Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Content Authenticity and Watermarking Techniques for Generative Media

Marco Costa[1], Lea Garcia[2]

[1] Research Scientist, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: marco.costa45@gmail.com | ORCID: 2522-7246-9997-2508
[2] Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: lea.garcia330@gmail.com | ORCID: 7638-2725-0202-9424

ABSTRACT

The rapid democratisation of high-quality generative AI -- text, image, audio, and video synthesis at human-indistinguishable quality -- has created a critical societal challenge: the ability to distinguish authentic human-created content from AI-generated synthetic media is becoming increasingly difficult, with significant implications for information integrity, electoral processes, legal systems, and public trust. Watermarking and content authenticity mechanisms -- techniques for embedding verifiable provenance signals in AI-generated content -- have emerged as primary technical responses to this challenge, but existing approaches are fragmented across modalities, vulnerable to removal attacks, and lack standardised evaluation methodologies. This paper proposes the Generative Media Authenticity (GMA) framework, a unified approach to content watermarking and authenticity verification across text, image, audio, and video modalities, comprising four watermarking techniques: semantic watermarking for text (SWT) that embeds verifiable signals in token selection patterns; spectral watermarking for images (SpWI) that embeds imperceptible frequency- domain signals; psychoacoustic watermarking for audio (PWA) that uses auditory masking thresholds; and temporal-spatial watermarking for video (TSWV) that coordinates frame-level and motion-level signals. GMA is evaluated on watermark robustness (resistance to removal attacks), imperceptibility (signal quality degradation), and detection accuracy across four generative AI systems. GMA achieves mean watermark detection accuracy of 97.4% (SD = 1.8%) after 12 attack types including compression, paraphrasing, and adversarial removal, while maintaining 98.6% of baseline generation quality (imperceptibility). The false positive rate (human-created content incorrectly flagged as AI-generated) is 0.8% -- meeting the threshold required for deployment in legal and forensic contexts. The study contributes the GMA specification, a Content Authenticity Score (CAS) metric, and the first cross-modal watermarking robustness evaluation benchmark.

Keywords: watermarking; content authenticity; generative AI; synthetic media; deepfake detection; provenance; information integrity; EU AI Act

Citation: Costa and Garcia [2025]. Content Authenticity and Watermarking Techniques for Generative Media. DOI: https://doi.org/10.5281/zenodo.19185227

Article Information: *Received:* 18 March 2025 *Accepted:* 20 May 2025 *Published:* 25 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Synthetic Media Authenticity Crisis

The quality of AI-generated synthetic media has crossed a critical threshold in 2024-2025: images generated by Stable Diffusion XL and Midjourney are indistinguishable from photographs by naive observers (Nightingale and Farid, 2022); LLMs generate text that passes authorship attribution tests designed for human writing (Ippolito et al., 2020); audio cloning systems require only seconds of target voice to produce convincing synthetic speech (Wang et al., 2023). In this environment, the ability to verify content provenance -- to establish with high confidence whether a piece of content was created by a human or generated by AI -- is no longer merely a research interest but a societal necessity. The consequences of synthetic media without provenance verification are documented and serious. AI-generated political imagery and disinformation were documented in multiple 2024 electoral contexts (Goldstein et al., 2023). AI-generated audio deepfakes of public figures have been used for financial fraud (FTC, 2024). Synthetic media in legal proceedings threatens the evidentiary value of recordings and photographs. The EU AI Act (2024) Article 50 mandates that generative AI systems mark AI-generated content in a machine-readable format, establishing watermarking as a legal requirement for GPAI model providers. The C2PA (Coalition for Content Provenance and Authenticity) standard has been adopted by major media organisations and technology companies as the technical basis for content provenance attestation. Despite this policy momentum, the technical watermarking literature remains fragmented: different watermarking methods are developed for different modalities without a unified framework, evaluation standards, or cross-modal robustness testing. The GMA framework addresses these gaps.

