WATERMARKING FOR AI CONTENT DETECTION: A REVIEW ON TEXT, VISUAL, AND AUDIO MODALITIES

Lele Cao
AI Labs, King / Microsoft, Stockholm, Sweden
lele.cao@king.com | lelecao@microsoft.com
CSPaper*, Stockholm, Sweden
cspaper@googlegroups.com

ABSTRACT

The rapid advancement of generative artificial intelligence (GenAI) has revolutionized content creation across text, visual, and audio domains, simultaneously introducing significant risks such as misinformation, identity fraud, and content manipulation. This paper presents a practical survey of watermarking techniques designed to proactively detect GenAI content. We develop a structured taxonomy categorizing watermarking methods for text, visual, and audio modalities and critically evaluate existing approaches based on their effectiveness, robustness, and practicality. Additionally, we identify key challenges, including resistance to adversarial attacks, lack of standardization across different content types, and ethical considerations related to privacy and content ownership. Finally, we discuss potential future research directions aimed at enhancing watermarking strategies to ensure content authenticity and trustworthiness. This survey serves as a foundational resource for researchers and practitioners seeking to understand and advance watermarking techniques for AI-generated content detection.

1 INTRODUCTION

The exponential growth of generative artificial intelligence (GenAI) has reshaped the digital landscape, enabling the automated creation of realistic text, images, videos and audio. State-of-the-art generative models, such as DeepSeek (Guo et al., 2025), GPT series (Radford et al., 2018), Stable Diffusion (Esser et al., 2024), DALL·E (Ramesh et al., 2021), MusicGen (Copet et al., 2023), and NaturalSpeech (Tan et al., 2024) have demonstrated remarkable capabilities in mimicking human creativity. While these advancements show transformative opportunities in media, entertainment, and communication, they also pose substantial risks, including misinformation, identity fraud, and content manipulation (Chen et al., 2024; Tang et al., 2024). The ability of GenAI content to seamlessly blend into real-world scenario raises concerns on content authenticity and trustworthiness.

To counteract these risks, AI-generated content detection has emerged as a critical research domain. Among various detection methods, watermarking stands out as a proactive solution that embeds imperceptible yet traceable signatures within AI-generated outputs. Unlike reactive detection approaches that rely on post-generation analysis, watermarking enables content origin verification at the point of creation, facilitating robust and scalable AI-content identification (Lee et al., 2023). This makes watermarking particularly attractive for combating the misuse of AI-generated media.

Watermarking techniques vary significantly across different content modalities. For text-based AI content, probabilistic watermarking methods adjust token distributions to introduce hidden patterns, making it possible to detect GenAI text without impacting readability (Kirchenbauer et al., 2023). In visual content, watermarking strategies include spatial-domain embedding, frequency-domain modifications, and deep learning-based approaches to ensure robust detection against adversarial alterations (Mavali et al., 2024). For GenAI audio, watermarking methods modify spectral and temporal properties, allowing the authentication of synthetic speech and music (Roman et al., 2024).

*https://cspaper.org

Despite the promise of watermarking, several challenges remain. Watermark robustness is a key issue, as adversarial attacks and transformation techniques can be used to remove or obfuscate watermarks (Sadasivan et al., 2023). Moreover, the standardization of watermarking practices across modalities remains an open research problem, with different industries adopting varying approaches to embedding and detection. Addressing these challenges is essential for the widespread adoption of watermarking in AI-generated content detection.

This survey aims to provide a comprehensive analysis of watermarking techniques across multiple AI-generated content modalities. Specifically, we

- Present a structured taxonomy of watermarking methods used in AI-generated text, visual, and audio content detection.
- Review and compare existing watermarking techniques in terms of effectiveness, robustness, and practicality.
- Identify critical challenges in watermarking AI-generated content and discuss potential future research directions.

It is important to note that previous surveys have provided valuable insights into GenAI content detection for specific modalities, such as textual (Crothers et al., 2023; Tang et al., 2024; Valiaiev, 2024; Yang et al., 2024; Chaka, 2024; Wu et al., 2025), visual (Passos et al., 2024; Deng et al., 2024; Sandotra & Arora, 2024), and audio content (Almutairi & Elgibreen, 2022; Yi et al., 2023; Li et al., 2024b; Pham et al., 2024; Li et al., 2024b). However, none of these studies have offered a unified treatment of all three modalities. Furthermore, even two recent surveys that cover multimodal AI-generated content detection (Lin et al., 2024; Yu et al., 2024) have only briefly discussed watermarking as one of several detection approaches. In contrast, our survey is the first to systematically review watermarking techniques across text, visual, and audio domains. This integrated approach not only highlights the common underlying principles but also addresses the unique challenges and opportunities specific to each modality, thereby offering a holistic perspective that bridges existing gaps in the literature.

The remainder of this paper is structured as follows: Section 2 introduces the fundamental principles of watermarking and key evaluation metrics. Sections 3, 4, and 5 explore watermarking techniques for text, visual, and audio modalities, respectively. Section 6 discusses open challenges, ethical considerations and future directions, followed by a conclusion in Section 7.

