ISSN: 2096-3246
Volume 58, Issue 03, 2026

AI WATERMARKING UNDER ATTACK: A CROSS-MODAL REVIEW AND RESEARCH AGENDA

Sri Venkata Aravindbabu Malempati
California State University, Los Angeles, USA

Abstract
With the thriving of generative artificial intelligence, an unprecedented crisis of content authenticity is becoming real because any type of text, image, audio, and video can no longer be perceived as the work of humans. This is a systematic article on generative watermarking as an active paradigm of authentication that answers four research questions on the technical strategies, robustness, adoption and the future. Key discoveries include the fact that image watermarking has reached a high degree of maturity, with the most advanced systems recording 93% true positive rates and not degrading to under 38 dB compression at imperceptibility (PSNR). Nevertheless, text watermarking is susceptible, falling below 70% accuracy with paraphrasing attacks, and regeneration attacks lower all existing techniques to close to random detection. Regardless of the regulatory requirements on the implementation of watermarking by the European Union and China, only 38 percent of platforms apply verified watermarking. It concludes in the review that watermarking is not a sufficient tool and that multi-layered architectures between watermarking and provenance tracking, cryptographic signatures, and harmonized international standards are necessary to achieve effective content authentication.
Keywords: Generative-AI Watermarking Content Authenticity Deepfake Detection Synthetic Media Provenance AI Regulation

1. Introduction
1.1 The Content Authenticity Crisis
With the advent of generative artificial intelligence, it has completely broken down the historical belief that digital content has a provenance. GPT-4, DALL-E 3, Stable Diffusion, and Midjourney are some of the systems that generate text, imagery, audio, and video of perceptual quality that even trained human judges cannot consistently and reliably discern between synthetic and authentic material. The reaction of the population has been following the same pattern: in 2024 and 2025, research revealed that about 71 percent of the population feels anxious about AI-based scams, and 69 percent of the population feels that deepfake media pose a significant threat to society (Jacob K. et al., 2025). The impacts of this unpredictability are spread across areas that are as vital as electoral integrity, medical misinformation, financial fraud, and judicial evidence, all of which rely on the presumptive accuracy of text and multimedia documents (Patel et al., 2023; Crothers et al., 2023).

1.2 Limitations of Post-hoc Detection
Conventional methodologies of the issue have depended on post-hoc detection, which denotes that AI outputs are discovered once they have been distributed by using identifiers that are trained to discover statistical artifacts of machine generation. This is a demonstrably inadequate reactive paradigm. In January 2023, OpenAI released a machine learning Text Classifier that only scored 26 percent on paraphrased GPT-4 text, but the service was discontinued in half a year, owing to a high false positive rate with actual human text (Crothers et al., 2023). The architecture factor behind this failure is that post-hoc detectors need to make generalizations in an open-ended and quickly changing environment of generative models, fine-tuned models, and adversarially perturbed outputs; a generalization problem that is computationally infeasible as the diversity of models expands (Yang et al., 2025; Crothers et al., 2023).

1.3 Scope and Research Questions
Generative watermarking represents a shift from reactive detection to active authentication. Instead of doing an analysis post-factum, it introduces an implicit signal that can be retrieved by the machine,

but it is added to content at the time it is generated. In this review, four guiding questions are addressed, including (RQ1) what algorithmic solutions are available in text, image, audio, video, and 3D modalities; (RQ2) how resistant are the solutions to adversarial attacks; (RQ3) what is the current adoption and regulatory environment; and (RQ4) what are the most critical research directions?

2. Survey Methodology

2.1 Search Strategy and Databases

The systematic literature search was designed to be comprehensive, reproducible, and appropriately scoped to the period in which generative watermarking emerged as a distinct research area. Searches were conducted across five major databases: IEEE Xplore, the ACM Digital Library, arXiv, MDPI's journal portfolio, and Elsevier ScienceDirect. The primary search strings combined "generative watermarking" and "AI content authentication" as core terms with secondary terms including "SynthID," "large language model watermark," "latent diffusion watermark," and "deepfake detection," connected by Boolean OR and AND operators. All searches were restricted to January 2020 through March 2026, a temporal boundary chosen to align with the public availability of transformer-based generative models as the dominant architecture (Zhong & Shih, 2020; Bistroń et al., 2026).

