Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Generated Content Detection: A Cross-Modal Survey of Methods, Challenges, and Future Directions

Mohamed Mady

mohamed.mady@tum.de

Technical University of Munich

Yupei Li

Imperial College London

Björn W. Schuller

Technical University of Munich

Berrak Sisman

Johns Hopkins University

Johannes Reschke

OTH Regensburg

Research Article

Keywords: AI-generated content detection, Deepfakes, Synthetic media, Multimedia forensics, Cross-modal detection, Adversarial robustness

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10864156/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-Generated Content Detection: A Cross-Modal Survey of Methods, Challenges, and Future Directions

Mohamed Mady[1,2*], Yupei Li[3*], Björn W. Schuller[1,3], Berrak Sisman[4], Johannes Reschke[2]

[1*]Chair of Health Informatics, Technical University of Munich (TUM), Munich, 81675, Germany.
[2]Smart Embedded Systems Lab, OTH Regensburg, Regensburg, 93053, Germany.
[3]GLAM – Group on Language, Audio & Music, Imperial College London, London, United Kingdom.
[4]Center for Language and Speech Processing (CLSP), Johns Hopkins University, Baltimore, MD, USA.

*Corresponding author(s). E-mail(s): mohamed.mady@tum.de; yl7622@ic.ac.uk;
Contributing authors: schuller@tum.de; sisman@jhu.edu; johannes.reschke@oth-regensburg.de;

Abstract

The explosive growth of generative artificial intelligence (AI) has dramatically accelerated the production of increasingly realistic synthetic text, images, audio, and video, unlocking new creative and operational capabilities while introducing acute risks to information integrity, digital security, and societal trust. This survey provides a comprehensive, up-to-date synthesis of AI-generated content detection research across text, image, audio, and video. Rather than treating each modality in isolation, we organise the literature around six shared families of detection techniques: likelihood-based scoring, frequency-domain analysis, artefact and fingerprint forensics, transformer-based representation learning, consistency and semantic checks, and active watermarking and cryptographic provenance, showing, for each family, where the technique transfers cleanly across modalities and where modality-specific assumptions cause it to break.

We then instantiate this view in a unified four-axis cross-modal taxonomy (objective, paradigm, signal type, and generalisation regime), operationalised through a master classification of representative methods. We further quantify the laboratory-to-deployment gap, critically assess adversarial robustness, cross-model and cross-domain generalisation, dataset scarcity, and fairness, and distil open research gaps toward robust, generalisable, and interpretable detection.

Keywords: AI-generated content detection, Deepfakes, Synthetic media, Multimedia forensics, Cross-modal detection, Adversarial robustness

1 Introduction

This section motivates the problem of AI-generated content detection, situates our survey against existing reviews, states our contributions, and outlines the technical challenges and research gaps that structure the remainder of the article.

1.1 Motivation and Importance

Recent breakthroughs in generative artificial intelligence (AI) have transformed digital content creation, enabling increasingly realistic synthesis of text, images, audio, and video (Zellers et al. 2019; Mitchell et al. 2023; Epstein et al. 2023; Shen et al. 2018). Despite differences in modality, many detection approaches rely on shared underlying principles, including statistical irregularities, model-specific artefacts, and inconsistencies in physical or semantic coherence. The widespread accessibility of powerful generative tools, accelerated by landmark releases such as ChatGPT in November 2022, has triggered an unprecedented surge in AI-generated content. The commercial scale of this shift is striking: Bloomberg Intelligence projects the generative AI market to grow from $40 billion in 2022 to approximately $1.3 trillion by 2032, a compound annual growth rate of roughly 42% (Bloomberg Intelligence 2023). While these developments expand opportunities for creativity, education, and operational efficiency, they also enable high-impact misuse, including political manipulation, financial fraud, academic plagiarism, identity theft, and large-scale misinformation (Kietzmann et al. 2020; Tolosana et al. 2020; Mirsky and Lee 2021; Westerlund 2019; Ajder et al. 2019). Reliable detection of synthetic media is therefore one concrete prerequisite for the broader goal of trustworthy, human-aligned AI systems (Sun et al. 2026).

This surge is not merely commercial. Identity-verification data indicate that deepfake fraud incidents grew roughly tenfold worldwide between 2022 and 2023 (Sumsub 2023), and the technology's reach now extends into peer-reviewed science: a population-level analysis of nearly one million papers estimated that by February 2024 up to 17.5% of sentences in computer-science abstracts had been modified by large language models (Liang et al. 2024). The financial stakes are concrete. In early 2024, a finance employee at the engineering firm Arup was deceived into transferring about $25 million after a video call in which the purported chief financial officer and colleagues were all real-time deepfakes (Cable News Network (CNN) 2024). Such incidents

are expected to proliferate: Deloitte's Center for Financial Services projects that generative-AI-enabled fraud losses in the United States could rise from $12.3 billion in 2023 to $40 billion by 2027, a compound annual growth rate of about 32% (Deloitte Center for Financial Services 2024). Collectively, these trends underscore an urgent need for robust detection and provenance mechanisms to preserve digital authenticity, maintain public trust, and mitigate emerging societal harms. As Figure 1 illustrates, the difficulty is not raw detector accuracy in the laboratory but the widening gap between generative capability and detection that survives real deployment.

The technical frontier of 2025–2026 makes this gap concrete. Commercial text-to-video systems now produce clips whose visual quality exceeds that of the open-source generators on which most academic detectors were trained (Feng et al. 2026), and speech-synthesis toolkits cover more than forty languages at million-clip scale (Huang et al. 2025a). Detection has not kept pace outside the laboratory: on deepfakes actually circulated online during 2024, the AUC of state-of-the-art open-source detectors decreases on average by 50% for video, 48% for audio, and 45% for image models relative to the academic benchmarks on which they were originally evaluated (Chandra et al. 2026). At the same time, the surrounding infrastructure is maturing rapidly: production-scale text watermarking has been validated across millions of live responses (Dathathri et al. 2024), provenance tooling that layers watermarks with C2PA content credentials is entering deployment (OpenAI 2026), community challenges now institutionalise robust in-the-wild evaluation (Gushchin et al. 2026), and the transparency obligations of the EU AI Act, which require machine-readable marking of synthetic content, apply from August 2026 (European Parliament and Council of the European Union 2024). Detection research therefore sits at the intersection of rapidly improving generators, demonstrably fragile detectors, and regulation that presumes reliable marking and detection are feasible.

In addition to low-level forensic traces, realistic synthetic media must satisfy *high-level* consistency constraints that are difficult to perfectly emulate at scale. For images and video, this includes physically plausible illumination, shadows, and reflections across objects and frames, where inconsistencies can reveal manipulation (Wu et al. 2022; Guo et al. 2023b). For audio, synthetic speech is often vulnerable to *acoustic-environment* mismatch, such as implausible reverberation and room characteristics under movement or changing recording conditions. For person-centric content, *behavioural cues* (e. g., gaze patterns and motion dynamics) provide complementary signals that are harder to control than pixel-level appearance alone (He et al. 2024).