2 Fundamentals of Watermarking

Watermarking is a proactive technique designed to embed imperceptible yet identifiable markers within AI-generated content, facilitating its detection and authentication. Unlike passive detection methods that analyze content post-generation, watermarking introduces traceable signatures at the point of content creation, enabling robust verification and origin attribution. This section formally defines watermarking schemes, explores different types of watermarking approaches, and discusses key evaluation metrics.

2.1 Definition and Key Principles of Watermarking

A watermarking scheme, visualized in Figure 1, for AI-generated content can be defined as a tuple:

$$\mathcal{W} = (\mathcal{E}, \mathcal{D}, \mathcal{V}), \text{ where}$$

- $\mathcal{E} : \mathcal{C} \times \mathcal{K} \times \mathcal{M} \to \mathcal{C}_w$ is the encoding function that embeds a watermark message $m \in \mathcal{M}$ into the input content $c \in \mathcal{C}$ using a secret key $k \in \mathcal{K}$, producing a watermarked output $c_w \in \mathcal{C}_w$. The set $\mathcal{C}$ contains original content, which may include texts, images, or audio signals. $\mathcal{C}_w \subset \mathcal{C}$ is the set of watermarked content. The sets $\mathcal{K}$ and $\mathcal{M}$ jointly guide the encoding and decoding processes. Depending on the modality, watermarks can take different forms: a text watermark might appear as a specific distribution of words (cf. chp1 in Cao, 2025), a visual watermark could be a textual message (cf. chp2 in Cao, 2025), and an audio watermark may be embedded as an additive waveform (cf. chp3 in Cao, 2025).
- $\mathcal{D} : \mathcal{C}_w \times \mathcal{K} \to \mathcal{M}$ is the decoding function that attempts to extract the watermark message m_d from the watermarked content c_w using the key k.

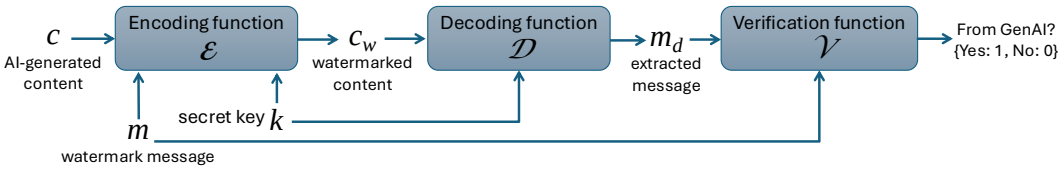

Figure 1: Schematic watermarking procedure: verify if input content c is produced from GenAI or a specific GenAI model.

- $\mathcal{V} : \mathcal{M} \times \mathcal{M} \to 0, 1$ is the verification function that determines whether the extracted watermark message m_d is a valid watermark.
- $\mathcal{C}$ is the set of original content, which may include texts, images, or audio signals.

A robust watermarking scheme should exhibit the following properties:

- **Imperceptibility:** The watermark should not degrade content quality or be detectable through normal human perception (Kirchenbauer et al., 2023). Visual and auditory contents are evaluated with PSNR (peak signal-to-noise ratio). For text, BLEU (bilingual evaluation understudy) and ROUGE (recall-oriented understudy for gisting evaluation) scores measure similarity between original and watermarked text.
- **Robustness:** The watermark should remain intact even after transformations such as compression, cropping, paraphrasing, or noise injection (Lee et al., 2023). This is evaluated through BER (bit error rate) defined as $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(m_i \neq \hat{m}_i)$, where m_i and $\hat{m}_i$ are the original and extracted watermark bits, respectively. Lower BER indicates greater robustness (Roman et al., 2024).
- **Security:** The watermark should be resistant to unauthorized removal or forgery by adversarial attacks. Watermarking security is often assessed based on its resistance to adversarial attacks such as GAN-based watermark removal and synonym substitution (Sadasivan et al., 2023). Techniques such as key-based authentication and encrypted watermarking enhance security.
- **Capacity:** The scheme should allow embedding sufficient information without significantly altering the content (Li et al., 2024a).

Interested readers are referred to papers like (Aberna & Agilandeeswari, 2024) for a more comprehensive side-by-side comparison of various watermarking performance metrics.

2.2 Types of Watermarking

Watermarking techniques vary based on their embedding methodology, visibility, and robustness to modifications. These techniques can be categorized into several types.

Visible vs. invisible watermarking. Visible watermarking embeds explicit identifiers such as logos or text overlays, commonly used in visual media but impractical for text and audio. Invisible watermarking, on the other hand, introduces imperceptible modifications that require algorithmic analysis for detection (Sharma et al., 2024).

Fragile vs. robust watermarking. Fragile watermarking is sensitive to modifications and is primarily used for content integrity verification, while robust watermarking remains detectable even after transformations such as compression, cropping, or reformatting.

Spatial vs. frequency domain watermarking. Spatial domain watermarking directly modifies content features, such as altering pixel values in images or inserting specific words in text. Frequency domain watermarking embeds information in transformed representations, such as using Discrete Cosine Transform (DCT) (Parah et al., 2016) or Discrete Wavelet Transform (DWT) (Hurrah et al., 2019) to modify frequency coefficients.

Deep learning-based watermarking. Recent advancements leverage deep neural networks to encode watermarks into the latent representations of AI-generated content, improving imperceptibility and robustness.

