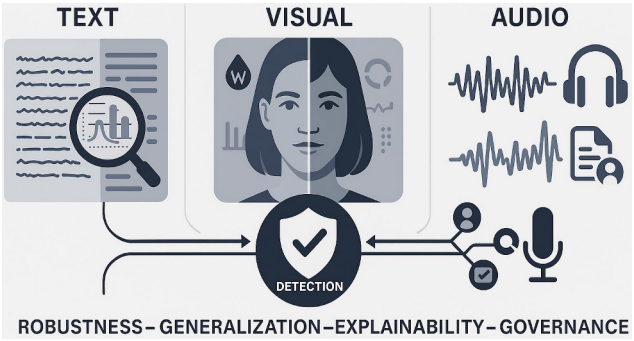
A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content

Lele Cao
AI Labs, King/Microsoft
R&D, CSPaper
Stockholm, Sweden
https://orcid.org/0000-0002-5680-9031



***Abstract*—Advances in AI-generated content have led to wide adoption of large language models, diffusion-based visual generators, and synthetic audio tools. However, these developments raise critical concerns about misinformation, copyright infringement, security threats, and the erosion of public trust. In this paper, we explore an extensive range of methods designed to detect and mitigate AI-generated textual, visual, and audio content. We begin by discussing motivations and potential impacts associated with AI-based content generation, including real-world risks and ethical dilemmas. We then outline detection techniques spanning observation-based strategies, linguistic and statistical analysis, model-based pipelines, watermarking and fingerprinting, as well as emergent ensemble approaches. We also present new perspectives on robustness, adaptation to rapidly improving generative architectures, and the critical role of human-in-the-loop verification. By surveying state-of-the-art research and highlighting case studies in academic, journalistic, legal, and industrial contexts, this paper aims to inform robust solutions and policymaking. We conclude by discussing open challenges, including adversarial transformations, domain generalization, and ethical concerns, thereby offering a holistic guide for researchers, practitioners, and regulators to preserve content authenticity in the face of increasingly sophisticated AI-generated media.**

***Index Terms*—Generative AI, AI-generated content detection, Deepfake detection, Large language models, Diffusion models, Generative adversarial networks, Watermarking, Multimedia forensics, Text analysis, Image forensics, Audio forensics**

I. Introduction

Generative AI technologies have drastically changed how digital content is created, consumed, and transmitted worldwide [10], [33]. Large language models (LLMs) now consistently produce text that emulates human writing style [62], [69], diffusion and GAN (Generative Adversarial Networks) based frameworks yield photorealistic images and videos [8], [22], and advanced text-to-speech (TTS) systems synthesize voices that rival real speakers [34], [68]. While these breakthroughs offer profound benefits to creative industries, data augmentation, and accessibility solutions, they also open the floodgates for significant societal challenges.

Unlike previous surveys that have largely focused on single modalities or isolated detection strategies, our work uniquely integrates detection approaches across textual, visual, and audio domains. We provide a comprehensive analysis of state-of-the-art techniques while also highlighting emerging trends such as watermarking, self-supervised learning, generative model fingerprinting, and human-in-the-loop verification. This holistic perspective, reinforced by practical case studies, differentiates our survey and underscores its relevance for both research and industrial applications.

Maliciously generated or manipulated content can propagate disinformation, orchestrate social engineering attacks, breach security protocols, undermine academic integrity, and violate intellectual property rights [16], [66]. From deceptively plausible news articles to convincing audio recordings impersonating public figures, adversaries increasingly create forgeries that elude basic detection.

Under increasing global concerns, the quest for robust detection of AI-generated textual, visual and audio content has intensified [69], [72]. Governments, academic institutions, technology platforms, and industries alike seek solutions to discern authentic content from synthetic. Addressing this need has led to the proliferation of detection strategies, ranging from visual artifact analysis to linguistic fingerprinting, watermark verification, and classification approaches.

The paper is organized as follows. Section II outlines the conceptual foundations of AI-generated content, including key generation approaches and threat models. Section III focuses on text-based detection strategies, from prompting-based classification to watermarking-based schemes. Section IV covers visual detection methods for AI-generated images and videos, emphasizing artifact analysis, watermarking, and advanced video forensics. Section V addresses synthetic audio detection, highlighting speech deepfake risks and music plagiarism.

This work was carried out with the support of CSPaper (https://cspaper.org), with contributions from Kai Xie, Lei You, Weiping Ding, Yong Du, Sven Salmonsson, Yumin Zhou, and Vilhelm von Ehrenheim.

2375-9259/25/$31.00 ©2025 IEEE
DOI 10.1109/ICDMW69685.2025.00170

Section VI provides real-world case studies in academia, journalism, and creative industries. Section VII surveys cross-modal insights, challenges, and future directions, including emerging self-supervised and generative model fingerprinting techniques. Finally, we conclude by underscoring the need for a multi-stakeholder and adaptive approach to preserve digital authenticity in the rapidly evolving generative AI landscape.