1.2 Context and Scope Definition

Existing surveys typically focus on single modalities or narrowly defined detection approaches (Verdoliva 2020; Westerlund 2019; Yang et al. 2025). However, the increasing complexity and multimodality of AI-generated content demand an integrated, cross-modal perspective. To address these multifaceted threats, this review surveys detection methods across text, images, audio, and video, emphasising methodological advances and evolving challenges published or publicly available between January 2022 and August 2026. Despite differences in modality, detection methods increasingly rely

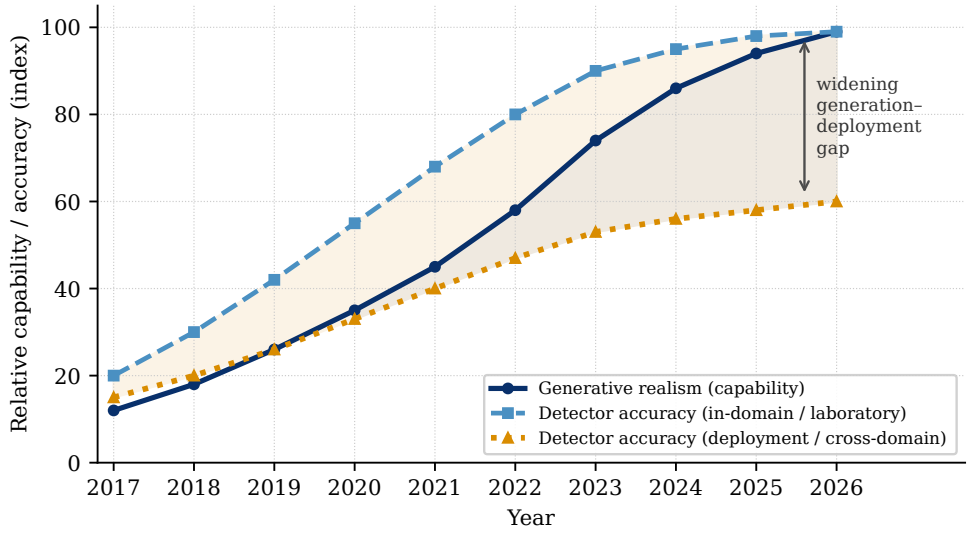


Fig. 1 The widening gap between generative capability and reliable detection. As generative realism has advanced, detector accuracy under controlled (in-domain) conditions has largely kept pace, but accuracy under realistic deployment (unseen generators, domain shift, compression, and post-editing) lags substantially behind. Closing this deployment gap, rather than maximising laboratory accuracy, is the central challenge we foreground in this survey. (Curves are illustrative of the qualitative trend rather than a single measured benchmark.)

on shared principles, including artefact-based signals, model-driven scoring, and consistency constraints across content and context. Unlike prior surveys that treat each modality in isolation, we organise the literature *by shared detection technique*, surfacing cross-domain connections, clarifying overlapping methodological trends, and highlighting opportunities for interdisciplinary progress in synthetic media detection.

1.3 Relation to Prior Surveys

Most existing surveys are organised *by modality* and stop short of analysing how detection signals transfer across modalities. The closest prior works also cover several modalities but retain a per-modality organisation: Lin et al. (2024) span text, image, audio, and video but structure methods by modality and by a pure-detection-versus-beyond-detection split; Xu et al. (2026) classify forensic cues as external or internal yet apply this taxonomy to each modality separately; and Nguyen-Le et al. (2026) pair a method-family review of image, video, and audio deepfake detection with an empirical generalisation study, while leaving text aside. Other surveys are narrower in scope: some focus on a single modality such as text (Wu et al. 2025a; Jawahar et al. 2020; Fariello et al. 2025), audio (Yi et al. 2023; Almutairi and Elgibreen 2022), or video deepfakes (Kaur et al. 2024), some on a single architecture family such as vision transformers for deepfake detection (Wang et al. 2024d), others on generation rather than detection (Cao et al. 2025b), on security and privacy of generative data (Wang et al. 2024a), or on watermarking alone, whether for images (Cao

et al. 2025a) or for large language models (Ye et al. 2026); the recent benchmark-and-survey of Pei et al. (2026) is similarly confined to facial image and video deepfakes. In contrast, this survey is organised *by detection technique*: each major family of methods (statistical/likelihood scoring, frequency-domain analysis, artefact and fingerprint forensics, transformer-based representation learning, consistency checks, and active watermarking/provenance) is examined across all four modalities, making explicit where a technique transfers cleanly and where modality-specific assumptions break down. Section 2 formalises this view as a four-axis taxonomy and instantiates it over the surveyed literature in Table 3. Table 1 summarises this comparison against the most closely related recent surveys.

1.4 Technical Challenges and Research Gaps

Despite rapid progress, current detection methodologies face critical limitations. Reported in-domain performance can exceed 99% for some detectors, yet substantial degradation has been documented under cross-domain, adversarial, and post-processing shift (Jisc 2025); Section 10 quantifies these gaps study by study. Persistent challenges include adversarial robustness, weak cross-model and cross-domain generalisation, limited and biased datasets, and systematic disparities that can elevate false positives for non-native writing styles and accented speech (Liang et al. 2023). Importantly, these technical shortcomings increasingly intersect with ethical and regulatory concerns (e. g., accountability, transparency, and compliance requirements), reinforcing the need for rigorous evaluation protocols and trustworthy detection pipelines beyond laboratory benchmarks. Bridging this laboratory-to-practice divide is essential for transitioning detection technologies from controlled benchmarks to robust real-world deployment.

1.5 Contributions and Roadmap

To address these gaps, this review makes the following contributions:

- A **technique-centric cross-modal organisation** of AI-generated content detection that analyses each major detection family across text, image, audio, and video, clarifying shared principles and transferable signals (Sections 4–9).
- A **four-axis cross-modal taxonomy** (objective, paradigm, signal type, generalisation regime), operationalised through a master classification of representative methods (see Section 2).
- A **quantitative characterisation of the laboratory-to-deployment gap**, using study-level comparisons of reported in-domain and cross-domain performance across modalities (Section 10).
- A **critical assessment of persistent technical, ethical, and legal challenges**, including adversarial robustness, open-set generalisation, and fairness treated as a cross-modal requirement.
- **Strategic recommendations and a forward-looking research roadmap** aimed at improving robustness, interpretability, and generalisability.

Table 1 Comparison with the most closely related recent surveys, ordered chronologically. The progression illustrates how the field has expanded from single-modality detection (text, then audio and video) toward broader coverage, while remaining organised *by modality*, *by architecture*, or *by mechanism*. Modality columns indicate substantive coverage of detection methods for that modality (✓ full, • partial/brief, ✗ not covered). Our survey differs less in which modalities it covers than in its *organising principle*: prior work treats each modality (or a single technique family) in isolation, whereas we organise the literature by shared detection technique *across* modalities, making cross-modal transfer the object of study rather than an afterthought.