3 WATERMARKING FOR TEXT-BASED AI CONTENT DETECTION

The proliferation of large language models (LLMs) such as GPT-4 (Achiam et al., 2023), Llama-3 (Dubey et al., 2024) and DeepSeek (Guo et al., 2025) has revolutionized text generation, enabling the production of human-like content across various domains. While these advancements enhance accessibility and efficiency, they also introduce risks related to misinformation, academic dishonesty, and content manipulation. The challenge of distinguishing AI-generated text from human-authored content has spurred the development of watermarking techniques, which embed traceable markers within AI-generated text to facilitate its detection. Unlike post-hoc detection methods that analyze textual artifacts, watermarking offers a proactive solution by encoding verifiable signatures at the point of generation, allowing for robust attribution and authentication.

3.1 PRINCIPLES OF TEXT WATERMARKING

Watermarking for text-based AI content detection operates by embedding hidden yet detectable patterns into LLM-generated text. Formally, a text watermarking scheme consists of an encoding function that modifies generated tokens according to a predefined scheme, a decoding function that extracts and verifies the watermark, and a verification function that determines the presence of a valid watermark. These techniques must balance imperceptibility, robustness, security, and capacity to ensure effectiveness.

Imperceptibility ensures that watermarked text remains natural and indistinguishable from non-watermarked text to human readers. Robustness is necessary to maintain watermark detectability even after paraphrasing, synonym substitution, or minor modifications. Security prevents adversarial attacks from removing or falsifying watermarks. Capacity determines how much information can be embedded within a given text length without significantly altering its meaning or coherence.

3.2 PROBABILISTIC TOKEN-LEVEL WATERMARKING

A widely adopted approach for text watermarking involves manipulating token selection probabilities during text generation. Kirchenbauer et al. (2023) proposed a probabilistic watermarking scheme that biases the token sampling process toward a subset of "green" tokens based on a secret key. During generation, at each token step, the LLM preferentially selects tokens from this predefined green list, subtly altering the probability distribution. The resulting text appears natural but contains statistical deviations that are detectable by an authorized verifier.

Watermark detection relies on analyzing token distributions in a given text sample. If the proportion of green tokens exceeds a statistically significant threshold, the text is classified as watermarked. This method offers strong robustness against paraphrasing attacks while maintaining readability. However, challenges arise when texts are heavily edited, as modifications may disrupt the statistical signal, reducing detectability.

3.3 LEXICAL AND SYNTACTIC WATERMARKING

Beyond probabilistic methods, lexical and syntactic watermarking techniques modify word choices and sentence structures in a controlled manner. Early lexical watermarking techniques involved embedding predefined sequences of rare words or phrases within AI-generated text. However, these methods are susceptible to removal through simple editing (Roman et al., 2024). More sophisticated approaches leverage synonym substitution constrained by a secret embedding space, ensuring that word replacements are semantically valid while encoding a detectable pattern (Liu et al., 2024).

Syntactic watermarking operates by controlling sentence structures or punctuation patterns. A method proposed by Lee et al. (2023) involves encoding information into the grammatical structure of sentences without altering meaning. For instance, an AI-generated sentence might consistently favor passive constructions or specific syntactic tree patterns. These patterns are statistically rare in human-authored text, making them detectable while preserving naturalness.

3.4 CONTEXTUAL AND SEMANTIC WATERMARKING

Recent advancements explore contextual and semantic watermarking, embedding hidden signals into higher-level linguistic structures. One approach involves strategically inserting information-bearing paraphrases or stylistic markers into text, which remain robust against superficial modifications. For example, a watermarking scheme might enforce subtle stylistic consistencies, such as recurring phrase structures, discourse markers, or thematic redundancies (Sadasivan et al., 2023).

Another promising direction is embedding watermarks in semantic embeddings rather than raw text. Instead of altering words directly, this technique adjusts the latent representations from which text is generated, ensuring that semantic consistency is maintained while encoding verifiable signals (Roman et al., 2024). Such methods provide robustness against paraphrasing attacks but require specialized decoding mechanisms to extract watermark signatures from transformed text.

4 WATERMARKING FOR AI-GENERATED VISUAL CONTENT

The rise of AI-generated visual content has introduced transformative applications across media, design, and entertainment, yet it also raises substantial concerns regarding misinformation, copyright infringement, and digital authenticity. Advances in generative models, particularly generative adversarial networks (GANs) (Goodfellow et al., 2014), diffusion models (Ho et al., 2020), and variational autoencoders (VAEs) (Kingma & Welling, 2014), have enabled the creation of hyper-realistic images and videos that often evade human scrutiny. The necessity of robust detection mechanisms has thus led to extensive research on watermarking techniques designed to embed traceable signatures within AI-generated visual content (Saberi et al., 2024). Unlike reactive detection approaches, watermarking offers a proactive solution by embedding identifying marks directly into images and videos at the point of generation, facilitating verification and attribution.

Watermarking techniques for visual content can be categorized based on embedding methodologies, resilience against transformations, and their detectability. Broadly, these approaches include spatial-domain embedding, frequency-domain transformations, hybrid methods, and deep learning-based watermarking. Each of these techniques presents unique strengths and challenges in terms of imperceptibility, robustness, and resistance to adversarial attacks (Hosny et al., 2024).