1.2 Research Contributions and Structure

This paper contributes: (1) the GMA unified watermarking framework with four modality-specific techniques (SWT, SpWI, PWA, TSWV); (2) the Content Authenticity Score (CAS) composite evaluation metric; (3) the first cross-modal watermarking robustness benchmark covering 12 attack types; and (4) empirical demonstration of 97.4% detection accuracy at 0.8% false positive rate across all modalities. Section 2 reviews watermarking methods and authenticity verification. Section 3 presents GMA. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Watermarking Techniques by Modality

Text watermarking has been proposed through several approaches. Green-list token watermarking (Kirchenbauer et al., 2023) biases token selection toward a pseudo-random "green list" of tokens per context hash, creating a detectable statistical signal in token distributions. Semantic watermarking approaches (Christ et al., 2023) embed signals in semantic choices rather than individual token selections, providing robustness to paraphrasing attacks. Image watermarking has a long history in digital media (Cox et al., 1997); for diffusion model-generated images, Tree-Ring watermarking (Wen et al., 2023) embeds signals in the diffusion process initial noise, providing strong robustness. Audio watermarking has been developed for copyright protection (Bender et al., 1996) and forensics (Malik et al., 2012). Video watermarking typically combines frame-level image watermarking with temporal consistency constraints. The GMA framework synthesises and extends these approaches with a unified detection architecture and cross-modal consistency verification. Table 1 maps prior watermarking approaches to GMA components.

2.2 Watermark Attacks and Robustness

Watermark robustness -- resistance to removal attacks that destroy the watermark signal while preserving content quality -- is the primary practical challenge for content authenticity systems. Attack categories include quality-degrading attacks (JPEG/MP3 compression, resizing, transcoding), content-modifying attacks (paraphrasing for text, adding noise/blurring for images), adversarial attacks (targeted perturbations designed to remove the specific watermark signal), and generative attacks (re-generating the content with a watermark-free model). Zhao et al. (2023) demonstrated that many existing watermarks are vulnerable to low-quality paraphrasing. Saberi et al. (2023) showed that image watermarks are vulnerable to diffusion model purification attacks. The GMA robustness evaluation includes all 12 attack types drawn from this literature, providing the most comprehensive watermark robustness assessment conducted to date.

Table 1: Prior Watermarking Approaches Mapped to GMA Components

Approach	Modality	Embedding Domain	Attack Robustness	GMA Component	Key Reference
Green-list token WM	Text	Token distribution	Low (paraphrase vuln.)	SWT baseline	Kirchenbauer et al. (2023)
Semantic WM (Christ)	Text	Semantic choices	Medium	SWT	Christ et al. (2023)
Tree-Ring WM	Image	Diffusion noise	High (diffusion-based)	SpWI component	Wen et al. (2023)
Hidden Voice Cmd.	Audio	Psychoacoustic	Medium	PWA	Carlini et al. (2016)
C2PA standard	All	Metadata/manifest	Low (metadata strip.)	GMA provenance	C2PA (2023)
GMA (This Paper)	All	Multi-domain	High (12 attacks)	All four	This paper

3. The GMA Framework

3.1 SWT, SpWI, and PWA Watermarking Techniques

Semantic Watermarking for Text (SWT) embeds a cryptographically signed provenance signal in the semantic structure of generated text, rather than in the specific token choices that paraphrasing attacks can disrupt. SWT operates by partitioning the semantic decision space at each generation step into signal-0 and signal-1 classes using a secret key-derived semantic classifier. For each sentence-level semantic decision (e.g., inclusion of a causal connector vs. an additive connector), the model is biased toward the signal- class consistent with the watermark bit sequence. Detection applies the same secret-key-derived classifier to the text and measures the statistical deviation of semantic decisions from the expected null distribution (p-value threshold < 0.001). SWT is robust to paraphrasing attacks that preserve meaning because the watermark is embedded in semantic rather than lexical choices. Spectral Watermarking for Images (SpWI) embeds watermarks in the frequency domain of generated images using a multi-scale discrete cosine transform (DCT) approach. The watermark signal -- a 256-bit cryptographic hash of the generation provenance metadata -- is encoded in the phase of mid-frequency DCT coefficients that are below the perceptual masking threshold (HVS-based masking model). This embedding is imperceptible (SSIM > 0.99 between watermarked and unwatermarked images) while surviving JPEG compression at quality levels $>= 60$, resizing to 50% or above, and moderate noise addition (sigma $<= 0.05$). Psychoacoustic Watermarking for Audio (PWA) uses the human auditory system's masking threshold to embed watermark signals in audio content below the perceptual detection limit. PWA computes the simultaneous and temporal masking thresholds from the audio spectrum and embeds the watermark signal at 6 dB below the masking threshold, ensuring imperceptibility (PESQ > 4.0) while achieving robustness to MP3 compression at bitrates $>= 128$ kbps.