Contributions and Unifying Taxonomy

What is new. Beyond synthesizing prior work, we contribute:

- *A cross-modal taxonomy* unifying textual, visual, and audio detectors into five families: (i) *signal/statistical* (stylometry, pixel/spectral cues), (ii) *model-likelihood* (perplexity/curvature), (iii) *supervised classifiers* (fine-tuned detectors), (iv) *provenance mechanisms* (watermarking/fingerprinting and content credentials), and (v) *retrieval- and consistency-based* (nearest-neighbor retrieval, cross-modal agreement checks).
- *A threat model matrix* mapping attacker capabilities to defenses across modalities (Table I), clarifying realistic deployment choices and gaps.
- *Actionable deployment notes* for retrieval-based text detection (mitigating paraphrasing), temporal/physiological video forensics, and real-time multilingual audio detection.
- *Reproducible resources* linking datasets, leaderboards, and open-source implementations (Appendix A).

II. Background on AI-Generated Content

To contextualize the approaches for detection, we provide a brief overview of the primary mechanisms behind AI-generated content and typical threat models.

A. Generative Modeling Paradigms

Three major paradigms underlie most modern AI-based content generation:

- Autoregressive models: Exemplified by GPT-type architectures, these models predict tokens sequentially to generate coherent text, code snippets, lyrics, or even image patches (by ViT - Vision Transformer) [7], [62]. Because large parameter counts are typically needed for strong performance, they are often referred to as LLMs in NLP (natural language processing) domain.
- GANs: Operate on the principle of a generator and discriminator in competition, widely used for producing images, style transfer, and deepfake videos [8].
- Diffusion models: Gradually denoise random noise into highly realistic images or audio. Known for generating crisp visuals (e.g., Stable Diffusion [22] and DALL-E[1]) and increasingly refined speech and music [36].

These architectures have been refined to handle specialized tasks: style-based face generation, voice cloning, text-to-music composition, etc. They can also be composited, e.g., LLM-driven prompts controlling diffusion image generation, to produce multi-modal AI outputs.

[1]https://openai.com/index/dall-e-3

B. Threat Models and Malicious Use Cases

Disinformation campaigns involve automated text production at scale, which can flood social media with politically motivated messages, conspiracy theories, or phony press releases, effectively shaping public opinion [16], [33]. **Identity theft and fraud** can be facilitated by synthetic voice or video mimicking a corporate executive or family member, enabling financial scams, social engineering, or blackmail [68], [73]. **Academic dishonesty** arises when students or researchers generate entire essays, lab reports, or dissertations using LLMs [66], undermining educational integrity. **Plagiarism and copyright infringement** may occur as AI composition tools create music or artwork that partially copies existing works or inadvertently violates intellectual property, posing a serious challenge [35], [37]. **Manipulation in e-commerce or branding** is also a concern, as product images or videos might be faked to damage a competitor's brand or mislead customers about product quality [39].

The overall challenge is that as AI-based generation quality improves, naive or conventional authenticity checks fail. In response, a vibrant research ecosystem focuses on robust detection measures, as discussed next.

III. Detecting AI-Generated Text

A. Key Motivations and Constraints in Textual Detection

The written word remains a cornerstone of communication in academia, journalism, business, and everyday conversations. LLM-generated content can be harnessed productively, but the ease of generating vast, coherent text also creates new vulnerabilities, including misinformation spread, spam, impersonation, and academic plagiarism [10], [69].

Several major constraints hinder effective textual detection. **Language diversity** poses a significant challenge, as tools must function across different languages, dialects, and text domains (technical, legal, creative). **Model evolution** further complicates detection – systems trained on older text generators such as GPT-2 [50] often underperform when confronted with more advanced models like GPT-4 [1] or Bard [47]. Additionally, **paraphrasing attacks** exploit simple rephrasings to mask many stylistic signals of machine-generated text [30].

B. Approaches

1) LLM prompting and zero-shot methods: One popular approach leverages an external LLM to classify whether a piece of text is AI-generated or not [7]. With carefully crafted prompts, these zero-shot methods can achieve moderate accuracy quickly. However, sensitivity to prompt design and the adversarial gap between the LLM detector and the text generator are common issues [61]. Recent zero-shot probabilistic curvature methods (e.g., DetectGPT and Fast-DetectGPT) trade accuracy and compute in different ways [6], [44].

TABLE I

ATTACKER CAPABILITIES VS. DEFENSES ACROSS MODALITIES (TEXT / IMAGE-VIDEO / AUDIO). ✓ LIKELY EFFECTIVE; △ PARTIALLY EFFECTIVE OR CONTEXT-DEPENDENT; ✗ INEFFECTIVE. "PROV." INCLUDES WATERMARKING/FINGERPRINTING AND PROVENANCE STANDARDS (E.G., C2PA).