Survey	**Year**	**Txt**	**Img**	**Aud**	**Vid**	**Organising principle**	**Primary focus / scope**
Jawahar et al. (2020)	2020	✓	✗	✗	✗	by method (text)	Machine-generated *text* detection.
Almutairi & Elgibreen (Almutairi and Elgibreen 2022)	2022	✗	✗	✓	✗	by method (audio)	Modern *audio* deepfake detection methods.
Yi et al. (2023)	2023	✗	✗	✓	✗	by method (audio)	*Audio* deepfake detection.
Kaur et al. (2024)	2024	✗	•	✗	✓	by method (video)	*Video* deepfake detection.
Wang et al. (2024d)	2024	✗	✓	✗	✓	by architecture (ViT)	ViT-based deepfake detectors only.
Wang et al. (2024a)	2024	•	•	•	•	by security property	Security/privacy of generative data (ACM CSUR).
Lin et al. (2024)	2024	✓	✓	✓	✓	by modality	Closest prior work; per-modality, pure-vs-beyond split.
Wu et al. (2025a)	2025	✓	✗	✗	✗	by method (text)	Methods and directions for *LLM-text* detection.
Cao et al. (2025b)	2025	✓	✓	•	•	by modality (*gen.*)	AIGC *generation*, not detection (ACM CSUR).
Cao et al. (2025a)	2025	✗	✓	✗	✗	by mechanism (watermark)	*Image watermarking* for provenance only.
Xu et al. (2026)	2026	✓	✓	✓	✓	by mechanism, per modality	External vs internal forensic cues; modalities treated separately (ACM CSUR).
Ye et al. (2026)	2026	✓	✗	✗	✗	by mechanism (watermark)	*LLM* watermarking and fingerprinting for provenance (ACM CSUR).
Pei et al. (2026)	2026	✗	•	✗	✓	by generation task	Facial deepfake generation and detection benchmark (ACM CSUR).
Nguyen-Le et al. (2026)	2026	✗	✓	✓	✓	by method family	Passive deepfake detection with an empirical generalisation study (AI Review).
This survey	2026	✓	✓	✓	✓	**by technique (cross-modal)**	Each detection family analysed *across* all four modalities.

Detection techniques (rows) analysed across every modality (columns)

	Text	Image	Audio	Video	Section
Statistical & model-based scoring	●	●	●	◐	§4
Frequency-domain analysis	—	●	●	●	§5
Artefact & fingerprint forensics	—	●	●	●	§6
Transformer representation learning	●	●	●	●	§7
Consistency & semantic checks	●	●	●	●	§8
Active watermarking / provenance	●	●	●	●	§9

● treated in depth ◐ treated briefly — sparse or not applicable; an open research direction

Fig. 2 The organising principle of this survey. Rather than treating each modality in a separate chapter, we structure the literature by *detection technique* (rows; one per technical section) and show how each technique recurs across text, image, audio, and video (columns). This technique-by-modality view makes cross-modal transfer, and the sparse regions where a technique has not yet been applied to a modality, directly visible.

The remainder of this survey is organised as follows. Section 2 defines the problem setting, threat model, and the four-axis cross-modal taxonomy. Section 3 consolidates the datasets and benchmarks used across modalities. Sections 4–9 then survey detection methods technique by technique across all four modalities, before Section 10 synthesises the cross-modal picture and the dominant open challenges, Section 11 outlines future directions, and Section 12 concludes. Figure 2 summarises this technique-by-modality organisation at a glance.

2 Problem Setting and Cross-Modal Taxonomy

2.1 Definitions and Scope

To systematically analyse AI-generated content detection across modalities, we introduce a unified problem formulation and cross-modal taxonomy that captures common objectives, signals, and deployment settings. We use *AI-generated content* to denote media whose primary content is synthesised by a generative model from a prompt or conditioning signal (e. g., LLM text generation; text-to-image; text-to-audio; text-to-video). We use *AI-edited* (or *AI-manipulated*) content to denote media that originates from a human-captured source but is partially modified through generative editing operations, such as inpainting/outpainting, face swapping, voice conversion, paraphrasing, style transfer, or background replacement. In real deployments, harmful content frequently arises from *partial* and *localised* edits rather than fully synthetic generation, which elevates the importance of *localisation* and *robustness to post-processing*.

Our scope covers detection and provenance methods for text, image, audio, and video, with emphasis on modern generative pipelines (notably diffusion-based

image/video synthesis and instruction-tuned LLMs) and on deployment-relevant evaluation settings (open-set generators, domain shift, and post-editing). Detection cannot be understood in isolation from the generation it must keep pace with; for the audio modality in particular, the threat surface is shaped by rapid advances in neural speech synthesis (Tan et al. 2021) and voice conversion (Sisman et al. 2021), and the resulting detection landscape has been surveyed in its own right (Yi et al. 2023).

2.2 Problem Formulations and Output Requirements

Given an input instance x (a document, image, audio segment, or video), the fundamental task is *binary detection*: decide whether x is human-originated or AI-generated/edited. Practical systems, however, often require richer outputs:

- **Open-set detection:** the generator/manipulation is unknown or unseen during training.
- **Attribution:** identify the model, family, or manipulation type (e. g., GAN vs diffusion; voice conversion vs TTS).
- **Localisation:** identify manipulated spans (text), regions (image), or segments (audio/video), optionally with pixel/temporal masks.
- **Provenance verification:** validate origin claims and transformation history via credentials, signatures, and traceable editing chains.

These tasks differ in supervision requirements, failure modes, and evaluation metrics; accordingly, we treat them as distinct objectives rather than minor variants of binary detection.

2.3 Threat Model and Attacker Capabilities

We assume an adaptive attacker whose objective is to maximise perceived authenticity and evade automated screening. Capability levels span:

- **Off-the-shelf generation:** direct use of publicly available models and prompts.
- **Benign post-processing:** compression, resizing, cropping, filtering, re-encoding, noise injection, or format conversion.
- **Deliberate evasion:** adversarial perturbations, detector-aware optimisation, paraphrasing/rewriting loops (for text), regeneration, model fine-tuning, and watermark/provenance stripping or spoofing.

This threat model motivates evaluating detectors under realistic transformation pipelines and distribution shifts, rather than relying exclusively on curated laboratory benchmarks.

2.4 Detection Paradigms: Passive, Active, and Hybrid

AI-generated content detection methods can be broadly categorised into three complementary paradigms. This categorisation forms one axis of our cross-modal taxonomy. We distinguish:

- **Passive forensics (content-only):** post-hoc inference without generator cooperation. Signals include model artefacts (e. g., spectral traces, noise residual inconsistencies, temporal jitter), and semantic/consistency cues (e. g., discourse coherence in text, audio–visual synchronisation in video).
- **Active signals (provenance/watermarking):** verifiable marks or metadata attached at generation/capture time, such as embedded watermarks, signed generation logs, or content credentials.
- **Hybrid systems:** combine passive cues with active verification when available; fall back to forensics when metadata is missing, stripped, or untrusted.