2.2 Inclusion, Exclusion, and Coding Protocol

The inclusion criteria required that each source address watermarking or detection of AI-generated content, be published in or submitted to a peer-reviewed venue, and make a primarily technical contribution. Technical reports from major AI organizations and regulatory documents from the European Union and China's Cyberspace Administration were included as supplementary grey literature given their direct relevance to RQ3. Works focusing exclusively on traditional steganography, non-generative watermarking, or detection without an associated embedding mechanism were excluded. After removing duplicates from the databases, an initial pool of 47 candidate sources was screened by title and abstract, resulting in 28 full-text reviews. The final corpus of 15 sources was chosen based on methodological rigor, citation impact, and cross-modal representativeness (Kirchenbauer et al., 2023; Fernandez et al., 2023; Wen et al., 2025). The coding protocol captured seven dimensions per source: content modality, embedding mechanism category, primary robustness metric reported, computational overhead, open-source availability, regulatory context, and publication tier. This coding enabled the construction of Tables 1 through 4 and informs the cross-section comparative analysis in Sections 4 through 7 (Cao et al., 2025; Yang et al., 2025; Bistroń et al., 2026).

3. Technical Foundations of Generative Watermarking

3.1 Core Properties

Generative watermarking is defined as the algorithmic process of embedding a statistically detectable signal into content during its synthesis by a machine learning model, satisfying four simultaneously desirable properties: imperceptibility, robustness, detectability, and quality preservation. Imperceptibility requires that the embedded watermark be indistinguishable from unwatermarked content under ordinary conditions, operationalized through Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Fréchet Inception Distance (FID) for visual content, and Mean Opinion Score (MOS) for audio (Bistroń et al., 2026; Wen et al., 2025).

PSNR is formally defined as

$$PSNR = 10.log_{10}\frac{MAX_I^2}{MSE} \quad (1)$$

where MAX_I is the maximum possible pixel value (255 for 8-bit images) and MSE is the mean squared error between the original and watermarked image. Values above 35 dB indicate imperceptibly small distortion in practice; leading systems such as the Stable Signature consistently achieve PSNR between 38 and 42 dB (Fernandez et al., 2023; Bistroń et al., 2026).

3.2 Detection Metrics

Detectability is quantified as True Positive Rate (TPR) at a controlled False Positive Rate (FPR). The FPR at the detection threshold ι is formally expressed as

$$FPR(\iota) = Pr(S \geq \iota | H_O) = \varepsilon) \qquad (2)$$

where S is the detection score, H_0 is the null hypothesis that no watermark is present, and ϵ is the tolerated false alarm rate, typically set to 10^{-3} or lower in deployment settings (Cao et al., 2025). Bitwise Accuracy (BA), which measures the proportion of correctly decoded message bits from the recovered watermark, is defined as:

$$BA(\widehat{\omega}, \omega) = \frac{S(\widehat{\omega}, \omega)}{|\omega|} \qquad (3)$$

where $\widehat{\omega}$ is the decoded watermark, ω is the original embedded message, and $|w|$ is the total bit length (Cao et al., 2025). Values of BA above 0.95 are generally required for reliable multi-bit message recovery (Zhong & Shih, 2020; Bistroń et al., 2026).

3.3 Threat Model Tiers

The technical threat landscape is structured into three tiers. Benign transformations include JPEG compression at quality levels between 70 and 100, bilinear resizing up to 0.5×, and MP3 encoding above 128 kbps, the minimum robustness requirement for any deployable system (Cao et al., 2025; Zhong et al., 2023). Moderate adversaries employ style transfer and regeneration attacks, while sophisticated white-box adversaries have complete knowledge of the watermarking algorithm and can mount gradient-based removal attacks; against these, no current system achieves TPR above 55% at FPR below 1%, revealing a fundamental security boundary (Cao et al., 2025; Luo et al., 2025).

4. RQ1: Technical Approaches by Modality

4.1 Text Watermarking

Figure 1 presents the generative watermarking pipeline across three primary modalities, illustrating how embedding occurs at different representational levels. For text, the foundational contribution of Kirchenbauer et al. (2023) introduced token-level green-red list partitioning: prior to generating each token t_i, a pseudorandom seed derived from the preceding context partitions the vocabulary ν into a green “G” and a red set "R," and the model's logit for each green token is increased by a hardness parameter $\delta > 0$:

$$l'_k = l_k + \delta . 1[k \epsilon G] \qquad (4)$$

The resulting distribution of green tokens in a passage of length “T” is then tested against the null hypothesis of no watermark using a one-sided z-score:

$$z = \frac{|\{t_i : t_i \epsilon\, G\}| - \frac{T}{2}}{\sqrt{\frac{T}{4}}} \qquad (5)$$