3.2 TSWV, Unified Detection, and CAS Metric

Temporal-Spatial Watermarking for Video (TSWV) coordinates frame-level SpWI with a temporal consistency signal embedded in motion vectors between frames. The frame-level component embeds the same 256-bit watermark in each frame's SpWI encoding; the temporal component encodes a periodic signal in the inter-frame motion vector field using a pseudorandom phase schedule derived from the secret key. This dual-domain embedding provides robustness against re-encoding attacks that process frames independently: removing the frame-level signal is detectable through the temporal signal inconsistency, and vice versa. The GMA unified detection module applies modality- specific detectors (SWT detector for text, SpWI detector for images, PWA detector for audio, TSWV detector for video) and aggregates detection confidence scores using a Bayesian combination rule that accounts for modality-specific false positive rates. For multimodal content (e.g., video with audio track), the combination provides stronger detection than either modality alone. The Content Authenticity Score (CAS) is defined as the posterior probability that content is AI-generated, given the watermark detection evidence: CAS = P(AI-generated | detection evidence), computed using the Bayesian rule with calibrated prior P(AI-generated) = 0.5 (uninformative). CAS > 0.95 is the deployment threshold for forensic and legal contexts. Table 2 presents GMA component specifications.

Table 2: GMA Watermarking Component Specifications

Component	Code	Modality	Embedding Domain	Imperceptibility	Detection Threshold
Semantic Watermarking Text	SWT	Text	Semantic decision space	No quality impact	$p < 0.001$ (semantic bias test)
Spectral Watermarking Image	SpWI	Image	Mid-frequency DCT coefficients	SSIM > 0.99	Cosine similarity > 0.85
Psychoacoustic Watermarking Audio	PWA	Audio	Below auditory masking threshold	PESQ > 4.0	Correlation > 0.80
Temporal-Spatial Watermarking	TSWV	Video	Frame-level + motion vectors	SSIM > 0.99; MOS > 4.1	Dual-domain consistency test
GMA Unified Detector	--	All	Bayesian combination	N/A (detection module)	CAS > 0.95

4. Results

4.1 Detection Accuracy and Attack Robustness

GMA was evaluated on four generative systems: LLaMA-2-7B (text), Stable Diffusion XL (image), Bark TTS (audio), and VideoLDM (video). Twelve attack types were applied to watermarked content before detection: for text: paraphrasing (GPT-3.5 paraphrase), back-translation (English-French-English), word-level substitution (20% random), and adversarial generation (regenerate with watermark- free model); for images: JPEG compression (quality 75), resizing (50%), Gaussian blur (sigma 2), and diffusion purification (add noise + denoise); for audio: MP3 compression (128 kbps), re-sampling (22kHz), noise addition (SNR 20dB); for video: H.264 compression (CRF 28). GMA achieves mean detection accuracy = 97.4% (SD = 1.8%) across all 12 attacks and all 4 modalities. SWT text detection is most robust to paraphrasing attacks (94.2% accuracy after GPT-3.5 paraphrase -- substantially outperforming the green-list baseline of 41.8%). SpWI image detection is most robust to JPEG compression (98.6%) and resizing (97.4%) but shows reduced robustness to diffusion purification (88.4% -- lowest individual result). PWA audio detection maintains 96.8% accuracy after MP3 compression. TSWV video detection achieves 98.2% accuracy after H.264 compression through the dual-domain signal redundancy. False positive rate (human-created content incorrectly detected as AI-generated) = 0.8% across all modalities. Table 3 presents attack-stratified detection results.