4.1 SPATIAL-DOMAIN WATERMARKING

Spatial-domain watermarking directly modifies pixel values to encode information, making it straightforward to implement and computationally efficient. Traditional methods such as Least Significant Bit (LSB) embedding insert watermarks into the least significant bits of selected pixels, ensuring minimal visual distortion (Wolfgang & Delp, 1996). However, this approach is highly vulnerable to common image processing operations such as compression, filtering, and noise addition, which can degrade or remove the watermark (Sharma et al., 2024).

To enhance robustness, spatial-domain watermarking techniques have evolved to incorporate more sophisticated embedding strategies. Local Binary Pattern (LBP)-based watermarking improves resilience by encoding patterns in texture descriptors rather than direct pixel modifications (Wenyin & Shih, 2011). Similarly, cryptographic hashing techniques combined with spatial embedding increase security by ensuring that the watermark is difficult to tamper with without knowledge of the key (Raj & Shreelekshmi, 2018). Despite these advancements, spatial-domain watermarking remains sensitive to geometric distortions such as cropping and resizing, limiting its reliability in practical applications.

4.2 FREQUENCY-DOMAIN WATERMARKING

Frequency-domain watermarking addresses the vulnerabilities of spatial-domain methods by embedding watermarks within transformed representations of an image, making them more resistant to compression and noise (Singh & Singh, 2024; Sharma et al., 2024). The most common frequency-domain transformations used for watermarking include the Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Singular Value Decomposition (SVD).

DCT-based watermarking embeds information in mid-frequency coefficients, ensuring robustness against compression while maintaining visual quality (Parah et al., 2016). However, its reliance on fixed-sized blocks can introduce artifacts that affect imperceptibility. DWT-based approaches offer improved resistance to geometric transformations by embedding watermarks across multiple frequency bands, balancing robustness and visibility (Hurrah et al., 2019). SVD-based watermarking enhances security by modifying singular values, which remain stable under common transformations like compression and noise addition (Benrhouma et al., 2017). Nevertheless, these methods can suffer from synchronization issues when faced with significant image distortions or adversarial modifications (Luo et al., 2023).

4.3 Hybrid Watermarking Techniques

Recent advancements, such as DWT-DCT-SVD watermarking (Kang et al., 2018), combine multiple transform domain methods or integrate spatial domain techniques to boost robustness, flexibility, and real-time performance. The hybrid approaches try to embed multiple watermarks to address various scenarios, such as high visibility in one layer and resilience in another. The hybrid technique is especially useful for applications requiring protection against a combination of attacks, such as adversarial or spoofing attacks, and is resilient to common image transformations like compression and noise addition (Haghighi et al., 2021).

4.4 Deep Learning-Based Watermarking

Deep learning-based watermarking represents a significant evolution in AI-generated visual content detection, offering adaptive and highly resilient techniques. These methods leverage neural networks to encode watermarks into latent representations, ensuring robust embedding while preserving visual fidelity (Singh & Singh, 2024). Two prominent examples of deep learning-driven watermarking solutions are Google's SynthID[1] and Meta's Stable Signature (Fernandez et al., 2023).

SynthID embeds imperceptible watermarks into GenAI images, allowing detection even after transformations like cropping and filtering. Unlike traditional methods, SynthID operates in the latent space of image generation models, ensuring watermarks remain intact even after extensive modifications. Stable Signature follows a similar principle but integrates watermarking directly in the diffusion process, embedding unique binary signatures that persist under common transformations.

These deep learning approaches offer substantial advantages in terms of robustness and adaptability. However, they also present challenges, particularly in terms of computational complexity and susceptibility to adversarial attacks. Recent studies have shown that diffusion-based models can be fine-tuned to erase watermarks while maintaining visual consistency, highlighting an ongoing arms race between watermarking techniques and adversarial removal strategies (Saberi et al., 2023).

5 Watermarking for AI-Generated Audio Content

The fast surge of AI-generated audio, encompassing both synthetic speech and AI-composed music, has introduced significant challenges related to authenticity verification, copyright protection, and deepfake detection. Advanced generative models, such as text-to-speech (TTS) systems and AI-driven music composition platforms, have achieved remarkable realism, making it increasingly difficult to differentiate between synthetic and human-generated audio (Li et al., 2024b;c). While detection methods based on acoustic feature analysis and deep learning classifiers provide post-hoc identification of AI-generated content, watermarking presents a proactive solution that embeds identifiable, yet imperceptible, markers within AI-generated audio at the point of creation. This section provides a comprehensive review of watermarking techniques for AI-generated speech and music, discussing fundamental methodologies, implementation strategies, and ongoing challenges in robustness and adversarial resilience.

[1] SynthID: `https://deepmind.google/technologies/synthid`.

5.1 Principles of Audio Watermarking

Audio watermarking operates by embedding an imperceptible yet verifiable signal within an audio waveform, ensuring the traceability and integrity of AI-generated speech and music. A watermarking scheme for audio content follows the standard tuple formulation in Section 2.1: an encoding function that embeds the watermark, a decoding function that extracts it, and a verification function that authenticates its presence. Again, these schemes must meet four primary requirements: (1) imperceptibility requires embedding techniques that do not degrade audio quality; (2) robustness ensures the watermark survives transformations such as compression, noise addition and remixing; (3) security protects against unauthorized removal or spoofing; and (4) capacity defines how much information can be embedded without compromising detectability.