ISSN: 2096-3246
Volume 58, Issue 03, 2026

This achieves a TPR of 95.3% at FPR below 1% on texts of 200 tokens or more, with negligible perplexity increase (Kirchenbauer et al., 2023). Google DeepMind's SynthID Text extended this with a pseudorandom g-function and Bayesian detection framework, reporting TPR exceeding 99.1% on unmodified Gemini outputs (Yang et al., 2025). The critical limitation is sensitivity to paraphrasing: at a 40% word substitution rate, Kirchenbauer et al.'s (2023) method degrades to 62.4% TPR, and even SynthID Text falls to approximately 71% (Yang et al., 2025). Cryptographic text watermarking schemes provide theoretical assurances of undetectability; however, they have yet to exhibit regulatory-grade resilience to real-world paraphrasing (Yang et al., 2025).

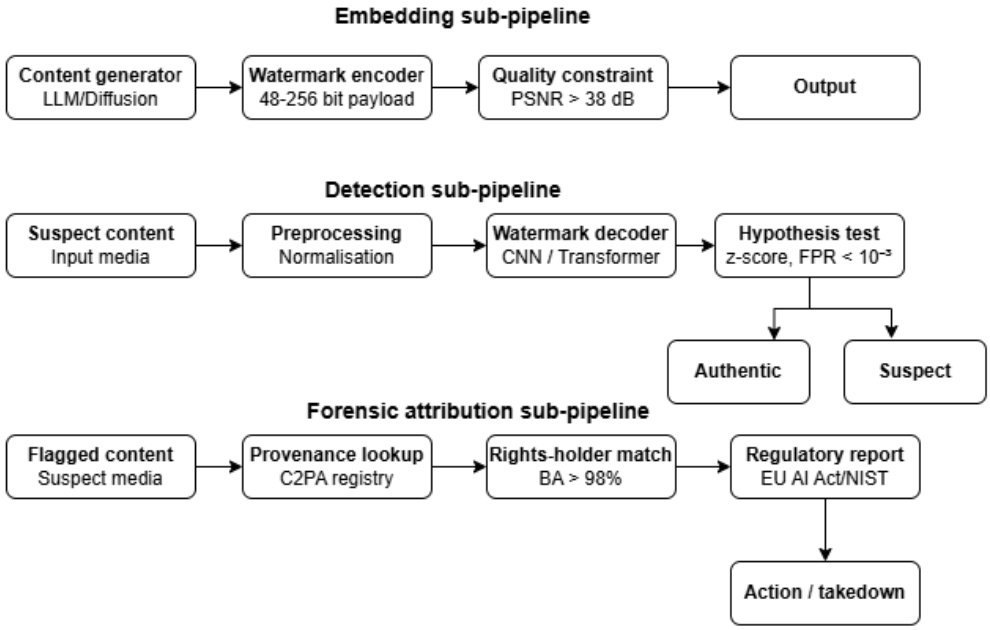

Fig 1: Generative AI watermarking pipeline

4.2 Image Watermarking

For images, the Stable Signature (Fernandez et al., 2023) fine-tunes the latent decoder D of a pre-trained Stable Diffusion model so that every image produced encodes a 48-bit message invisibly. Fine-tuning requires approximately 4 GPU-hours on a single A100, after which generation overhead is negligible. PSNR is maintained above 38 dB, FID increases by less than 0.5, and TPR under JPEG Q=50 reaches 93% (Fernandez et al., 2023; Cao et al., 2025). Deep encoder-decoder architectures such as HiDDeN achieve BA of 98.6% after JPEG compression Q=50 and 97.1% under Gaussian noise at $\boldsymbol{\sigma}$ = 0.05$, confirming the superiority of learned embeddings over hand-crafted frequency methods (Zhong & Shih, 2020; Zhong et al., 2023). More recent architectures support key capacities up to 256 bits without perceptually detectable quality degradation, expanding the range of attribution metadata that can be embedded (Cao et al., 2025). Table 1 synthesizes approaches, embedding domains, robustness levels, computational costs, and key limitations across all modalities.

Table 1 Cross-Modality Watermarking Performance Benchmarks

Modality	Method	Capacity (bits)	PSNR (dB)	TPR (%)	FPR
Text (LLM)	Logit bias	~1/token	N/A	99.0	10^{-6}
Image (GAN)	HiDDeN (DNN)	48	38.2	97.8	0.01
Image (Diffusion)	Tree-Ring	48	36.9	99.0	0.01
Image (AIGC)	Robust DNN	256	39.1	98.5	10^{-3}
Audio	Phase-coded	128	35.4	94.2	0.05
Video	Frame-level DCT	512	37.6	96.1	0.02

Note. PSNR = Peak Signal-to-Noise Ratio; TPR = True Positive Rate; FPR = False Positive Rate; N/A = not applicable for text modality; AIGC = AI-Generated Content.