Attacker capability	**Text defenses**			**Image/Video defenses**			**Audio defenses**		
	Stat./LLM	Sup./Ensemble	Prov.	Pixel/Freq.	Temporal	Prov.	Pipeline/NN	End-to-end	Prov.
Zero-effort (raw model output)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Paraphrasing / style transfer	△	△	△*	△	△	△	△	△	△
Heavy compression / resampling	△	△	✓	△	△	✓	△	△	✓
Cropping / rotation / time-stretch	△	△	✓	△	△	△	△	△	△
Model-switching (unmarked models)	△	△	✗	△	△	✗	△	△	✗
Adversarial perturbations	△	△	△	△	△	△	△	△	△
Cross-lingual / domain shift	△	△	△	△	△	△	△	△	△
Splicing / synthetic-real mixing	△	△	△	△	△	△	△	△	△

* Several works show that paraphrasing can substantially degrade watermark detection, though more robust schemes continue to be proposed [23], [28], [31], [32], [52], [75]. Provenance standards like C2PA provide complementary, opt-in authenticity metadata [14].

2) Linguistic and statistical signatures: Traditional stylometric features (e.g., function words, syntax complexity, average phrase length) have long been used in authorship attribution [21], [44]. More modern detection focuses on computing perplexity or log-likelihood using reference language models, observing that LLM-generated text tends to show distinctive probability distributions. Additionally, specialized white-box methods can measure rank ordering of tokens if the generating model is partially known [60].

3) Supervised classification (training-based): Labeled corpora of AI versus human text enable fine-tuning of large pre-trained transformers like RoBERTa or T5 to discriminate synthetic text [12], [41]. Researchers improve robustness with adversarial training sets that contain paraphrased or AI-generated passages shifted in style. Tools like GPTZero [63] and RADAR [24] exemplify advanced supervised detectors. However, assembling high-quality, representative training data remains a challenge, especially as new generator architectures emerge frequently.

4) Watermarking for AI text: Cooperative watermarking modifies text generation at token selection time, embedding an imperceptible pattern in the distribution of words or punctuation [28], [56]. A verifier can detect such patterns after the fact. While promising for major industrial LLMs that adopt the standard, watermarking fails if malicious or open-source models do not embed it, or if paraphrasing disrupts the signal [23], [32], [52], [75]. Complementary provenance standards (e.g., C2PA) aim to capture capture-chain information when available [14].

5) Ensemble and multi-feature systems: To mitigate single-method vulnerabilities, some frameworks combine perplexity-based signals, style analysis, embedding-based classification, and watermark checks [48], [74]. By fusing different perspectives, these ensembles often achieve higher accuracy. The trade-off is system complexity and the need for sufficiently large training resources.

6) Retrieval-augmented defenses against paraphrasing (deployment notes): Paraphrasing can significantly degrade both watermark-based and likelihood-based detectors, yet retrieval proves effective: by searching for near-duplicates of candidate passages in a large corpus of model outputs or web-scale text, paraphrase-derived copies can be surfaced [30]. In practice we recommend: (i) maintain an index of recent LLM outputs or prompt-completion pairs; (ii) use character n-gram shingles (e.g., 5∼13) and embedding-based ANN search; (iii) combine retrieval scores with a calibrated detector score (e.g., logistic fusion); (iv) perform chunked matching (200 to 400 tokens) with overlap to handle local paraphrases; (v) log decisions for human audit in high-stakes settings (academia/journalism). Fast zero-shot curvature detectors [6], [44] can act as efficient first-stage filters before expensive retrieval.

IV. DETECTING AI-GENERATED VISUAL CONTENT

A. Motivations and Real-World Impact

AI-generated images and videos, often created via GANs or diffusion models, enable powerful visual illusions [8], [39]. Notable concerns include **political misinformation**, where fabricated news images depict fictional events; **financial fraud**, involving misleading product visuals or manipulation of stock markets; **harassment and defamation**, as seen in deepfake pornography or face swaps designed to humiliate victims; and **intellectual property theft**, such as art or design forgery that undermines artists' livelihoods.

B. Detection Methodologies

1) Observation and manual inspection: Human experts can sometimes detect unnatural artifacts in lighting, shadows, perspective, or anatomical features [26]. Context-based checks (e.g., unrealistic historical detail) also help. However, manual inspection is subjective, time-consuming, and not scalable for large volumes of online images.

2) Model-based artifact analysis: Algorithms analyze pixel-level statistics or frequency-domain features. For instance, Fourier transform reveals periodic textures characteristic of upsampling procedures in GANs or consistent small-scale noise from diffusion [42], [55]. White-box strategies exploit knowledge of the generator's pipeline (e.g., measuring the likelihood under the reverse diffusion process).

3) Black-box deep learning classifiers: With large amount of labeled real/fake data, CNN (convolutional neural network) based classifiers (ResNet, EfficientNet) learn discriminative cues [57]. Ensemble approaches combine multiple model

outputs or domain-specific sub-networks (e.g., focusing on faces vs. backgrounds) to bolster accuracy.