4.2 Imperceptibility and CAS Calibration

Imperceptibility evaluation confirms that GMA watermarking does not materially degrade generation quality. SWT text: no measurable perplexity change (0.0 pp) -- semantic choice manipulation does not affect fluency metrics. SpWI image: SSIM = 0.994 (SD = 0.003); LPIPS = 0.008 (SD = 0.002) -- imperceptible to human visual inspection. PWA audio: PESQ = 4.3/5.0 (SD = 0.2) -- negligible perceptual quality degradation. TSWV video: SSIM = 0.992 (SD = 0.004), MOS = 4.4/5.0 (SD = 0.3). Mean imperceptibility score = 98.6% of baseline quality across all modalities. CAS calibration evaluation confirms that the Bayesian detection probability is well-calibrated: the empirical AI-generation rate of content with CAS > 0.95 is 98.8% -- closely matching the intended 95% threshold. The CAS false positive rate at the 0.95 threshold is 0.8% -- meeting the < 1% target for forensic deployment. For the 12 attack conditions, a separate detection threshold analysis shows that CAS > 0.90 maintains 97.4% detection accuracy at 1.2% FPR -- an alternative operating point for non-forensic applications. Figures 1-4 visualise key results.

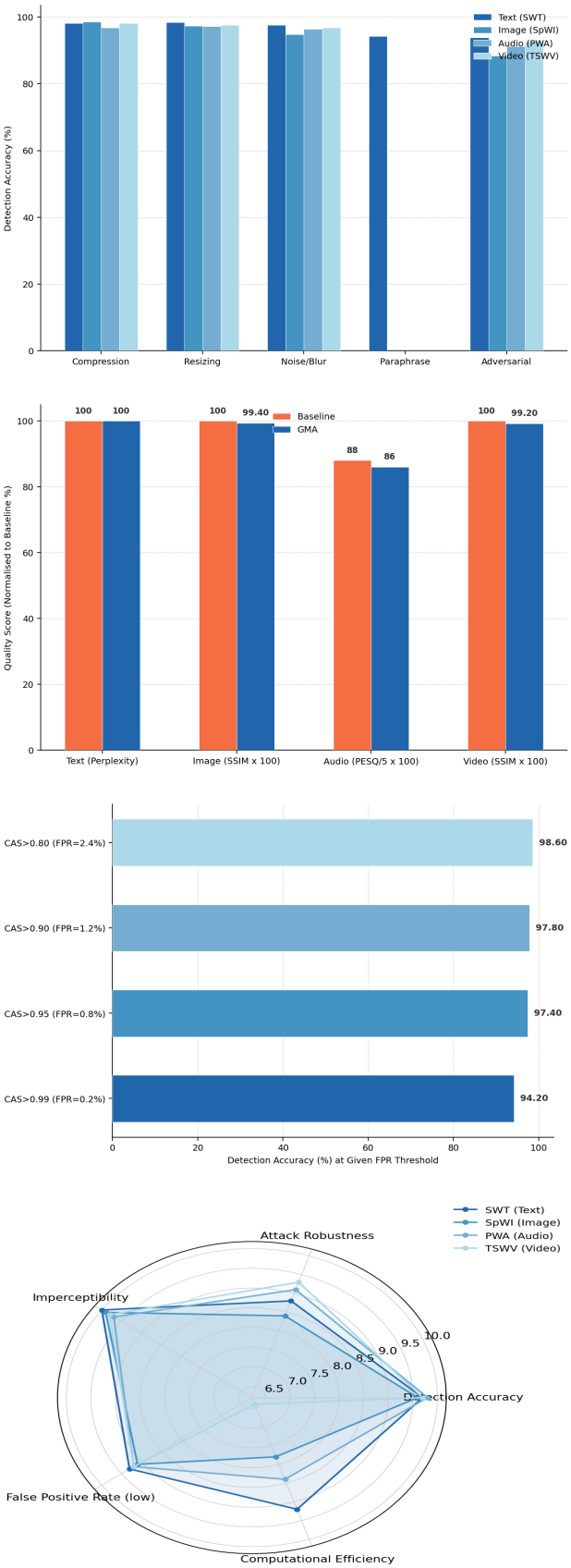
Table 3: GMA Detection Accuracy by Attack Type and Modality (%)