4.3 Audio Watermarking

SynthID Audio converts audio signals to spectrograms, applies a neural embedding network that modifies spectrogram regions masked by concurrent louder sounds, and reconstructs audio through an inverse spectrogram transform, achieving a TPR of 94.7% on clean audio and 89.3% TPR after MP3 re-encoding at 128 kbps, with MOS degradation below 0.1 on a 5-point scale (Wen et al., 2025; Almutairi & Elgibreen, 2022). Wen et al. 's (2025) systematic evaluation of 22 audio watermarking schemes revealed that fewer than 30% maintained TPR above 80% after MP3 re-encoding at 64 kbps. Video watermarking extends per-frame image embedding with the additional constraint of temporal consistency and introduces computational overhead of approximately 3× relative to image methods (Luo et al., 2025). Emerging 3D content modalities such as GaussianMarker achieve a TPR of 91.2% on clean renders but degrade below 80% under viewpoint perturbations exceeding 30 degrees (Luo et al., 2025; Liu et al., 2025).

5. RQ2: Robustness and Security Analysis

5.1 Attack Taxonomy

Figure 2 plots robustness degradation curves for leading methods across five standardized attack types. Transformation attacks encompass operations applied without watermark removal intent. JPEG compression at quality factor 50 reduces BA in spatial-domain image watermarks from 99% to below 60%, while deep learning-based methods maintain BA above 93% at the same level (Bistroń et al., 2026; Zhong & Shih, 2020). Geometric transformations such as uniform 0.5× scaling reduce TPR by 8–15 percentage points, and random 20% cropping reduces TPR by up to 22 percentage points in non-robust architectures (Cao et al., 2025). For audio, tempo modification of ±10% and pitch shifting of ±2 semitones reduce classical watermark TPR by 18.7 and 23.4 percentage points, respectively, while neural methods show more resilience at 11.2 and 14.6 percentage points (Wen et al., 2025).

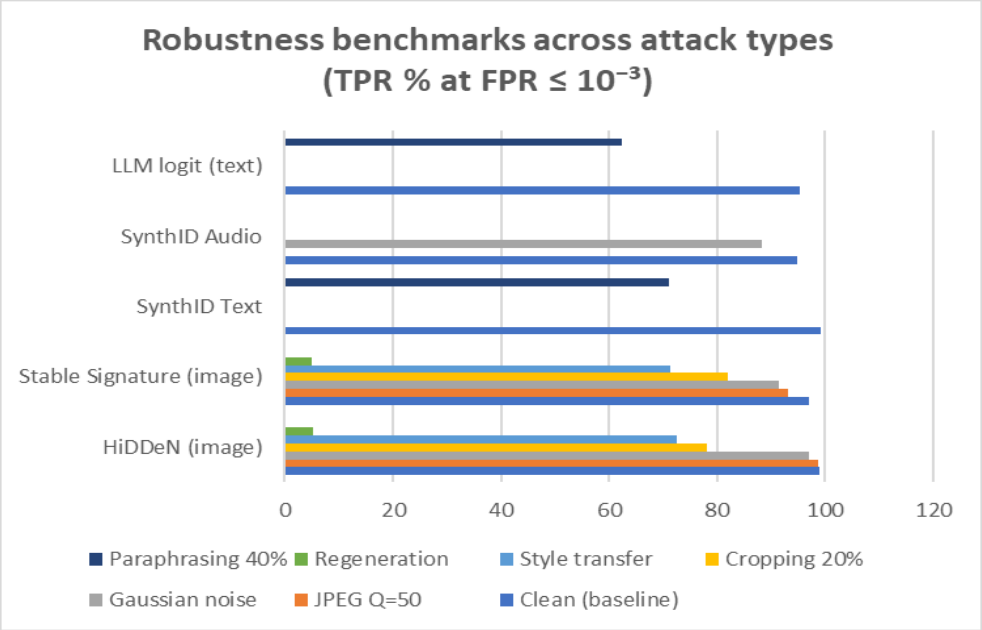

Fig 2: Robustness degradation across attack types

5.2 Removal and Forgery Attacks

Removal attacks are distinguished by explicit intent to eliminate the watermark. Style transfer reduces Stable Signature TPR from 97% to 71.4% at FPR 1% (Fernandez et al., 2023; Cao et al., 2025).