4) Watermarking for visual media: When the generative pipeline is compliant, watermarks are embedded. These vary from invisible spatial pixel encodings to frequency manipulations [28], [40]. However, simple image transformations, such as crop or rotation, can weaken naive watermarks [23], [59] unless specifically designed for robustness. Content credentials such as C2PA can attach tamper-evident provenance metadata on capture and edit chains [14].

5) Temporal/physiological forensics for video: Beyond per-frame artifacts, temporal and bio-signal cues are highly informative. Methods exploit phoneme-viseme inconsistencies in lip motion [2], remote photoplethysmography (rPPG) from subtle skin color changes [13], eye-blink and micro-expression patterns [38], optical flow coherence and head-pose/motion geometry [5], [11], [46], [65]. These cues are often more resilient to spatial filtering but can be sensitive to heavy compression and low frame rates.

C. Datasets and Benchmarks

Well-known datasets include CelebA-HQ [27] for facial images and LAION [58] for broad image domains. Large-scale synthetic benchmarks like GenImage (1M+ pairs; diverse generators and degradations) offer cross-generator and degraded-image evaluation [76], [77]. Community-maintained leaderboards and consolidated testbeds (e.g., AIGCDetect Benchmark) enable reproducible comparisons and track generalization to newer models [18], [20]. For video, FaceForensics++ [54], DFDC [17], and Celeb-DF are commonly used [64]. However, frequent advancement of the generative model requires continuous expansion of the detection dataset.

D. Limitations and Outlook

Compression sensitivity is a major challenge, as downsampling, scaling, or re-encoding can obscure forensic traces [55], [71]. **Generalization to new architectures** remains difficult since tools often lag behind novel generator types, such as next-generation diffusion or hybrid models. **Ethical implications** also complicate detection efforts: large-scale scanning of user images for potential deepfakes raises privacy concerns, while overly aggressive detection can flag benign content as suspicious, ultimately hurting user trust.

V. Detecting AI-Generated Audio Content

A. Risks and Use Cases

Synthetic audio has rapidly evolved thanks to neural TTS, voice conversion (VC), and audio diffusion. High-fidelity speech generation from minimal samples can facilitate voice impersonation [68], [73] or mislead detection systems in telephony security. Meanwhile, AI-composed music raises legal problems on originality, licensing, and plagiarism [36].

Use cases of audio detection span several domains. **Voice deepfake forensics** is essential in contexts such as law enforcement, banking, or enterprise authentication systems, where verifying speaker identity is critical [25], [45]. **Music authenticity** matters for streaming services or record labels that aim to detect GenAI music to safeguard artists' rights [35]. **Real-time moderation** is crucial for conference platforms that filter suspicious speech to prevent social engineering attacks.

B. Detection Techniques

1) Pipeline classifiers: Features such as Mel frequency cepstral coefficients (MFCC), Linear frequency cepstral coefficients (LFCC) spectrogram-based descriptors are extracted and then fed into machine learning models (SVM, XGBoost, or CNN) [34], [73]. They typically rely on analyzing subtle cues in pitch, timbre, and vocal fold dynamics.

2) End-to-end neural approaches: Powerful audio deepfake detectors ingest raw waveforms or full spectrograms using deep architectures like SincNet, Wav2Vec2.0 [4], or CRNN [49], [53]. These systems can learn complex patterns indicative of synthetic audio, such as unnatural transitions or missing microprosody. However, performance can degrade with domain shifts (e.g., new TTS pipelines, different languages) or background noise.

3) Music detection specifics: Music detection often analyzes melodic structure, chord progressions, or repetitive patterns [36], [51]. Large neural networks trained on real vs. AI-generated music can spot overly mechanical or simplistic progressions [15], [37].

4) Audio watermarking: Similar to the textual and visual domain, watermarking can be inserted into the synthetic audio during generation, using imperceptible frequency modulations or phase shifts [19], [40]. The watermark reveals the audio's origin; however, many transformation (tempo shift, reverb, denoising) can weaken naive watermarking signals.

C. Operational considerations: latency and multilingual robustness

Streaming/real-time: For telephony fraud prevention and live moderation, detectors must operate with very low latency. Sub-second chunking (e.g., 0.5–1.0 s hops with 2–3 s context windows) enables timely alerts, while causal CNN or Conformer backbones with look-ahead $\leq$200 ms help maintain responsiveness. Incorporating on-device VAD (voice activity detection) further reduces computational overhead and lowers false-alarm rates. **Calibration**: Detector performance should be tuned to in-domain EERs (equal error rates). Dynamic thresholding strategies that adapt to input SNR levels can significantly improve stability in noisy or variable conditions. **Multilingual/broad-accent**: Results from the ASVspoof evaluations, which include the *Logical Access (LA)* track for algorithmically generated speech, the *Physical Access (PA)* track for replayed audio, and the *Deepfake (DF)* track for highly realistic neural synthesis, consistently show sharp performance drops under cross-corpus and cross-condition settings. These findings underscore the importance of training with diverse languages and accents, followed by periodic domain adaptation [3], [67], [70]. Robustness can be further enhanced with self-supervised pretraining methods (e.g., HuBERT, Wav2Vec2) and test-time augmentations such as noise injection or codec simulation.