Passive methods offer broad backward compatibility, since they require no cooperation from the generator, but they tend to be brittle under post-editing and model evolution, degrading sharply when the test distribution departs from training (Zhu et al. 2023b). Active methods can be highly reliable when adopted, yet remain constrained by deployment incentives, interoperability, and demonstrated vulnerability to watermark removal and forgery attacks (An et al. 2024; Jiang et al. 2025). Hybrid designs, which fall back to forensic analysis when provenance metadata is absent or untrusted, are therefore increasingly attractive for real-world pipelines.

2.5 The Four-Axis Taxonomy

To structure the remainder of the survey, we adopt a cross-modal taxonomy along four axes:

1. **Objective:** detection, open-set detection, attribution, localisation, provenance verification.
2. **Paradigm:** passive, active, hybrid.
3. **Signal type:** artefact-based traces (e. g., frequency/noise/temporal cues), model-based scoring (e. g., likelihood/curvature), semantic/consistency signals (within- and cross-modal), and provenance metadata/signatures.
4. **Generalisation regime:** closed-set (seen models/domains), open-set (unseen generators), and cross-domain (domain shift, post-processing, distribution drift).

This taxonomy, summarised in Table 2, enables unified comparison across modalities and highlights transferable principles, e. g., shared forensic traces across images and video, or transformer-based scoring and representation learning across text and audio, while making explicit where modality-specific assumptions break down. We operationalise the taxonomy in Table 3, which classifies representative methods from all four modalities along these axes; the empty cells of that table directly motivate the open research gaps discussed in Section 10.

Table 2 The four-axis cross-modal taxonomy for AI-generated content detection. Every method, regardless of modality, is located by one choice on each axis; organising the literature this way exposes techniques that recur across text, image, audio, and video.

Objective	**Paradigm**	**Signal type**	**Generalisation regime**
What the detector outputs	*How evidence is obtained*	*Which cue is exploited*	*Which regime is targeted*
Detection	Passive (forensic)	Artefact	Closed-set
Open-set detection	Active (watermark/ provenance)	Model-based	Open-set
Attribution	Hybrid (passive + active)	Semantic/consistency	Cross-domain/robust
Localisation		Provenance metadata	
Provenance verification			

Table 3: Master cross-modal classification of representative detection and provenance methods along the four taxonomy axes of Section 2. Methods are grouped by **signal type** and interleaved across modalities to expose shared structure; the recurrence of the same paradigm and signal type across text, image, audio, and video is the cross-modal regularity we foreground in this survey. **Mod.**: T text, I image, A audio, V video. **Obj.** (objective): D detection, OS open-set, At attribution, L localisation, P provenance. **Par.** (paradigm): Pa passive, Ac active, Hy hybrid. **Gen.** (generalisation regime targeted): C closed-set, O open-set, X cross-domain/robustness. Empty or sparse regions (e. g., active provenance for text and audio, localisation across most modalities, and the scarcity of genuinely hybrid passive–active systems) mark open research gaps discussed in Section 10.

Method	**Mod.**	**Obj.**	**Par.**	**Signal type**	**Gen.**
Artefact-based traces					
GLTR (Gehrmann et al. 2019)	T	D	Pa	Token-rank / frequency artefacts	C
FreqNet (Tan et al. 2024a)	I	D, OS	Pa	Frequency-domain artefacts	O
NPR (Tan et al. 2024b)	I	D, OS	Pa	Up-sampling / pixel artefacts	O
Frank et al. (2020)	I	D	Pa	Frequency (DCT) artefacts	C
Corvi et al. (2023)	I	D, OS	Pa	Diffusion spectral artefacts	O
SPAI (Karageorgiou et al. 2025)	I	D, OS	Pa	Spectral self-supervision	O, X
APN (Meng et al. 2024)	I	D, OS	Pa	Purified frequency artefacts	O
UGAD (Alam et al. 2024)	I	D, OS	Pa	Radial / Fourier fingerprints	O
Synthbuster (Bammey 2024)	I	D	Pa	Spectral (diffusion) artefacts	O

continued on next page

Table 3 continued from previous page

Method	Mod.	Obj.	Par.	Signal type	Gen.
MesoNet (Afchar et al. 2018)	V	D	Pa	Mesoscopic CNN artefacts	C
GAN fingerprints (Marra et al. 2019)	I	At	Pa	Model-specific noise residual	C
Attributing GANs (Yu et al. 2019)	I	At	Pa	Learned GAN fingerprints	C
FingerprintNet (Jeong et al. 2022)	I	D, At	Pa	Generator fingerprint spectrum	O
LGrad (Tan et al. 2023)	I	D, OS	Pa	Gradient-based artefacts	O
PRNU (Lukáš et al. 2006)	I	At	Pa	Sensor pattern noise	C
AASIST (Jung et al. 2022)	A	D	Pa	Spectro-temporal graph attention	C
F0+spectrogram CM (Xue et al. 2022)	A	D	Pa	f_0 + spectral magnitude/phase	C
Vocoder-cue CM (Kawa et al. 2023)	A	D	Pa	Vocoder / Whisper-feature cues	O
Model-based scoring					
DetectGPT (Mitchell et al. 2023)	T	D, OS	Pa	Likelihood curvature	O
Fast-DetectGPT (Bao et al. 2024)	T	D, OS	Pa	Conditional-prob. curvature	O
DNA-GPT (Yang et al. 2024)	T	D, OS	Pa	Divergent n-gram likelihood	O
DetectLLM (Su et al. 2023a)	T	D, OS	Pa	Log-likelihood / log-rank	O
TDT (West et al. 2025)	T	D, OS	Pa	Temporal likelihood dynamics	O, X
OOD likelihood (Zeng et al. 2025)	T	OS	Pa	Out-of-distribution scoring	O, X
DIRE/ AEROBLADE (Ricker et al. 2024)	I	D, OS	Pa	Reconstruction error	O
LaRE^2 (Luo et al. 2024)	I	D, OS	Pa	Latent reconstruction error	O
DRCT (Chen et al. 2024)	I	D, OS	Pa	Recon. contrastive scoring	O
Learned representations (supervised / transformer / SSL)					
RoBERTa detector (Guo et al. 2023a)	T	D	Pa	Fine-tuned encoder features	C
ImBD (Chen et al. 2025a)	T	D, OS	Pa	Style-preference-aligned scoring	O