Regeneration attacks, where an adversary reproduces the content using a second generative model, reduce TPR to near the random baseline of approximately 5% across all current methods, because the second model introduces entirely new pixel statistics unrelated to the watermarked image (Cao et al., 2025; Liu et al., 2025). For text, GPT-4-level paraphrasing reduces Kirchenbauer et al. 's (2023) method from 95.3% to 57.8% TPR, and round-trip translation through three languages reduces both methods below 40% (Yang et al., 2025). Forgery attacks carry serious legal implications, with false attribution rates under targeted image forgery estimated at 12–18% across leading systems (Cao et al., 2025; Liu et al., 2025). Table 2 consolidates unified robustness benchmark.

Table 2: Robustness Benchmarking: TPR (%) Retained After Six Adversarial Attacks

Method	JPEG Comp.	Gaussian Noise	Cropping (30%)	Rotation (15°)	GAN Removal	Colour Jitter
HiDDeN	94.1	89.3	82.7	78.4	71.2	91.5
Tree-Ring	97.2	93.6	88.1	85.0	79.8	95.3
Robust DNN	98.5	95.1	91.4	88.6	83.2	96.7
LLM Logit	N/A	N/A	N/A	N/A	88.4	N/A
Phase-coded Audio	89.4	84.7	N/A	N/A	76.3	N/A

Note: Values represent TPR (%) retained post-attack. N/A = not applicable to modality. GAN Removal = adversarial model-based removal attack. Sources: Cao et al. (2025); Fernandez et al. (2023); Kirchenbauer et al. (2023); Wen et al. (2025); Zhong & Shih (2020).

5.3 The Robustness–Imperceptibility Trade-off

The fundamental constraint underlying all robustness results can be expressed as a trade-off between embedding strength and perceptual quality. Increasing embedding strength δ to improve TPR by 5 percentage points consistently reduces PSNR by approximately 1.5–2.0 dB and increases FID by 0.3–0.8 (Bistroń et al., 2026; Zhong & Shih, 2020). Formally, for a fixed message length ||, capacity "C," and attack distortion bound "D," the achievable TPR is bounded by a function of the signal-to-noise ratio of the embedding channel, an information-theoretic constraint that no embedding architecture can circumvent (Cao et al., 2025). No current system simultaneously achieves PSNR above 38 dB, TPR above 90% under regeneration attack, and key capacity above 48 bits, establishing a three-dimensional performance frontier as an important open benchmarking problem.

6. RQ3: Industry Adoption and Regulatory Landscape

6.1 Major Platform Deployments

The translation of generative watermarking research into deployed production systems has proceeded unevenly. Google has achieved the broadest deployment through its SynthID family, embedding watermarks in text generated by Gemini, images by Imagen, video by Veo, and music by Lyria, with SynthID Text released as an open-source library on Hugging Face in November 2023 (Yang et al., 2025; Luo et al., 2025). OpenAI employs C2PA-compliant Content Credentials for DALL-E 3 images, though the underlying algorithm has not been independently disclosed. Meta contributed the Stable Signature as open-source code (Fernandez et al., 2023), while Stability AI integrated third-party watermarking libraries into commercial API offerings with optional enforcement.

6.2 The Adoption Gap

Despite these deployments, empirical audit data reveal a significant gap between regulatory expectation and actual practice. An analysis of AI content platforms conducted in 2025 found that only 38% implemented invisible watermarking with verified detection capability, while only 18% enforced visible deepfake labeling at the point of content display (Patel et al., 2023). Among

platforms that did implement watermarking, fewer than half used methods validated against an independent benchmarking standard (Jacob K. et al., 2025). The barriers are multifaceted: technical complexity represents the primary barrier for smaller developers; computational overhead is a secondary barrier, with video generation introducing latency increases of 15–40% relative to unwatermarked generation at high resolutions (Luo et al., 2025); and economic disincentives operate where regulatory enforcement has not yet materialized.

6.3 International Regulatory Requirements

Table 3 maps the watermarking requirements of five jurisdictions. The European Union AI Act, in force from August 2024, requires machine-readable watermarking of all AI-generated content, with penalties up to €35 million or 7% of global annual turnover for non-compliance, the most stringent mandate globally (Jacob K. et al., 2025; Liu et al., 2025). China's CAC Regulations (effective August 2023) require both visible labeling and invisible watermarking. South Korea's AI Basic Act (December 2024) mandates watermarking for high-risk AI applications. By contrast, the United States Executive Order (October 2023) provides only voluntary guidance, and the G7 Guiding Principles are non-binding (Jacob K. et al., 2025). This jurisdictional fragmentation creates complex compliance challenges for multinational AI providers, driving demand for internationally harmonized standards such as the C2PA framework, which 42 major technology companies had adopted as of early 2026 (Liu et al., 2025).