VI. Case Studies in Practice

To illustrate how detection strategies apply in different domains, we highlight several real-world contexts where AI-generated content is already a pressing concern.

A. Academic Integrity and Higher Education

Universities face surging usage of LLM tools for assignments and research papers. In many cases, naive plagiarism checks fail to detect newly generated text. Some institutions adopt specialized systems (e.g., GPTZero[2], Turnitin's AI detection[3]) that combine perplexity measures, stylometric analysis, and partial reference matching [66], [69]. However, concerns about privacy (scanning entire student submissions) and false positives remain. Many universities are establishing policies requiring students to label or disclaim AI assistance.

B. Newsrooms and Journalistic Fact-Checking

Misinformation campaigns are increasingly complex due to auto-generated text, manipulated images (e.g., fabricated protest scenes), and deepfake videos of political leaders. Journalists use hybrid detection pipelines: a first-level automated classifier flags suspicious content, which is then reviewed by human fact-checkers [16]. They also rely on watermark or metadata checks if major AI model providers tag their outputs. Major media platforms and social networks have begun integrating these systems into their content moderation workflows, combining automated screening with expert review to minimize false positives.

C. Law Enforcement and Legal Proceedings

Synthetic audio or video can compromise evidence authenticity. Forensic experts apply advanced image, audio forensics, and motion analysis tools to verify recordings [25], [64]. They may also cross-reference biometric cues, such as lip movement, voice biometrics, or EKG-like signals from speech waveforms. Courts increasingly grapple with how to interpret detection tool outputs, calling for transparent and explainable detection methods.

D. Creative Industries and Content Platforms

Content creators worry about plagiarism from AI tools that replicate their style or incorporate copyrighted material [37]. Music streaming platforms experiment with classifier-based scanning of newly uploaded tracks for suspicious patterns. Some are exploring watermark enforcement with partial success [19]. Visual artists are also advocating for improved detection of unauthorized use of their work, prompting platforms to implement hybrid approaches that combine automated and manual review processes.

For researchers interested in experimental validation and reproducibility, numerous open-source frameworks and benchmark leaderboards (e.g., DFDC, FaceForensics++, Audio Deepfake datasets) are available. We reference these resources to facilitate further exploration of state-of-the-art methods, and encourage readers to consult [9] for a more practical and comprehensive guide.

[2]https://gptzero.me

[3]https://www.turnitin.com

VII. Cross-Modal Insights, Challenges, and Future Directions

A. Unified Themes Across Modalities

Despite modality differences, several consistent themes arise.

1) *Arms race with generators*: As generative models advance, older detection strategies degrade, requiring frequent retraining or adaptation [43], [62].
2) *Vulnerability to perturbations*: Slight paraphrasing in text, minor image edits (e.g., crop and color shift), or audio pitch/time adjustments can bypass naive methods [30], [55].
3) *Watermarking as partial solution*: Watermarking is useful if widely adopted by model providers, but it's easily circumvented by uncooperative or malicious providers [32], [52].
4) *Ensemble or multimodal methods*: Combining multiple cues (statistical, watermark, contextual) and bridging text–image–audio modalities yields more robust detection [48], [74].
5) *Ethical pitfalls*: Systemic large-scale scanning can infringe on user privacy, and erroneous misclassifications can harm reputations.

B. Deeper analysis of watermark robustness

Empirical and theoretical studies report mixed results on watermark robustness under paraphrase, editing, and translation. While Unigram/semantic variants and adaptive schemes seek provable or empirical robustness [23], [32], [75], other analyses demonstrate learnability or reverse-engineering of green-list partitions and sharp drops in detection rates under targeted paraphrasing [29], [52]. We therefore position watermarking as one ingredient in a broader provenance strategy alongside content credentials (C2PA) and out-of-band logging for cooperative platforms [14].

C. Concrete pathways to address open challenges

Adversarial robustness: Strengthening resilience requires transformation-consistent training (such as paraphrasing or back-translation for text, crop/resize/codec chains for images, and time-stretching or noise perturbations for audio) combined with adversarial example mining and confidence calibration through temperature scaling. **Domain generalization**: To cope with unseen generators and distribution shifts, promising strategies include generator-aware data augmentation, continual learning with replay of past samples, and retrieval-based regularizers that leverage nearest-neighbor consistency. **Explainability**: Enhancing interpretability involves highlighting salient tokens, regions, or frames and linking them to human-understandable cues, for example, viseme–phoneme alignment [2] in video or rPPG signals in facial imagery. **Governance**: Beyond algorithmic techniques, standardized disclosures and verifiable audit trails, implemented via provenance

metadata (e.g., C2PA manifests), should be paired with opt-in watermarking whenever feasible to ensure accountability and transparency.