continued on next page

Table 3 continued from previous page

Method	Mod.	Obj.	Par.	Signal type	Gen.
EAGLE (Bhattacharjee et al. 2024)	T	D, OS	Pa	Generator-invariant DG features	O, X
MGT-Prism (Liu et al. 2026)	T	D	Pa	Spectral DG representation	X
UnivFD (Ojha et al. 2023)	I	D, OS	Pa	Frozen CLIP features	O, X
C2P-CLIP (Tan et al. 2025)	I	D, OS	Pa	Category-aware CLIP	O, X
CO-Spy (Cheng et al. 2025a)	I	D, OS	Pa	Semantic + artefact fusion	O, X
B-Free (Guillaro et al. 2025)	I	D, OS	Pa	Bias-free self-supervision	O, X
FatFormer (Liu et al. 2024)	I	D, OS	Pa	Forgery-aware transformer	O
GAPL (Qin et al. 2026)	I	D, OS	Pa	Generator-aware prototypes	O, X
SSL-ViT (Nguyen et al. 2024)	I	D	Pa	Self-supervised ViT features	O
wav2vec2+AASIST (Tak et al. 2022)	A	D, OS	Pa	SSL speech representations	O, X
Spoof-SUPERB (Ali et al. 2026)	A	D	Pa	SSL front-end benchmark	O
Affective Bridge (Li et al. 2025a)	A	D, OS	Pa	Unified affective speech features	O, X
AIGI-Holmes (Zhou et al. 2025)	I	D, L	Pa	MLLM features + explanation	O
FakeShield (Xu et al. 2025b)	I	D, L	Pa	MLLM-explainable localisation	O
BusterX++ (Wen et al. 2025a)	V	D, At	Pa	MLLM cross-modal reasoning	O
GenD (Yermakov et al. 2026)	I	D, OS	Pa	CLIP LayerNorm adaptation	O, X
Semantic / consistency signals					
FakeCatcher (Ciftci et al. 2020)	V	D	Pa	Biological (rPPG) signals	C
LipForensics (Haliassos et al. 2021)	V	D, OS	Pa	Mouth-motion consistency	O
Lip-sync detect (Datta et al. 2025a)	V	D, L	Pa	Audio–visual sync	O
MM-Det (Song et al. 2024)	V	D	Pa	Audio–visual joint cues	O
Active: watermarking and provenance					
KGW watermark (Kirchenbauer et al. 2023)	T	P, At	Ac	Logit-bias watermark	O, X

continued on next page

Table 3 continued from previous page

Method	Mod.	Obj.	Par.	Signal type	Gen.
SynthID-Text (Dathathri et al. 2024)	T	P, At	Ac	Sampling watermark	O, X
Tree-Ring (Wen et al. 2023)	I	P, At	Ac	Latent-space watermark	O, X
RingID (Ci et al. 2024)	I	P, At	Ac	Multi-key latent watermark	O, X
WAVES benchmark (An et al. 2024)	I	P	Ac	Watermark robustness eval	X
AudioMarkNet (Zong et al. 2025)	A	P, At	Ac	Audio watermark	O, X
Video Seal (Fernandez et al. 2024)	V	P, At	Ac	Neural video watermark	O, X
VideoMarkBench (Jiang et al. 2025)	V	P	Ac	Video watermark robustness	X
C2PA credentials (Coalition for Content Provenance and Authenticity (C2PA) 2024)	V	P	Ac	Signed provenance metadata	X
OpenAI provenance (OpenAI 2026)	I	P	Ac	Provenance classifier + C2PA	X

Reading the table by *column sparsity* rather than by row turns it from a catalogue into a source of research questions: several cells are empty not because a technique is inapplicable, but because it has not yet been ported across modalities. Three examples recur in our synthesis. First, *localisation* (the L objective) appears almost exclusively in image methods built on multimodal LLMs (e. g., FakeShield (Xu et al. 2025b)), with no mature analogue for *temporal* localisation in audio or *span-level* localisation in text; yet real-world harm increasingly arises from *partial* edits, making "where was it manipulated?" an open cross-modal problem (Section 10). Second, *physiological* consistency cues such as rPPG are confined to a single video method (FakeCatcher (Ciftci et al. 2020)) and remain closed-set only, leaving open whether comparable biologically grounded signals exist for audio (e. g., breathing or vocal-tract dynamics) and whether they generalise. Third, *active provenance* is dense for text, image, and video watermarking but thin for audio attribution, where AudioMarkNet (Zong et al. 2025) is largely alone; robust, removal-resistant audio watermarking is therefore an identifiable gap rather than a solved problem. In each case the empty cell names a concrete direction, and we return to these in Section 10.

2.6 Literature Collection Methodology

To balance breadth with reproducibility, we followed a structured literature-search and selection protocol, informed by PRISMA principles but not intended as a formal systematic review. We queried six scholarly sources (*IEEE Xplore*, the *ACM Digital Library*, *arXiv*, *Scopus*, *Google Scholar*, and *Semantic Scholar*), covering the period from January 2022 to August 2026, supplemented by a small number of foundational pre-2022 works (e. g., early GAN-fingerprint and PRNU forensics) retained for historical context.

Search strings.

We combined a detection/provenance term with a modality term, for example:

> (`"AI-generated content detection"` OR `"deepfake detection"` OR `"synthetic media detection"` OR `"machine-generated text detection"` OR `"watermarking"` OR `"content provenance"`) AND (`text` OR `image` OR `audio` OR `video` OR `"cross-modal"`)

These were augmented with method-specific queries (e. g., *DetectGPT*, *diffusion reconstruction error*, *spectral artefact*, *CLIP detector*, *audio anti-spoofing*, *audio–visual synchronisation*, *C2PA*) and with backward/forward citation chasing from seminal papers and widely used benchmarks.

Inclusion and exclusion criteria.

We included (i) peer-reviewed papers from major venues in machine learning, computer vision, speech, natural language processing, and multimedia security; (ii) high-impact preprints (typically arXiv) when they introduced influential methods, benchmarks, or results not yet in archival form; and (iii) widely adopted datasets, benchmarks, and provenance standards. We excluded non-technical or non-peer-reviewed commentary, vendor marketing material, and works without a reproducible methodological contribution. When multiple papers addressed the same idea, we retained representative methods and those with the most complete evaluations. The resulting corpus emphasises detection *techniques* that recur across modalities, consistent with the cross-modal organisation of this survey; it is intended to be representative of the state of the art rather than exhaustive.

Selection pipeline.

Figure 3 summarises the selection process as a structured literature-selection flow. The initial queries returned approximately 1 240 records across the six databases. After removing 310 duplicates, approximately 930 records were screened on title and abstract, of which 620 were excluded as off-topic or unrelated to detection and provenance. The remaining 310 records underwent full-text assessment, after which 79 were excluded for lacking a reproducible methodological contribution, being superseded by later work, or not being peer reviewed. This yielded the 231 sources synthesised in this survey. The counts are approximate and reflect a structured rather than exhaustive protocol, but the pipeline is reported so that the selection can be understood and reproduced.

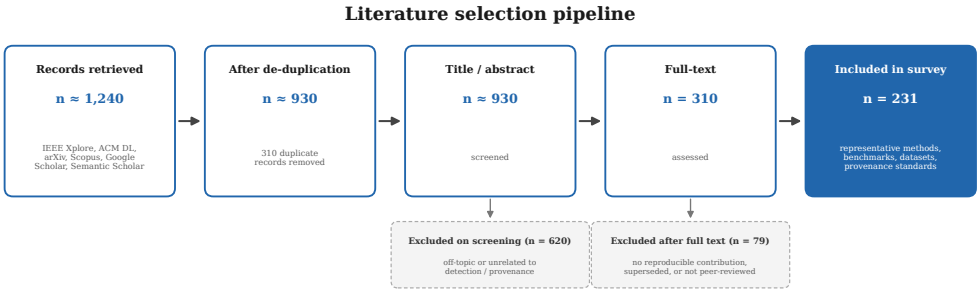


Fig. 3 Structured literature-selection flow, from records retrieved across six databases through de-duplication, title/abstract screening, and full-text assessment to the final set of sources included in the survey. Counts are approximate and reflect a structured rather than exhaustive protocol.