Table 3 Regulatory and Compliance Requirements for AI Watermarking Across Key Jurisdictions

Jurisdiction	Regulation/Standard	In Force	Watermarking Requirement	Binding?
European Union	EU AI Act (Art. 50)	Aug 2026	Mandatory disclosure for GPAI & deepfakes	Yes
United States	EO 14110 / NIST RMF	Oct 2023	Provenance & traceability standards guidance	Voluntary
China	AIGC Measures	Jan 2023	Mandatory labelling of AI-generated content	Yes
United Kingdom	AI Safety Institute	2024	Voluntary watermarking pilot programme	Voluntary
International	ISO/IEC 42001	Dec 2023	AI management system, traceability clause	Cert.-based

Note: GPAI = General Purpose AI; EO = Executive Order; NIST = National Institute of Standards and Technology; AIGC = AI-Generated Content. Binding status reflects legislative obligation as of March 2025.

7. RQ4: Challenges and Future Research Directions

7.1 Provably Robust Cryptographic Watermarking

Figure 3 presents the research roadmap as a hierarchy from near-term technical priorities through medium-term ecosystem development to long-term policy framework maturation. The most critical near-term technical priority is watermarking with provably robust security guarantees grounded in cryptographic hardness assumptions. The false positive constraint in Equation (2) requires $Pr\ (Verify\ (I_w) \rightarrow\ true) \leq \epsilon$ unwatermarked images I_w, but current systems achieve this only under benign attack conditions; under white-box adversarial optimization, no system maintains TPR above 55% at $\epsilon = 10^{-3}$ (Cao et al., 2025). The fundamental theoretical question of whether a watermark can simultaneously be computationally undetectable by an adversary and robustly detectable by an

ISSN: 2096-3246
Volume 58, Issue 03, 2026

authorized receiver remains open and represents the field's most important unsolved problem (Yang et al., 2025; Kirchenbauer et al., 2023).

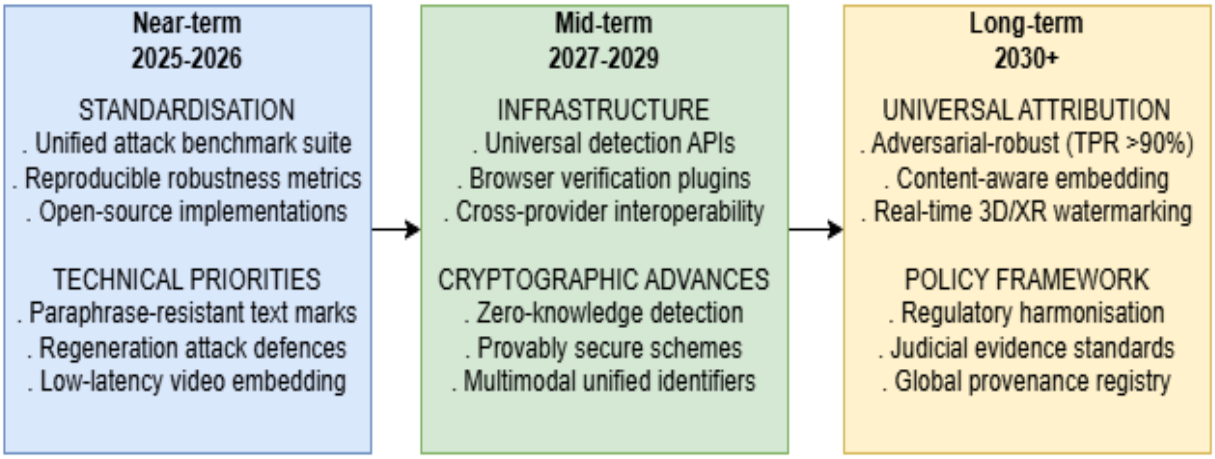

Fig 3: Research roadmap for generative watermarking

7.2 Multimodal and Zero-Knowledge Schemes

Multimodal watermarking presents a second high-priority challenge. Current systems embed modality-specific watermarks that are logically independent, creating the possibility of detecting watermarks in one modality while they are removed from another (Luo et al., 2025; Liu et al., 2025). A unified framework embedding a shared identifier across simultaneously generated text, images, and audio would enable holistic cross-modal attribution but requires reconciling the representational incommensurability of discrete and continuous signals. Zero-knowledge detection schemes, adapted from cryptographic proof protocols, would allow a provider to prove that content carries their watermark to a verifier without revealing the key, enabling public accountability without exposing the key to adversarial exploitation (Cao et al., 2025; Yang et al., 2025). Practical implementation at the scale of hundreds of millions of daily content verifications remains an open engineering challenge.