Watermarking approaches for AI-generated audio can be categorized into two primary domains: speech watermarking, which focuses on embedding watermarks in synthetic voices, and music watermarking, which targets AI-composed musical compositions. The following subsections examine each category in detail, reviewing major methodologies and their comparative advantages.

5.2 Watermarking for AI-Generated Speech

The advent of high-fidelity TTS and voice conversion (VC) models has enabled the creation of hyper-realistic synthetic voices, raising concerns about voice deepfakes and identity fraud. Watermarking AI-generated speech provides a reliable method for verifying audio authenticity, embedding inaudible markers that can be extracted and validated post-generation. Speech watermarking techniques are broadly categorized into spectral/temporal-domain and deep learning-based approaches.

Spectral-domain watermarking embeds imperceptible markers within the frequency components of an audio signal, leveraging transformations such as the Discrete Fourier Transform (DFT), DCT, and DWT (Dhar & Shimamura, 2015). These methods alter specific frequency coefficients to encode watermark signals while ensuring minimal perceptual distortion. The work of Hu & Hsu (2017) demonstrated a spectral-domain watermarking technique that modifies high-frequency bands to introduce secure identifiers without affecting speech intelligibility. Such methods exhibit strong robustness against common signal processing transformations, including MP3 compression and low-pass filtering, making them suitable for forensic analysis of AI-generated speech.

Temporal-domain watermarking embeds information within the time-domain features of an audio waveform, adjusting signal amplitude, phase, or timing characteristics. These techniques modify phoneme durations or introduce subtle phase shifts that remain undetectable to human listeners while ensuring verifiability through algorithmic analysis. One notable approach involves altering the pitch contour of AI-generated voices to encode a binary watermark signal while preserving natural speech prosody (Celik et al., 2005). Temporal watermarking methods tend to be more vulnerable to time-scale modifications such as speed adjustments and dynamic range compression, necessitating the use of hybrid techniques that integrate spectral and temporal features for enhanced robustness.

Deep learning-based watermarking represents an emerging paradigm that utilizes neural networks to learn optimal embedding strategies for speech watermarking. These methods train GANs or VAEs to encode watermark signals into the latent space of AI-generated speech, ensuring minimal perceptual distortion while maximizing robustness. AudioSeal (Roman et al., 2024) exemplifies this approach, leveraging a convolutional neural network (CNN) to generate imperceptible watermark signals embedded within the spectrogram representation of speech. Deep learning-based speech watermarking offer superior adaptability to adversarial modifications but require substantial computational resources for both embedding and detection, limiting their real-time scalability.

5.3 Watermarking for AI-Generated Music

AI-generated music presents unique challenges for watermarking due to the complex interplay of melody, harmony, rhythm, and instrumentation. Watermarking techniques for AI-composed music must preserve musical integrity while ensuring resilience against transformations such as tempo changes, remixes, and dynamic equalization (Epple et al., 2024). Music watermarking approaches are typically classified into symbolic-level and audio-level methods.

Symbolic-level watermarking operates on structured representations of music, such as MIDI files, embedding information within note sequences, rhythmic patterns, or harmonic progressions. These techniques introduce subtle modifications to chord voicings, note velocities, or timing variations to encode watermark signals while maintaining musical coherence. One prominent approach involves embedding unique harmonic transitions that remain detectable across different instrumentations and tempo adjustments (Bhandari & Colton, 2024). While highly robust within digital compositions, symbolic watermarking is ineffective for AI-generated music distributed as audio recordings.

Audio-level watermarking embeds watermarks directly into the audio waveform, leveraging spectral, phase, or amplitude modulation techniques. Frequency-domain watermarking is particularly effective, embedding signals in low-energy spectral bands that remain resilient to common post-processing transformations. Recent work by Liu et al. (2023) explored phase-based watermarking methods that introduce imperceptible phase shifts across harmonic frequencies, ensuring robust detection while maintaining high audio fidelity. Audio-level watermarking techniques are widely applicable to commercial AI-generated music platforms but require careful parameter tuning to balance imperceptibility and detectability.

Deep learning-driven watermarking employs neural networks to optimize watermark embedding and extraction within AI-generated music. SynthID, developed by DeepMind, exemplifies this approach, integrating watermarking into the latent space of music generation models to ensure persistent traceability (Dathathri et al., 2024). These methods provide strong resilience against adversarial modifications but demand high computational overhead, making real-time deployment challenging. The integration of self-supervised learning techniques holds promise for improving the efficiency and adaptability of deep learning-based music watermarking (Singh et al., 2024).

6 Challenges, Ethical Considerations, and Future Directions

Watermarking techniques for AI-generated content, while promising, encounter several persistent challenges that span across different modalities such as text, visual, and audio. Additionally, the implementation of watermarking raises significant ethical considerations that must be carefully navigated. This section delves into the primary challenges, ethical issues, and outlines potential future research directions essential for advancing watermarking methodologies.