3 Datasets and Benchmarks

Progress in detection is bounded by the data on which detectors are trained and evaluated. Because the central difficulty identified throughout this survey is *generalisation*, the most informative benchmarks are those that deliberately span multiple generators, domains, and post-processing conditions rather than a single source. This section consolidates the datasets and benchmarks most frequently used across the four modalities; Tables 4 and 5 summarise their scale, composition, and the evaluation challenge each is designed to expose. We organise the discussion by modality, but note that the field is increasingly moving toward unified, cross-modal resources, which we discuss last. Figure 4 maps the same resources by modality, scale, and year; the visible clustering of large multilingual and cross-modal corpora in 2025–2026 is the clearest structural signal of this shift.

3.1 Text

Text benchmarks have moved decisively from single-generator to multi-generator, multi-domain settings. HC3 (Guo et al. 2023a) established an early, widely used baseline by pairing human expert answers with ChatGPT responses, and HC3 Plus (Su et al. 2023b) extends it to semantic-invariant tasks such as summarisation, translation, and paraphrasing, which are harder to detect than question answering. Both remain single-generator, however, limiting what they reveal about transfer. M4 (Wang et al. 2024c) and MAGE (Li et al. 2024a) were introduced precisely to expose this limitation: both span many generators and domains (and, for M4, multiple languages, with roughly 147 k human–machine parallel pairs alongside over 10 million human texts), and both report sharp accuracy drops when detectors face an unseen domain or generator. RAID (Dugan et al. 2024) couples millions of generations with eleven adversarial attacks and decoding variations; its 2026 successor DetectRL-X (Wu et al. 2026b) extends robustness evaluation to eight languages and six misuse-prone domains, and TSM-Bench (Quaremba et al. 2026) demonstrates that task-specific generation, e. g., summarisation in real Wikipedia editing, evades current detectors far more effectively than generic prompting. Complementing these, we release two resources used in our own evaluations: an academic-domain corpus of 669 k arXiv, GPT-3.5, and Gemini

Table 4 Representative datasets and benchmarks for AI-generated content detection: text and image resources, ordered by year within each group (audio, video, and cross-modal resources follow in Table 5). "Real / Fake" gives the human-or-real versus AI-or-fake sample split where the resource reports one (a dash indicates a split that the benchmark does not define, typically evaluation-only or cross-modal sets); "Lang." gives the principal language(s) for text and speech resources (– for vision-only sets). "Focus" summarises the generalisation challenge each resource is designed to probe. Rows marked † are released by the present authors; the HC3 entries are among the most widely used text baselines and underpin several of our experiments.

Dataset	Mod.	Year	Scale	Real / Fake	Lang.	Focus
HC3 (Guo et al. 2023a)	Text	2023	85 k QA	58 k / 26 k	EN, ZH	Single-generator baseline (vs ChatGPT)
HC3 Plus (Su et al. 2023b)	Text	2023	210 k	105 k / 105 k	EN, ZH	Semantic-invariant tasks (summarise, translate)
M4 (Wang et al. 2024c)	Text	2024	147 k	147 k / 74 k	9 langs	Multi-generator, multi-domain, multilingual
MAGE (Li et al. 2024a)	Text	2024	447 k	154 k / 294 k	EN	27 LLMs, 7 domains, in the wild
RAID (Dugan et al. 2024)	Text	2024	6 M+	15 k / 509 k	EN	11 adversarial attacks, decoding shift
HC3-Gemini† (Mady 2025)	Text	2025	23 k	0 / 23 k	EN	Cross-generator probe (Gemini 2.0 Flash)
DetectRL-X (Wu et al. 2026b)	Text	2026	3.5 M	144 k / 144 k	8 langs	Multilingual misuse domains; core pairs plus 3.2 M attack and augmentation variants
TSM-Bench (Quaremba et al. 2026)	Text	2026	178 k	25 k / 153 k	EN, PT, VI	Task-conditioned generation (Wikipedia editing)
Academic-Text† (Mady et al. 2026a,b)	Text	2026	669 k	469 k / 200 k	EN	Academic (arXiv) domain shift; GPT, Gemini
GenImage (Zhu et al. 2023b)	Image	2023	2.7 M	1.3 M / 1.4 M	–	Cross-generator over 8 GAN/diffusion sources
ArtiFact (Rahman et al. 2023)	Image	2023	2.5 M	964 k / 1.5 M	–	25 generators, with real-world impairments
AI-Face (Lin et al. 2025)	Image	2025	1.65 M	401 k / 1.25 M	–	Demographically annotated; fairness
Chameleon (Yan et al. 2025a)	Image	2025	26 k	14.9 k / 11.2 k	–	Human-indistinguishable in-the-wild images
Community Forensics (Park and Owens 2025)	Image	2025	2.7 M	– / 2.7 M	–	Generator diversity (4 803 generators)
SID-Set (Huang et al. 2025b)	Image	2025	300 k	100 k / 200 k	–	Social-media detection and localisation
HydraFake (Tan et al. 2026)	Image	2026	100 k	50 k / 50 k	–	Hierarchical cross-model/forgery/domain tiers

Table 5 Representative datasets and benchmarks, continued from Table 4: audio, video, and cross-modal resources. Columns as in Table 4; a dash in "Real / Fake" marks a resource that does not report an aggregate real/fake split, e. g., question-based evaluation suites (LOKI), releases whose real counterpart is reused from external corpora (DF40), or benchmarks reporting per-modality rather than aggregate splits (Omni-Fake).