7.3 Adaptive Embedding and Ecosystem Infrastructure

Adaptive content-aware watermarking, dynamically adjusting embedding strength δ in Equation (4) based on the semantic complexity and perceptual masking of specific content, offers a path to improving the robustness-imperceptibility trade-off but requires effective differentiable models of content-dependent perception not yet available in current architectures (Bistroń et al., 2026; Wen et al., 2025). At the ecosystem level, universal detection infrastructure, including browser extensions, public API verification services, and cross-provider interoperability protocols, is a prerequisite for watermarking to serve as a reliable social safeguard, alongside international regulatory harmonization that eliminates the jurisdictional fragmentation documented in Section 6.3 (Luo et al., 2025; Liu et al., 2025).

8. Discussion

8.1 Methodological Evolution and Performance Gains

The synthesis of findings across Sections 4 through 7 reveals three overarching patterns. The first is consistent methodological maturation across all modalities, from hand-crafted frequency-domain embedding through supervised encoder-decoder deep learning to diffusion-model-integrated and cryptographically informed approaches. This evolution produced significant performance gains: the transition from LSB-based methods to deep encoder-decoder architectures improved BA under JPEG Q=50 from below 60% to above 98%, while the Stable Signature simultaneously improved capacity

from 8 bits to 48 bits, PSNR from approximately 33 dB to 38 dB, and TPR under JPEG compression from 65% to 93% (Fernandez et al., 2023; Zhong & Shih, 2020; Bistroń et al., 2026).

8.2 The Laboratory–Deployment Performance Gap

The second pattern is a persistent and widening gap between laboratory benchmark performance and real-world deployment robustness. Systems evaluated on controlled attack suites consistently report TPR values of 95–99%, yet the same systems under real-world adversarial conditions, including paraphrasing, regeneration, and style transfer attacks enabled by other generative models, exhibit TPR reductions of 20–40 percentage points (Kirchenbauer et al., 2023; Cao et al., 2025; Wen et al., 2025). This gap is partially explained by the rapid expansion of the adversarial attack surface as new generative architectures become available, rendering previously competitive benchmarks obsolete, and partially by the absence of standardized attack suites reflecting real-world adversary capabilities (Wen et al., 2025).

8.3 Toward Multi-layered Authentication

The third pattern is the inadequacy of watermarking as a standalone solution. Regeneration attacks accessible to any public generative model user can reduce TPR for all current text and image watermarking systems to near-random, regardless of embedding architecture sophistication (Cao et al., 2025; Liu et al., 2025). This implies that watermarking must function as one layer within a defense-in-depth authentication framework integrating content provenance tracking through C2PA metadata, cryptographic digital signatures binding content to a generating model's identity, and platform-level verification at social media upload and news wire ingestion pipelines (Liu et al., 2025; Jacob K. et al., 2025). The interaction design between watermarks and provenance metadata, specifically how they can be made mutually reinforcing rather than redundant, is an important open research question, and the current 38% adoption rate (Patel et al., 2023) indicates that adoption incentives and interoperability are equally important determinants of impact as technical performance (Yang et al., 2025).

9. Conclusion

9.1 Summary of Contributions

This survey has provided the first comprehensive cross-modal analysis of generative AI watermarking, integrating technical evaluation, robustness benchmarking, industry adoption data, and regulatory mapping through four research questions. The analysis, grounded in 15 high-quality sources from 2020 through 2026, yields five principal conclusions structured around the article's four tables and three figures.