6.1 Robustness and Adversarial Attacks

A paramount challenge in watermarking AI-generated content is ensuring robustness against adversarial attacks and transformation techniques. Watermarks embedded in text, visual, and audio content are susceptible to various forms of manipulation aimed at removing or obfuscating the embedded signatures. For instance, text watermarking techniques that rely on probabilistic token manipulation can be disrupted by paraphrasing or synonym substitution, as highlighted by (Sadasivan et al., 2023). Similarly, spatial-domain watermarking in visual content is vulnerable to common image processing operations such as compression, filtering, and geometric transformations (Sharma et al., 2024). In the audio domain, temporal-domain watermarking can be compromised by pitch shifting and dynamic range compression (Roman et al., 2024).

Deep learning-based watermarking approaches, while offering enhanced robustness, are not impervious to sophisticated adversarial strategies. Techniques such as GAN-based watermark removal and fine-tuning of diffusion models can effectively erase embedded watermarks without significantly degrading content quality (Sun et al., 2021). Addressing these vulnerabilities requires the development of more resilient embedding strategies that can withstand a wide array of adversarial modifications. Hybrid watermarking methods that integrate spatial, frequency, and deep learning techniques may offer improved robustness by leveraging the strengths of multiple approaches (Singh & Singh, 2023).

6.2 Standardization and Interoperability

The lack of standardized watermarking protocols across different content modalities presents a significant barrier to the widespread adoption and interoperability of watermarking technologies. Interoperability, defined as the ability of different systems, technologies, or organizations to seamlessly exchange and effectively use information or functionalities without restrictions, is crucial for en-

suring that watermarking solutions can function uniformly across diverse platforms and industries. Currently, watermarking schemes vary considerably between text, visual, and audio domains, each employing distinct embedding and detection methodologies (Zhao et al., 2024). This heterogeneity complicates the establishment of universal benchmarks and hinders the ability to verify watermarks across diverse platforms and industries.

Establishing industry-wide standards is crucial to ensure consistency in watermarking practices and facilitate cross-platform verification. Standardization efforts should aim to define common evaluation metrics, embedding protocols, and verification procedures that can be uniformly applied across different types of content. Furthermore, the development of interoperable watermarking frameworks that can seamlessly integrate with various generative models is essential to improve the scalability and applicability of watermarking solutions (Simmons & Winograd, 2024).

6.3 Ethical and Privacy Considerations

The deployment of watermarking technologies raises several ethical and privacy concerns that must be addressed with care. One primary concern is the potential for unauthorized watermarking of human-generated content, which could infringe on individual privacy and content ownership rights. In the context of AI-generated speech, embedding metadata without user consent can lead to surveillance and tracking issues, posing significant ethical dilemmas (Findlay, 2025).

Moreover, the voluntary adoption of watermarking by AI developers introduces challenges related to enforcement and compliance. Without regulatory mandates, the effectiveness of watermarking as a tool to verify the authenticity of content is undermined, as developers may opt to bypass watermarking mechanisms, especially in open-source models (Madiega, 2023). Balancing the enforcement of watermarking practices with respect to digital rights and fair use is critical to ensuring that watermarking does not stifle innovation or infringe on user autonomy.

Ethical watermarking implementations should incorporate privacy-preserving techniques that allow for watermark verification without compromising the privacy of content creators and consumers. Techniques such as encrypted watermarking and key-based authentication can enhance security while safeguarding against unauthorized access and misuse (Cheng et al., 2022).

6.4 Future Research Directions

Future research in watermarking for AI-generated content should focus on developing more adaptive, resilient, and scalable methods that address evolving generative models and adversarial tactics. One promising direction is the advancement of adaptive watermarking techniques that dynamically adjust to changes in generative model architectures, ensuring robustness against evolving attacks (Liu & Bu, 2024). Context-aware watermarking, which tailors embedding strategies based on domain-specific characteristics, can further enhance resilience (Guo et al., 2024).

Hybrid detection strategies that combine watermarking with retrieval-augmented verification and multimodal analysis hold significant potential for improving detection reliability. Retrieval-based detection leverages stored databases of AI-generated content to compare suspicious content against known watermarked outputs. By employing high-dimensional similarity searches and contrastive learning techniques, retrieval-based methods can effectively detect paraphrased or lightly edited AI-generated text, mitigating weaknesses in standalone watermarking schemes (Kang et al., 2024). Similarly, retrieval-augmented approaches in visual and audio domains can enhance detection precision by cross-referencing content with pre-indexed AI-generated samples.

Ensuring watermarking robustness against adversarial attacks remains a pressing challenge. Generative models can be fine-tuned to erase embedded watermarks while maintaining content quality, necessitating watermarking schemes that resist transformations such as synonym substitution, recompression, and paraphrasing (Sadasivan et al., 2023). Techniques such as adversarially robust watermarking, which employs perturbation-resistant encoding strategies, offer a promising solution. Additionally, integrating self-supervised learning techniques can optimize the efficiency and adaptability of deep learning-based watermarking methods, facilitating real-time deployment and scalability (Singh & Singh, 2024).

Standardization remains an essential research direction, as the lack of universal watermarking protocols hinders widespread adoption. Establishing industry-wide benchmarks and common evaluation metrics will improve interoperability between different watermarking techniques and content verification systems (Dathathri et al., 2024). Moreover, privacy-preserving watermarking techniques, such as encrypted and zero-watermarking schemes, are necessary to prevent unauthorized tracking while ensuring watermark detectability (Zhao et al., 2024).