D. Dataset gaps and mitigation

Current benchmarks are heavily skewed toward *faces* and *English* text, leaving major gaps in other domains. Non-face imagery (e.g., documents, user interfaces, medical scans, remote-sensing data) and low-resource languages remain underrepresented, limiting generalization to diverse real-world scenarios. In audio, multi-lingual, multi-accent, and codec-diverse corpora are still scarce despite notable progress from ASVspoof [67], [70]. To close these gaps, we recommend three complementary strategies: (i) targeted data collection in underrepresented modalities and languages; (ii) synthetic hard-negative generation to expose detectors to adversarially challenging examples; and (iii) cross-temporal benchmarks that explicitly test robustness against newly released generators, such as GenImage and evolving AIGCDetect testbeds [18], [76], [77].

VIII. Conclusion and Broader Reflections

As AI generation systems become more sophisticated and widespread, detection is a critical fortress to uphold authenticity, trust, and accountability in the digital ecosystem. This paper has surveyed leading techniques for distinguishing AI-generated text, images, video, voice, and music, with attention to the motivations, current approaches, challenges, and ethical underpinnings across each domain. Our investigation underscores the following overarching lessons:

- No silver bullet: Each detection category, text, visual, or audio, relies on complementary signals (statistical, watermark-based, manual observation), yet none are foolproof against adaptive adversaries.
- Continual adaptation: Generative models evolve quickly. Detectors must be regularly updated, often requiring new training data from newly released generation architectures.
- Watermarking potential and pitfalls: Watermarks show promise if major platforms adopt them systematically, but they are easily circumvented by malicious or open-source models.
- Contextual and human-AI collaboration: Real-world detection extends beyond pure algorithmic classification. Human oversight, context checks, retrieval-based cross-referencing, and specialized domain knowledge remain pivotal, especially in high-stakes use cases.
- Ethical complexity: Overly invasive or inaccurate detection can harm user privacy and trust, while under-detection fuels misinformation and fraud. Balancing these risks requires responsible governance.

Looking forward, deeper integration across modalities, advanced retrieval-based or self-supervised approaches, improved adversarial robustness, and generative model fingerprinting will define the next generation of AI-content detectors. Collaboration among academia, industry, policymakers, and civil society is paramount to develop globally recognized standards and frameworks ensuring that generative AI can flourish as a positive force, while preserving integrity and truthfulness in digital media.

For further learning, we suggest to explore open-source tools, benchmark datasets, and books like [9] to accelerate progress in this fast-moving field.

Appendix A
Reproducibility Resources (Datasets, Detectors, Scripts)

To support reproducibility and practical experimentation, we summarize widely used datasets, detectors, and code resources across modalities in Table II. This table centralizes pointers that can serve as entry points for researchers and practitioners.

TABLE II
KEY REPRODUCIBILITY RESOURCES FOR AI-GENERATED CONTENT DETECTION.

Modality	Representative Resources
Text	DetectGPT, Fast-DetectGPT (zero-shot curvature) [6], [44]; retrieval-based defenses [30]; semantic/robust watermarks [23], [32], [75]
Image /Video	FaceForensics++ [54]; DFDC [17]; GenImage (1M+ pairs) [76], [77]; community leaderboards [18], [20]; temporal/physiological cues such as viseme mismatch [2], rPPG [13], blink patterns [38]
Audio	ASVspoof 2019/2021 datasets and challenge reports [3], [67], [70]; DeepMind SynthID audio watermark [19]
Provenance /Policy	C2PA technical specification and overview [14]

Acknowledgments

We thank the reviewers of the **MMAI Workshop on Multimodal AI**, held in conjunction with the **IEEE International Conference on Data Mining (ICDM) 2025**, for their constructive feedback. Their comments directly motivated the addition of a unified taxonomy, a threat matrix, temporal-forensics notes, retrieval deployment guidance, and a consolidated resource appendix.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] S. Agarwal and H. Farid, "Detecting deep-fake videos from phoneme–viseme mismatches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[3] ASVspoof Challenge Organizers, "ASVspoof 2021 challenge website," https://asvspoof.org/, 2021, accessed 2025-09-26.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[5] J. Bai, M. Lin, G. Cao, and Z. Lou, "AI-generated video detection via spatial-temporal anomaly learning," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2024.

[6] F. Bao, M. Post, C. Raffel, J. He, Z. He, and C. Callison-Burch, "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via self-consistency," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[7] A. Bhattacharjee and H. Liu, "Fighting fire with fire: Can chatgpt detect ai-generated text?" *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 2, pp. 14–21, 2024.