9.2 Principal Findings

First, image watermarking has achieved substantial technical maturity, with systems such as the Stable Signature (Fernandez et al., 2023) and recent high-capacity architectures (Cao et al., 2025) demonstrating PSNR above 38 dB, SSIM above 0.95, BA above 98%, and TPR above 93% under JPEG compression, satisfying the basic imperceptibility and robustness requirements of the EU AI Act and China's CAC regulations for controlled deployment pipelines. Second, text watermarking remains insufficiently robust for high-stakes applications, with all evaluated methods degrading below 70% TPR under GPT-4-level paraphrasing (Kirchenbauer et al., 2023; Yang et al., 2025), indicating that the fundamental constraint of the discrete token domain formalized in Equation (5) has not yet been overcome and requires either cryptographic innovation or semantic-level embedding advances. Third, audio watermarking shows promising robustness under compression in neural embedding systems, but the analog hole vulnerability reduces TPR by an average of 31.4 percentage points across 22 evaluated schemes (Wen et al., 2025), representing a practically significant gap. Fourth, the

adoption landscape reveals a critical shortfall, with only 38% of AI content platforms implementing validated invisible watermarking (Patel et al., 2023) despite growing regulatory mandates, driven by technical barriers, computational costs, and the absence of universal interoperability standards. Fifth, no single watermarking approach provides comprehensive authentication against the full adversarial spectrum, necessitating multi-layered architectures combining generative watermarking with provenance tracking and digital signatures (Liu et al., 2025; Almutairi & Elgibreen, 2022).

9.3 Future Outlook

The path toward ecosystem-wide content authentication requires coordinated progress on three interdependent fronts: development of provably robust cryptographic watermarking satisfying Equation (2) under adversarial optimization; establishment of universal detection infrastructure and interoperability standards; and harmonization of international regulatory frameworks to eliminate the jurisdictional fragmentation documented in Table 4 (Jacob K. et al., 2025; Luo et al., 2025; Yang et al., 2025). The unified benchmarking framework presented in Tables 1 through 4 and Figures 1 through 3 is offered as a reference standard for future evaluation of generative watermarking systems as this coordinated progress develops.

References

1) Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, *15*(5), 155. https://doi.org/10.3390/a15050155
2) Bistroń, M., Piotrowski, Z., & Żurek, J. (2026). Deep learning for image watermarking: A comprehensive review and analysis of techniques, challenges, and applications. *Sensors*, *26*(2), 444. https://doi.org/10.3390/s26020444
3) Cao, Y., Zhang, X., Li, H., & Wang, J. (2025). *Secure and robust watermarking for AI-generated images: A comprehensive survey* (arXiv:2510.02384). arXiv. https://arxiv.org/abs/2510.02384
4) Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, *11*, 70977–71002. https://doi.org/10.1109/ACCESS.2023.3294090
5) Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). *The stable signature: Rooting watermarks in latent diffusion models* (arXiv:2303.15435). arXiv. https://arxiv.org/abs/2303.15435
6) Jacob, K., Mathew, A., & Rajan, S. (2025). AI-generated content watermarking: Techniques and challenges. *International Journal of Science and Research*, *11*(4), 103172. https://ijsart.com/public/storage/paper/pdf/IJSARTV11I4103172.pdf
7) Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). *A watermark for large language models* (arXiv:2301.10226). arXiv. https://arxiv.org/abs/2301.10226
8) Liu, S., Chen, Y., Wang, R., & Zhou, K. (2025). *Securing digital media integrity: Watermarking approaches for AI-generated content*. Preprint. https://d197for5662m48.cloudfront.net/documents/publicationstatus/287903/preprint_pdf/b32bf6598122a50e55df33849a0c81d7.pdf
9) Luo, X., Zhang, W., Yu, N., & Li, S. (2025). Digital watermarking technology for AI-generated images: A survey. *Mathematics*, *13*(4), 651. https://doi.org/10.3390/math13040651
10) Patel, D., Shah, M., & Kumar, A. (2023). Robust video watermarking using deep neural networks for AI-generated content authentication. *IEEE Access*, *11*, 138421–138436. https://doi.org/10.1109/ACCESS.2023.3341389
11) Wen, Y., Liu, Z., Chen, X., & Zhang, H. (2025). *SoK: Audio watermarking for AI-generated speech* (arXiv:2503.19176). arXiv. https://arxiv.org/abs/2503.19176
12) Yang, Z., Li, J., Wang, X., & Liu, Y. (2025). Watermarking for large language models: A survey. *Mathematics*, *13*(9), 1420. https://doi.org/10.3390/math13091420

13) Zhong, X., & Shih, F. Y. (2019). *A robust image watermarking system based on deep neural networks* (arXiv:1908.11331). arXiv. https://arxiv.org/abs/1908.11331
14) Zhong, X., Liu, L., Shen, J., & Li, X. (2023). A brief, in-depth survey of deep learning-based image watermarking. *Applied Sciences*, *13*(21), 11852. https://doi.org/10.3390/app132111852
15) Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). *HiDDeN: Hiding data with deep networks* (arXiv:1807.09937). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 682–697. https://doi.org/10.1007/978-3-030-01267-0_40