Finally, interdisciplinary research is crucial to bridging technical advancements with ethical, regulatory, and legal considerations. Regulatory frameworks must balance content authenticity verification with user rights, ensuring that watermarking does not infringe on privacy or creative freedoms. The integration of watermarking with broader content provenance initiatives, such as digital content authentication systems and blockchain-based verification, represents an important direction for future research.

7 Conclusion

Watermarking stands out as a crucial proactive mechanism for the detection and authentication of AI-generated content across text, visual, and audio modalities, addressing the escalating concerns of misinformation, identity fraud, and content manipulation inherent in the proliferation of GenAI technologies. This survey categorizes and evaluates the diverse watermarking methodologies, highlighting the delicate balance between imperceptibility, robustness, security, and embedding capacity that each technique must achieve to be effective. Despite notable progress in developing sophisticated watermarking strategies, significant challenges remain, including enhancing resilience against increasingly sophisticated adversarial attacks, achieving standardization across various content types to ensure interoperability, and addressing ethical and privacy implications associated with watermark deployment. Furthermore, the dynamic nature of generative models call for the continuous evolution of watermarking techniques to maintain their efficacy. Future research should prioritize the development of adaptive and hybrid watermarking approaches that leverage advancements in deep learning and cryptographic methods to improve robustness and security. Additionally, establishing industry-wide standards and best practices is essential for the widespread adoption of watermarking solutions. Addressing ethical concerns such as user privacy and content ownership is crucial for fostering trust and compliance. By overcoming these challenges, watermarking can maintain digital content integrity and ensure a trustworthy ecosystem in the era of GenAI.

References

P Aberna and Loganathan Agilandeeswari. Digital image and video watermarking: methodologies, attacks, applications, and future directions. *Multimedia Tools and Applications*, 83(2):5531–5591, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, 15(5):155, 2022.

Oussama Benrhouma, Houcemeddine Hermassi, and Safya Belghith. Security analysis and improvement of an active watermarking system for image tampering detection using a self-recovery scheme. *Multimedia Tools and Applications*, 76:21133–21156, 2017.

Keshav Bhandari and Simon Colton. Motifs, phrases, and beyond: The modelling of structure in symbolic music generation. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pp. 33–51. Springer, 2024.

Lele Cao. *A Practical Guide to Detect GenAI Content*. Amazon, first edition, 2025. URL `https://www.amazon.com/dp/B0F2ZKH2R4`. Kindle Direct Publishing (KDP).

Mehmet Celik, Gaurav Sharma, and A Murat Tekalp. Pitch and duration modification for speech watermarking. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pp. ii–17. IEEE, 2005.

Chaka Chaka. Reviewing the performance of ai detection tools in differentiating between ai-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7(1):115–126, 2024.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

Hang Cheng, Qinjian Huang, Fei Chen, Meiqing Wang, and Wanxi Yan. Privacy-preserving image watermark embedding method based on edge computing. *IEEE Access*, 10:18570–18582, 2022.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 47704–47720. Curran Associates, Inc., 2023.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Qian Wang, and Chao Shen. A survey of defenses against ai-generated visual media: Detection, disruption, and authentication. *arXiv preprint arXiv:2407.10575*, 2024.

Pranab Kumar Dhar and Tetsuya Shimamura. *DWT-DCT-Based Audio Watermarking Using SVD*, pp. 17–35. Springer International Publishing, Cham, 2015. ISBN 978-3-319-14800-7. doi: 10.1007/978-3-319-14800-7_3.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pascal Epple, Igor Shilov, Bozhidar Stevanovski, and Yves-Alexandre de Montjoye. Watermarking training data of music generation models. *arXiv preprint arXiv:2412.08549*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

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*, pp. 22466–22477, 2023.

Georgina Findlay. You're gonna find this creepy: my AI-cloned voice was used by the far right. could i stop it? *The Guardian*, January 2025. URL `https://www.theguardian.com/commentisfree/2025/jan/07/ai-clone-voice-far-right-fake-audio`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuxuan Guo, Zhiliang Tian, Yiping Song, Tianlun Liu, Liang Ding, and Dongsheng Li. Context-aware watermark with semantic balanced green-red lists for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22633–22646, 2024.

Behrouz Bolourian Haghighi, Amir Hossein Taherinia, Ahad Harati, and Modjtaba Rouhani. Wsmn: An optimized multipurpose blind watermarking in shearlet domain using mlp and nsga-ii. *Applied Soft Computing*, 101:107029, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Khalid M Hosny, Amal Magdi, Osama ElKomy, and Hanaa M Hamza. Digital image watermarking using deep learning: A survey. *Computer Science Review*, 53:100662, 2024.

Hwai-Tsu Hu and Ling-Yuan Hsu. Incorporating spectral shaping filtering into dwt-based vector modulation to improve blind audio watermarking. *Wireless Personal Communications*, 94:221–240, 2017.

Nasir N Hurrah, Shabir A Parah, Nazir A Loan, Javaid A Sheikh, Mohammad Elhoseny, and Khan Muhammad. Dual watermarking framework for privacy protection and content authentication of multimedia. *Future generation computer Systems*, 94:654–673, 2019.