Dataset	Mod.	Year	Scale	Real / Fake	Lang.	Focus
ASVspoof 2021 (Yamagishi et al. 2021; Liu et al. 2023)	Audio	2021	1.7 M	168 k / 1.57 M	EN	LA, PA, and DF tracks; codec and compression shift
WaveFake (Frank and Schönherr 2021)	Audio	2021	134 k	16 k / 118 k	EN, JA	Vocoder artefacts, 6 systems
ASVspoof 5 (Wang et al. 2024b, 2026)	Audio	2024	1 M+	189 k / 815 k	EN	Crowdsourced speech, codecs, adversarial
Codecfake (Xie et al. 2025)	Audio	2025	1.06 M	132 k / 926 k	EN, ZH	Neural-codec (ALM) speech, 7 codecs
SONICS (Rahman et al. 2025)	Audio	2025	97 k songs	48 k / 49 k	EN	Fully AI-generated songs (Suno, Udio)
SpeechFake (Huang et al. 2025a)	Audio	2025	3.6 M	349 k / 3.3 M	46 langs	40 generation tools, multilingual
FaceForensics++ (Rössler et al. 2019)	Video	2019	5 k vid.	1 k / 4 k	–	4 manipulations, compression levels
DFDC (Dolhansky et al. 2020)	Video	2020	128 k vid.	23 k / 104 k	–	Large-scale, in-the-wild face swaps
DF40 (Yan et al. 2024)	Video	2024	1 M+	– / 1.1 M	–	40 forgery techniques; real data reused from FF++/Celeb-DF
Celeb-DF++ (Li et al. 2025c)	Video	2025	54 k vid.	590 / 53 k	–	High-quality forgeries; generalisation
GenVidBench (Ni et al. 2026)	Video	2026	6.8 M	431 k / 6.35 M	–	Fully AI-generated (text-to-video)
CoCoVideo (Feng et al. 2026)	Video	2026	26 k vid.	13 k / 13 k	–	Paired clips from 13 commercial generators
FakeAVCeleb (Khalid et al. 2021)	A/V	2021	20 k+ vid.	500 / 20 k	EN	Audio–visual disagreement
LOKI (Ye et al. 2025)	Multi	2024	18 k+	–	EN	MLLMs as cross-modal detectors
MVAD (Hu et al. 2025a)	V/A	2025	121 k	61 k / 60 k	EN	Joint video–audio detection; four real/fake pairings
AV-Deepfake1M++ (Cai et al. 2025)	A/V	2025	2.1 M vid.	628 k / 1.4 M	EN	Perturbed manipulations; temporal localisation
Ivy-Fake (Jiang et al. 2026)	I/V	2026	111 k	78 k / 53 k	–	Unified, explainable detection
Deepfake-Eval-2024 (Chandra et al. 2026)	Multi	2026	in the wild	2 949 / 2 882	52 langs	Deepfakes circulated online in 2024
Omni-Fake (Li et al. 2026a)	Multi	2026	1.3 M	–	multi.	Unified four-modality benchmark; OOD split
TriDF (Jiang-Lin et al. 2026)	Multi	2026	65 k QA	1:1 pairs	EN	MLLM perception, detection, hallucination

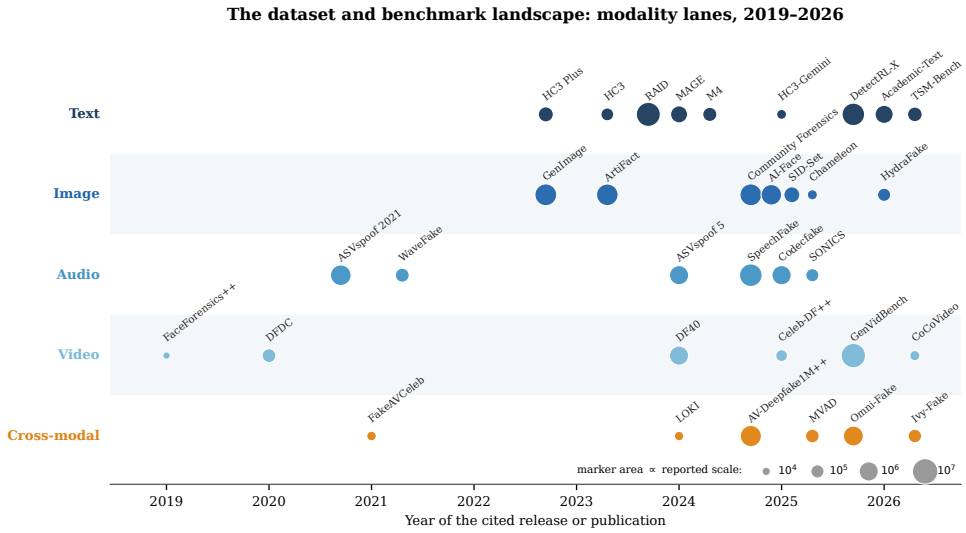


Fig. 4 The dataset and benchmark landscape, consolidating Tables 4 and 5. Each marker is one resource, placed in a lane per modality at the year of its cited release or publication; the marker area is proportional to the reported sample scale (logarithmic). Two structural trends are visible at a glance: overall scale has grown by roughly three orders of magnitude since 2019, and the 2025–2026 region is dominated by multilingual audio corpora and unified cross-modal benchmarks rather than single-modality sets.

paragraphs (Mady et al. 2026a,b) that isolates distribution shift in scientific writing, and a focused set of Gemini 2.0 Flash responses to HC3 questions for cross-generator probing (Mady 2025). Together, these benchmarks operationalise the open-set and adversarial fragility discussed in Section 10.

3.2 Image

Image detection is anchored by two large-scale benchmarks. GenImage (Zhu et al. 2023b) provides over a million images spanning eight GAN and diffusion generators and has become the standard testbed for cross-generator transfer, with the customary protocol of training on one generator (e. g., Stable Diffusion v1.4) and testing on the rest. ArtiFact (Rahman et al. 2023) is larger and more diverse still, with roughly 2.5 million images from 25 generators across GAN, diffusion, and miscellaneous families, and it additionally applies real-world impairments such as JPEG compression and downscaling, making it well suited to robustness evaluation. AI-Face (Lin et al. 2025) adds a complementary axis: as the first million-scale, demographically annotated dataset of AI-generated faces, it underpins the fairness evaluation we return to in Section 10. The Chameleon benchmark, introduced alongside a broader sanity check of the field, curates genuinely human-indistinguishable synthetic images on which nearly all published detectors misclassify fakes as real, recalibrating claims of solved performance (Yan et al. 2025a).

3.3 Audio

Audio anti-spoofing has a mature evaluation tradition centred on the ASVspoof challenge series. Its logical-access track (Yamagishi et al. 2021) remains a standard reference point, while the more recent ASVspoof 5 (Wang et al. 2024b, 2026) raises difficulty substantially by incorporating crowdsourced speech, codec compression, and adversarial attacks; this is where many self-supervised systems exhibit the largest degradation (Section 10). WaveFake (Frank and Schönherr 2021) complements these challenge corpora by isolating vocoder artefacts across six synthesis architectures and two languages, supporting controlled study of frequency-domain cues. Two 2025 corpora close the generator-recency gap: Codecfake extends the controlled-artefact paradigm to the neural-codec speech produced by audio language models, on which vocoder-trained detectors degrade sharply (Xie et al. 2025), and SpeechFake scales coverage to more than three million clips from 40 recent synthesis toolkits across 46 languages (Huang et al. 2025a).

3.4 Video

Video benchmarks reflect the field's shift from face manipulation toward fully synthetic generation. FaceForensics++ (Rössler et al. 2019) and the large-scale DeepFake Detection Challenge corpus (Dolhansky et al. 2020) established the face-swap setting at scale and across compression levels, and Celeb-DF++ (Li et al. 2025c) was constructed specifically to stress cross-dataset generalisation with high-quality forgeries. DF40 (Yan et al. 2024) sharpens the generalisation challenge further by assembling 40 distinct forgery techniques spanning face swapping, reenactment, entire-face synthesis, and editing, explicitly so that detectors can be tested against generators unseen during training. As text-to-video generators have matured, GenVidBench (Ni et al. 2026) extends evaluation to millions of fully AI-generated clips, a setting distinct from the manipulation-centric datasets that preceded it, and CoCoVideo (Feng et al. 2026) raises the bar further by pairing real footage with clips from high-quality commercial generators, whose visual quality exceeds the open-source systems underlying earlier corpora.