[8] A. Borji, "Qualitative failures of image generation models and their application in detecting deepfakes," *Image and Vision Computing*, vol. 136, p. 104771, 2023.

[9] L. Cao, *A Practical Guide to Detect GenAI Content*, 1st ed. Amazon, 2025, kindle Direct Publishing (KDP). [Online]. Available: https://www.amazon.com/dp/B0F2ZKH2R4

[10] C. 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*, vol. 7, no. 1, 2024.

[11] C. Chang, Z. Liu, X. Lyu, and X. Qi, "What matters in detecting ai-generated videos like sora?" in *arXiv preprint arXiv:2406.19568*, 2024.

[12] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, and B. Raj, "GPT-sentinel: Distinguishing human and chatgpt generated content," *arXiv preprint arXiv:2305.07969*, 2023.

[13] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, early version: arXiv:1901.02212 (2019).

[14] Coalition for Content Provenance and Authenticity (C2PA), "C2PA - verifying media content sources," https://c2pa.org/, 2025, accessed 2025-09-26.

[15] L. Comanducci, P. Bestagini, and S. Tubaro, "FakeMusicCaps: A dataset for detection and attribution of synthetic music generated via text-to-music models," in *arXiv preprint arXiv:2409.10684*, 2024.

[16] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, vol. 11, pp. 70 977–71 002, 2023.

[17] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[18] FDMAS Research Group, "Awesome AIGC detection benchmark (aigcdetect)," https://fdmas.github.io/AIGCDetect, 2023, accessed 2025-09-26.

[19] Google DeepMind, "Synthid for AI-generated audio," https://deepmind.google/science/synthid/ai-generated-audio/, 2024, accessed 2025-09-26.

[20] Gray Dove, "Awesome AIGC image detection," https://github.com/graydove/Awesome-AIGC-Image-Detection, 2023, accessed 2025-09-26.

[21] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi *et al.*, "Spotting llms with binoculars: zero-shot detection of machine-generated text," in *International Conference on Machine Learning (ICML)*, 2024.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[23] L. Hou, M. Wang, X. Liu, B. Li, and Q. Li, "Semstamp: A semantic watermark for large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[24] X. Hu, P.-Y. Chen, and T. Y. Ho, "RADAR: Robust ai-text detection via adversarial learning," in *Advances in Neural Information Processing Systems*, 2023, pp. 15 077–15 095.

[25] M. Hussain *et al.*, "Forensic audio authentication in law enforcement and court proceedings," *ArXiv e-prints*, 2022, arXiv:2210.11273.

[26] N. Kamali, K. Nakamura, A. Chatzimparmpas, J. Hullman, and M. Groh, "How to distinguish ai-generated images from authentic photographs," *arXiv preprint arXiv:2406.08651*, 2024.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[28] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023.

[29] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, and T. Goldstein, "On the reliability of watermarks for large language models," *arXiv preprint arXiv:2306.04634*, 2023.

[30] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," in *NeurIPS*, 2024.

[31] K. Krishna, V. Zayats, S. Agrawal, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text," *arXiv preprint arXiv:2303.13408*, 2023.

[32] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang, "Robust distortion-free watermarks for language models," *arXiv preprint arXiv:2307.15593*, 2023.

[33] L. Li, P. Wang, K. Ren, T. Sun, and X. Qiu, "Origin tracing and detecting of llms," *arXiv preprint arXiv:2304.14072*, 2023.

[34] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[35] Y. Li, H. Li, L. Specia, and B. W. Schuller, "M6: Multi-generator, multi-domain, multi-lingual and cultural, multi-genres, multi-instrument machine-generated music detection databases," *arXiv preprint arXiv:2412.06001*, 2024.

[36] Y. Li, M. Milling, L. Specia, and B. W. Schuller, "to ai-generated music detection: A pathway and overview," *arXiv preprint arXiv:2412.00571*, 2024.

[37] Y. Li, Q. Sun, H. Li, L. Specia, and B. W. Schuller, "Detecting machine-generated music with explainability: A challenge and early benchmarks," *arXiv preprint arXiv:2412.13421*, 2024.

[38] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai-created fake videos by detecting eye blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[39] L. Lin, N. Gupta, Y. Zhang, H. Ren, C. H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu, "Detecting multimedia generated by large AI models: A survey," *arXiv preprint arXiv:2402.00045*, 2024.

[40] A. Liu, L. Pan, X. Hu, S. Meng, and L. Wen, "A semantic invariant robust watermark for large language models," *ICLR*, 2024.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[42] S. Mavali, J. Ricker, D. Pape, Y. Sharma, A. Fischer, and L. Schönherr, "Fake it until you break it: on the adversarial robustness of ai-generated image detectors," *arXiv preprint arXiv:2410.01574*, 2024.

[43] N. S. Mireshghallah, J. Mattern, S. Gao, R. Shokri, and T. Berg-Kirkpatrick, "Smaller language models are better zero-shot machine-generated text detectors," *Proceedings of the EACL*, pp. 278–293, 2024.