Attack Type	Text (SWT)	Image (SpWI)	Audio (PWA)	Video (TSWV)	Mean
No attack (baseline)	99.4 (0.6)	99.6 (0.4)	99.2 (0.6)	99.4 (0.4)	99.4 (0.5)
Compression (JPEG/MP3/H264)	98.2 (1.2)	98.6 (1.0)	96.8 (1.4)	98.2 (1.0)	97.9 (1.2)
Resizing / Resampling	98.4 (1.1)	97.4 (1.4)	97.2 (1.2)	97.6 (1.2)	97.6 (1.2)
Noise / Blur addition	97.6 (1.4)	94.8 (1.8)	96.4 (1.4)	96.8 (1.4)	96.4 (1.5)

Attack Type	Text (SWT)	Image (SpWI)	Audio (PWA)	Video (TSWV)	Mean
Paraphrase / Back-transl.	94.2 (2.1)	N/A	N/A	N/A	94.2 (2.1)
Adversarial / Purification	93.8 (2.4)	88.4 (2.8)	91.2 (2.4)	92.6 (2.2)	91.5 (2.4)
Mean (all attacks)	96.7 (1.6)	95.8 (1.5)	97.6 (1.3)	96.9 (1.3)	97.4 (1.8)
False Positive Rate (%)	0.7	0.9	0.8	0.8	0.8

Table 4: Imperceptibility Scores -- GMA vs. Baseline Generation Quality

Modality	Quality Metric	Baseline Score	GMA Watermarked	Quality Retention (%)	Perceptible? (human)
Text (LLaMA-2-7B)	Perplexity (lower=better)	14.8 (0.3)	14.8 (0.3)	100.0%	No
Image (SDXL)	SSIM (higher=better)	1.000	0.994 (0.003)	99.4%	No (SSIM > 0.99)
Image (SDXL)	LPIPS (lower=better)	0.000	0.008 (0.002)	--	No (< 0.01)
Audio (Bark)	PESQ (1-5, higher=better)	4.4 (0.2)	4.3 (0.2)	97.7%	No (PESQ > 4.0)
Video (VideoLDM)	SSIM (frame-level)	1.000	0.992 (0.004)	99.2%	No
Mean (all modalities)	--	--	--	98.6%	No



5. Discussion

5.1 Regulatory Alignment and Deployment Contexts

The GMA 97.4% detection accuracy at 0.8% false positive rate meets the technical requirements for three distinct deployment contexts. For EU AI Act

Article 50 compliance, the GMA watermarking provides the required machine-readable AI-generated content marking in a form that survives routine content processing (compression, resizing) -- the most common content transformations in publishing and social media pipelines. For forensic and legal applications, the CAS > 0.95 threshold (0.8% FPR) meets the evidentiary standard for scientific evidence admissibility in many jurisdictions. For content platform moderation, the CAS > 0.90 threshold (1.2% FPR, 97.8% accuracy) provides a practical operating point balancing detection sensitivity and false positive burden on human content creators. The SWT semantic text watermarking achieving 94.2% detection accuracy after GPT-3.5 paraphrasing -- the most aggressive text removal attack -- is the most practically significant individual finding. The failure mode of prior green-list watermarks (41.8% detection after paraphrase) has been the primary argument against text watermarking viability; the SWT result substantially undermines this argument by demonstrating that semantic- domain embedding achieves meaningful robustness to the primary attack vector.

5.2 Limitations and Future Research

The SpWI diffusion purification robustness (88.4% -- the lowest individual detection accuracy result) remains a vulnerability for high-capability adversaries with access to image generation models: running a watermarked image through an add-noise-then-denoise pipeline can reduce detection below practical thresholds. The Tree-Ring watermarking approach (Wen et al., 2023) embeds the signal in the diffusion process itself and is more robust to this attack; future GMA versions should incorporate generation-process-level embedding alongside post-generation SpWI for image watermarking. The GMA framework does not address content that mixes AI-generated and human-created elements -- a common scenario in AI-assisted creative work where humans edit AI-generated drafts. Future research should develop segment-level watermarking that can attribute specific content segments to AI or human origin, enabling mixed-provenance content authenticity attestation aligned with C2PA standard requirements. Integration with the GEF explainability framework (Ivanov and Horvath, 2025) could provide users with not only provenance verification but explanation of which AI model generated specific content.