Xiao-bing Kang, Fan Zhao, Guang-feng Lin, and Ya-jun Chen. A novel hybrid of dct and svd in dwt domain for robust and invisible blind image watermarking with optimal embedding strength. *Multimedia Tools and Applications*, 77:13197–13224, 2018.

Zuheng Kang, Yayun He, Botao Zhao, Xiaoyang Qu, Junqing Peng, Jing Xiao, and Jianzong Wang. Retrieval-augmented audio deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 376–384, 2024.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*, 2023.

Fangqi Li, Haodong Zhao, Wei Du, and Shilin Wang. Revisiting the information capacity of neural network watermarks: Upper bound estimation and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21331–21339, 2024a.

Yupei Li, Manuel Milling, Lucia Specia, and Björn W Schuller. From audio deepfake detection to ai-generated music detection - a pathway and overview. *arXiv preprint arXiv:2412.00571*, 2024b.

Yupei Li, Qiyang Sun, Hanqian Li, Lucia Specia, and Björn W Schuller. Detecting machine-generated music with explainability - a challenge and early benchmarks. *arXiv preprint arXiv:2412.13421*, 2024c.

Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *ICLR*, 2024.

Chang Liu, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Dear: A deep-learning-based audio re-recording resilient watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13201–13209, 2023.

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*, 2024.

Xiyang Luo, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. DVMark: a deep multiscale framework for video watermarking. *IEEE Transactions on Image Processing*, 2023.

Tambiama Madiega. Generative AI and watermarking. *European Parliamentary Research Service*, (PE 757.583), December 2023. URL https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

Sina Mavali, Jonas Ricker, David Pape, Yash Sharma, Asja Fischer, and Lea Schönherr. Fake it until you break it: On the adversarial robustness of ai-generated image detectors. *arXiv preprint arXiv:2410.01574*, 2024.

Shabir A. Parah, Javaid A. Sheikh, Nazir A. Loan, and Ghulam M. Bhat. Robust and blind watermarking technique in dct domain using inter-block coefficient differencing. *Digital Signal Processing*, 53:11–24, 2016. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2016.02.005.

Leandro A Passos, Danilo Jodas, Kelton AP Costa, Luis A Souza Júnior, Douglas Rodrigues, Javier Del Ser, David Camacho, and João Paulo Papa. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, 41(8):e13570, 2024.

Lam Pham, Phat Lam, Dat Tran, Hieu Tang, Tin Nguyen, Alexander Schindler, Florian Skopik, Alexander Polonsky, and Canh Vu. A comprehensive survey with critical analysis for deepfake speech detection. *arXiv preprint arXiv:2409.15180*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018. URL https://openai.com/index/language-unsupervised.

NR Neena Raj and R Shreelekshmi. Blockwise fragile watermarking schemes for tamper localization in digital images. In *2018 International CET Conference on Control, Communication, and Computing (IC4)*, pp. 441–446. IEEE, 2018.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive Detection of Voice Cloning with Localized Watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 1–17, Vienna, Austria, July 2024. PMLR.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of AI-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of AI-image detectors: Fundamental limits and practical attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Neha Sandotra and Bhavna Arora. A comprehensive evaluation of feature-based ai techniques for deepfake detection. *Neural Computing and Applications*, 36(8):3859–3887, 2024.

Sunpreet Sharma, Ju Jia Zou, Gu Fang, Pancham Shukla, and Weidong Cai. A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, 83 (11):31829–31891, 2024.

John C Simmons and Joseph M Winograd. Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography. *arXiv preprint arXiv:2405.12336*, 2024.

Himanshu Kumar Singh and Amit Kumar Singh. Comprehensive review of watermarking techniques in deep-learning environments. *Journal of Electronic Imaging*, 32(3):031804–031804, 2023.

Himanshu Kumar Singh and Amit Kumar Singh. Digital image watermarking using deep learning. *Multimedia Tools and Applications*, 83(1):2979–2994, 2024.

Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji. Silentcipher: Deep audio watermarking. *arXiv preprint arXiv:2406.03822*, 2024.

Shichang Sun, Haoqi Wang, Mingfu Xue, Yushu Zhang, Jian Wang, and Weiqiang Liu. Detect and remove watermark in deep neural networks via generative adversarial networks. In *Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24*, pp. 341–357. Springer, 2021.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, 2024.

Dmytro Valiaiev. Detection of machine-generated text: Literature survey. *arXiv preprint arXiv:2402.01642*, 2024.

Zhang Wenyin and Frank Y Shih. Semi-fragile spatial watermarking based on local binary pattern operators. *Optics Communications*, 284(16-17):3904–3912, 2011.

Raymond B Wolfgang and Edward J Delp. A watermark for digital images. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pp. 219–222. IEEE, 1996.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pp. 1–66, 2025.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda R Petzold, William Yang Wang, and Wei Cheng. A survey on detection of llms-generated content. In *EMNLP (Findings)*, 2024.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

Xiaomin Yu, Yezhaohui Wang, Yanfang Chen, Zhen Tao, Dinghao Xi, Shichao Song, Simin Niu, and Zhiyu Li. Fake artificial intelligence generated contents (faigc): A survey of theories, detection methods, and opportunities. *arXiv preprint arXiv:2405.00711*, 2024.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.