[44] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023.

[45] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *INTERSPEECH 2022*, 2022.

[46] T. D. Nguyen, S. Fang, and M. C. Stamm, "Videofact: Detecting video forgeries using attention, scene context, and forensic traces," in *WACV*, 2024, pp. 8563–8573.

[47] E. P. Nyberg, A. E. Nicholson, K. B. Korb, M. Wybrow, I. Zukerman, S. Mascaro, S. Thakur, A. Oshni Alvandi, J. Riley, R. Pearson *et al.*, "Bard: A structured technique for group elicitation of bayesian networks to support analytic reasoning," *Risk Analysis*, vol. 42, no. 6, pp. 1155–1178, 2022.

[48] I. Ong and B. K. Quek, "Applying ensemble methods to model-agnostic machine-generated text detection," *arXiv preprint arXiv:2406.12570*, 2023.

[49] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, "Towards generalisable and calibrated audio deepfake detection with self-supervised representations," in *Interspeech*, 2024, pp. 4828–4832.

[50] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[51] K. S. Rao and P. P. Das, "Melody extraction from polyphonic music by deep learning approaches: a review," *arXiv preprint arXiv:2202.01078*, 2022.

[52] S. Rastogi and D. Pruthi, "Revisiting the robustness of watermarking to paraphrasing attacks," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 18 100–18 110.

[53] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[54] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[55] M. Saberi, V. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi, "Robustness of ai-image detectors: fundamental limits and practical attacks," *ICLR*, 2024.

[56] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?" *arXiv preprint arXiv:2303.11156*, 2023.

[57] R. A. F. Saskoro, N. Yudistira, and T. N. Fatyanosa, "Detection of ai-generated images from various generators using gated expert convolutional neural network," *IEEE Access*, 2024.

[58] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[59] S. Sharma, J. J. Zou, G. Fang, P. Shukla, and W. Cai, "A review of image watermarking for identity protection and verification," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 31 829–31 891, 2024.

[60] J. Su, T. Zhuo, D. Wang, and P. Nakov, "DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text," *Findings of the Association for Computational Linguistics: EMNLP*, pp. 12 395–12 412, 2023.

[61] K. Taguchi, Y. Gu, and K. Sakurai, "The impact of prompts on zero-shot detection of ai-generated text," *arXiv preprint arXiv:2403.20127*, 2024.

[62] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated text," *Communications of the ACM*, vol. 67, no. 4, pp. 50–59, 2024.

[63] Y. Tian, H. Chen, X. Wang *et al.*, "Multiscale positive-unlabeled detection of ai-generated texts," in *International Conference on Learning Representations (ICLR)*, 2024.

[64] A. K. Tiwari, A. Sharma, P. Rayakar, and M. K. Bhavriya, "AI-generated video forgery detection and authentication," in *IEEE I2CT*, 2024, pp. 1–8.

[65] D. S. Vahdati, T. D. Nguyen, H. Azizpour, and M. C. Stamm, "Beyond deepfake images: detecting ai-generated videos," in *CVPR*, 2024.

[66] D. Valiaiev, "Detection of machine-generated text: Literature survey," *arXiv preprint arXiv:2402.01642*, 2024.

[67] X. Wang, J. Yamagishi, A. Nautsch, N. Evans, T. Kinnunen, M. Todisco, H. Delgado, M. Sahidullah, V. Vestman, K. A. Lee *et al.*, "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge," *arXiv preprint arXiv:1911.01601*, 2019.

[68] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.-A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," in *Computer Speech & Language*, vol. 64. Elsevier, 2020, p. 101114.

[69] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, "A survey on llm-gernerated text detection: Necessity, methods, and future directions," *arXiv preprint arXiv:2310.14724*, 2024.

[70] J. Yamagishi, N. Evans, M. Todisco, H. Delgado, X. Wang, T. Kinnunen *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proceedings of the ASVspoof 2021 Workshop*, 2021.

[71] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie, "A sanity check for AI-generated image detection," *arXiv preprint arXiv:2406.19435*, 2024.

[72] X. Yang, L. Pan, X. Zhao, H. Chen, L. Petzold, W. Y. Wang, and W. Cheng, "A survey on detection of llms-generated content," *arXiv preprint arXiv:2310.15654*, 2023.

[73] J. Yi, C. Wang, J. Tao, X. Zhang, C. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[74] Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, and G. Chen, "Detecting ai-generated sentences in realistic human-ai collaborative hybrid texts: Challenges, strategies, and insights," in *arXiv preprint arXiv:2403.03506*, 2024.

[75] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang, "Provable robust watermarking for AI-generated text," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[76] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "GenImage: A million-scale benchmark for detecting AI-generated image," *arXiv preprint arXiv:2306.08571*, 2023.

[77] ——, "GenImage: A million-scale benchmark for detecting AI-generated image," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2023.