND FUTURE DIRECTIONS

Several challenges remain unresolved. First, robustness evaluation is fragmented: different papers apply different distortion sets, strengths, and imperceptibility budgets, which makes cross-method comparison unreliable; standardised benchmarks such as WAVES are a step toward addressing this but have not yet been universally adopted [7]. Second, the robustness-quality-capacity trade-off is fundamental rather than incidental — increasing payload capacity or embedding strength to survive stronger edits tends to reduce perceptual quality or increase false-positive rates, and every scheme reviewed here occupies a different point on this trade-off curve [2], [3]. Third, regeneration and diffusion-purification attacks represent a qualitatively different threat than classical signal-processing distortions because they resample rather than merely perturb the content, and current defenses against this class of attack remain limited [4]. Fourth, cross-lingual and cross-modal robustness is underexplored: text watermarks are typically evaluated in English and degrade under translation, and few works evaluate whether a watermark surviving one modality-specific edit (e.g., video re-encoding) also survives edits more typical of another distribution channel (e.g., social media re-compression combined with cropping) [8], [14]. Finally, certified or provable robustness guarantees, analogous to certified adversarial robustness in classification, remain rare; distortion-free and certified watermarking schemes are promising early steps but currently offer guarantees only under restricted attack models [10], [25].

X. CONCLUSION

Watermarking has become the most actively developed technical approach for tracing the provenance of AI-generated images, video, text, and audio, but robustness to real-world editing and compression is not yet a solved problem. This review has organised the landscape into post-hoc and in-generation embedding paradigms, summarised the non-malicious and adversarial threat models used to evaluate robustness, and compared representative methods across modalities. The evidence surveyed indicates that in-generation watermarking, which couples the embedded signal to the generative process itself, currently offers the strongest resilience to routine editing and compression, while remaining susceptible to adaptive, white-box, and regeneration-style attacks. Continued progress will likely depend on standardised, adversarially inclusive benchmarks; semantically grounded embedding strategies; and formal robustness guarantees that hold under clearly specified and realistic threat models.

REFERENCES

[1] X. Zhong, P.-C. Huang, S. Mastorakis, and F. Y. Shih, "An automated and robust image watermarking scheme based on deep neural networks," IEEE Transactions on Multimedia, vol. 23, pp. 1951–1961, 2021.
[2] H. Fang, Z. Jia, Z. Ma, E.-C. Chang, and W. Zhang, "PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network," in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 2267–2275.

[3] X. Luo, R. Zhan, H. Chang, F. Yang, and P. Milanfar, "Distortion agnostic deep watermarking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 13548–13557.

[4] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, and A. Emami, "ReDMark: Framework for residual diffusion watermarking based on deep networks," Expert Systems with Applications, vol. 146, p. 113157, 2020.

[5] Z. Jia, H. Fang, and W. Zhang, "MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression," in Proc. 29th ACM Int. Conf. Multimedia, 2021, pp. 41–49.

[6] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 14448–14457.

[7] N. Yu, V. Skripniuk, D. Chen, L. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[8] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze, "Watermarking images in self-supervised latent spaces," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2022, pp. 3054–3058.

[9] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in Proc. ACM Int. Conf. Multimedia Retrieval (ICMR), 2017, pp. 269–277.

[10] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in Proc. 27th USENIX Security Symp., 2018, pp. 1615–1631.

[11] B. D. Rouhani, H. Chen, and F. Koushanfar, "DeepSigns: An end-to-end watermarking framework for ownership protection of deep neural networks," in Proc. 24th Int. Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2019, pp. 485–497.

[12] F. Boenisch, "A systematic review on model watermarking for neural networks," Frontiers in Big Data, vol. 4, art. 729663, 2021.

[13] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in Proc. 40th Int. Conf. Mach. Learn. (ICML), 2023, pp. 17061–17084.

[14] K. Yoo, W. Ahn, J. Jang, and N. Kwak, "Robust multi-bit natural language watermarking through invariant features," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL), 2023, pp. 2092–2115.

[15] A. Hou et al., "SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation," arXiv:2310.03991, 2023.

[16] M. Christ, S. Gunn, and O. Zamir, "Undetectable watermarks for language models," arXiv preprint arXiv:2306.09194, 2023.

[17] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust," in Proc. NeurIPS, 2023, arXiv:2305.20030.

[18] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The Stable Signature: Rooting Watermarks in Latent Diffusion Models," in Proc. IEEE/CVF ICCV, 2023, pp. 22466–22477, arXiv:2303.15435.

[19] J. Zhu, R. Kaplan, J. Johnson, and F. Li, "HiDDeN: Hiding Data with Deep Networks," arXiv:1807.09937, 2018.

[20] " J. Zhao, X. Wang, and W. Zhang, "Invisible image watermarks are provably removable using generative AI," arXiv preprint arXiv:2306.01953, 2023.

[21] A. Grinbaum and L. Adomaitis, "The ethical need for watermarks in machine-generated language," arXiv preprint arXiv:2209.03118, 2022.

[22] V. Vukotić, V. Chappelier, and T. Furon, "Are deep neural networks good for blind image watermarking?" in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[23] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense," in Proc. NeurIPS, 2023.

[24] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang, "Robust Distortion-Free Watermarks for Language Models," arXiv:2307.15593, 2023.

[25] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang, "Provable Robust Watermarking for AI-Generated Text," arXiv:2306.17439, 2023.