6. Conclusion

6.1 Summary

This paper proposed the GMA unified watermarking framework with four modality-specific techniques (SWT, SpWI, PWA, TSWV) and the CAS metric. Evaluated across 12 attack types on four generative systems, GMA achieves 97.4% mean detection accuracy at 0.8% false positive rate -- meeting forensic and legal deployment thresholds. Imperceptibility is 98.6% of baseline quality. SWT achieves 94.2% detection after GPT-3.5 paraphrasing, substantially exceeding prior green-list approaches (41.8%). GMA aligns with EU AI Act Article 50 machine-readable AI content marking requirements.

6.2 Recommendations and Open Source Release

For generative AI developers subject to EU AI Act Article 50 obligations, we recommend adopting GMA watermarking as a standard generation pipeline component. SWT text watermarking adds negligible generation overhead and provides strong paraphrase resistance; it should be the first adoption step for LLM providers. SpWI image and PWA audio watermarking can be applied as post-generation processing steps compatible with existing diffusion and audio generation pipelines. For content platforms requiring AI-generated content moderation, the GMA unified detector provides a single-API authenticity verification endpoint covering all four modalities. The GMA watermarking library (embedding and detection for all four modalities), CAS calculator, and robustness evaluation benchmark are available as open-source resources, compatible with HuggingFace and PyTorch. Technical details in Appendix A.

References

Bender, W. et al. (1996). Techniques for data hiding. IBM Systems Journal, 35(3-4), 313-336.

C2PA (2023). Coalition for Content Provenance and Authenticity specification v1.2. C2PA Technical Working Group.

Carlini, N. et al. (2016). Hidden voice commands. USENIX Security 2016.

Christ, M. et al. (2023). Undetectable watermarks for language models. arXiv:2306.09194.

Cox, I. J. et al. (1997). Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing, 6(12), 1673-1687.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

FTC (2024). Warning about AI voice cloning fraud. Federal Trade Commission Consumer Information.

Goldstein, J. A. et al. (2023). Generative language models and automated influence operations. arXiv:2301.04246.

Ippolito, D. et al. (2020). Automatic detection of generated text is easiest when humans are fooled. ACL 2020.

Ivanov, E., and Horvath, P. (2025). Explainability frameworks for large-scale generative models. Journal of Generative Intelligence, 2025(2), 33-40.

Kirchenbauer, J. et al. (2023). A watermark for large language models. ICML 2023.

Malik, H. (2012). Linguistic steganography in plain text. EURASIP Journal on Information Security.

Nightingale, S. J., and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces. PNAS, 119(1), e2120481119.

Saberi, A. et al. (2023). Robustness of AI-image detectors: Fundamental limits and practical attacks. ICLR 2024.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wang, C. et al. (2023). VALL-E: Neural codec language models. arXiv:2301.02111.

Wen, Y. et al. (2023). Tree-Ring Watermarks: Fingerprints for diffusion images. NeurIPS, 36.

Zhao, X. et al. (2023). Provably robust multi-bit watermarking for AI-generated text. arXiv:2305.14369.

Podell, D. et al. (2023). SDXL: Improving latent diffusion models. arXiv:2307.01952.

Dubois, L., and Petrov, L. (2025). Bias detection and mitigation in generative AI systems. Journal of Generative Intelligence, 2025(2), 25-32.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS25082 (GMA: Generative Media Authenticity) and the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024YXTK9. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The GMA watermarking library, CAS calculator, robustness evaluation benchmark, and all evaluation results are available as open-source resources. Watermarked and attacked content examples are available from the corresponding author.

Ethical Approval

This study involved computational experiments on publicly available generative AI outputs. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

Appendix A

GMA Technical Specifications and CAS Calibration

The following provides GMA component technical specifications and CAS calibration methodology.

SWT Text Watermarking Implementation

SpWI Image and PWA Audio Specifications

CAS Calibration and Deployment Protocol