3.5 Cross-Modal and In-the-Wild Resources

The clearest recent trend is toward resources that span modalities or capture realistic deployment. FakeAVCeleb (Khalid et al. 2021) pairs audio and visual deepfakes, enabling detectors to exploit modality disagreement of the kind discussed in Section 8, and AV-Deepfake1M++ scales this paradigm to over two million clips with real-world perturbations and temporal-localisation labels (Cai et al. 2025). LOKI (Ye et al. 2025) evaluates multimodal large language models as general-purpose detectors across modalities, while Ivy-Fake (Jiang et al. 2026) and MVAD (Hu et al. 2025a) unify image-and-video and video-and-audio detection, respectively, under single benchmarks. Finally, Deepfake-Eval-2024 (Chandra et al. 2026) collects deepfakes actually circulated online during 2024, offering an in-the-wild measure of the

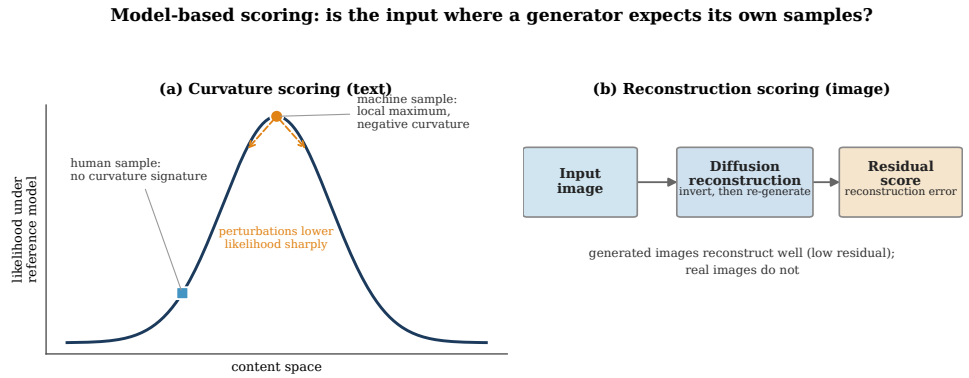


Fig. 5 The shared principle of model-based scoring. (a) Curvature scoring for text: machine-generated samples sit near local maxima of the reference model's likelihood surface, so small perturbations sharply lower their likelihood, whereas human text exhibits no such curvature signature. (b) Reconstruction scoring for images: an image is inverted and re-generated by a diffusion model, and a low reconstruction error indicates that the image lies where the generator expects its own samples. Both forms ask the same question of different generators, and both weaken as generators close the gap to natural data.

laboratory-to-deployment gap that curated datasets, by construction, tend to understate. These cross-modal resources are the natural evaluation substrate for the technique-by-modality view we advocate in this survey.

4 Statistical and Model-Based Scoring

A foundational family of detectors asks a single cross-modal question: *how typical is the input under a reference generative model?* In text this takes the form of likelihood, perplexity, and curvature statistics; in images it reappears as diffusion reconstruction-error scoring; and in audio it underlies the simplest statistical acoustic cues. The shared appeal is that these methods need no labelled training set and, in their zero-shot forms, no supervised classifier at all. The shared weakness is equally cross-modal: model-based scores are fragile under domain mismatch, short inputs, and light post-editing or paraphrasing, and they degrade as generators evolve to better match the reference distribution. Figure 5 sketches the shared principle in its two dominant forms.

4.1 Likelihood, Perplexity, and Token-Rank Signals (Text)

A classic approach exploits differences in likelihood structure under a reference language model, using perplexity or related likelihood statistics. Such signals can separate some synthetic samples from human text, especially when the generator is weaker or closely matched to the detector model. Early work framed this as an arms race between generator and detector: the Grover model showed that the best discriminator of neural fake news was often the generator itself (Zellers et al. 2019), while feature-based classifiers built on carefully engineered linguistic statistics offered a lightweight alternative to model-based scoring on GPT-2/GPT-3/Grover text (Fröhling and Zubiaga 2021). In practice, these methods are fragile: performance can collapse under

domain mismatch, short inputs, or stylistic variation, and can be strongly affected by light post-editing or paraphrasing (Radford et al. 2019; Sadasivan et al. 2025). Human-in-the-loop tools such as GLTR visualise token-rank patterns and sampling artefacts (Gehrmann et al. 2019), but modern instruction-tuned LLMs can suppress obvious rank-based footprints, limiting standalone reliability.

4.2 Zero-Shot Curvature and Perturbation Methods (Text)

A major shift since 2022 is the emergence of *zero-shot* detectors that avoid training a supervised classifier. DetectGPT hypothesises that machine-generated text lies in regions of *negative curvature* of the log-probability surface of a language model; by applying controlled perturbations and comparing likelihood changes, it enables zero-shot detection without supervised training (Mitchell et al. 2023). Fast-DetectGPT improves the computational efficiency of curvature-based detection via more efficient estimation, making zero-shot detection more practical at scale (Bao et al. 2024). A parallel line dispenses with perturbation entirely: DetectLLM exploits log-rank statistics (Su et al. 2023a), and DNA-GPT measures divergence between truncated-and-regenerated continuations (Yang et al. 2024). More recent work moves beyond a single summary score, Temporal Discrepancy Tomography analyses token-level anomaly signals across morphological, syntactic, and discourse scales via wavelet transforms (West et al. 2025), Lastde++ mines the local dynamics of the token-probability sequence as a time series, improving zero-shot robustness to paraphrasing (Xu et al. 2025a), while a complementary reframing treats human text as the diverse out-of-distribution class rather than assuming a coherent human distribution, improving generalisation to unseen authors (Zeng et al. 2025); a further reframing scores text under the implicit reward model recoverable from preference-tuned LLMs, targeting the alignment stage rather than the pre-training distribution (Liu et al. 2025a). Nevertheless, robustness to domain shift and strong post-editing remains a central limitation under realistic deployment conditions (Sadasivan et al. 2025; Lu et al. 2024).

4.3 Reconstruction-Error Scoring (Image)

The same model-based principle transfers to images through *diffusion* dynamics. Because diffusion models introduce traces that GAN-era detectors miss (Rombach et al. 2022), a complementary line of work scores how well an image is reconstructed by a diffusion process: reconstruction-error-based approaches such as DIRE leverage diffusion reconstruction characteristics to expose synthetic origin (Croitoru et al. 2023). This idea has rapidly matured: AEROBLADE performs training-free detection using the autoencoder reconstruction error inherent to latent diffusion models (Ricker et al. 2024), LaRE^2 shifts the analysis to latent space for efficiency (Luo et al. 2024), DRCT applies contrastive training on reconstruction pairs to learn generator-invariant features that transfer across unseen diffusion models (Chen et al. 2024), and FIRE hardens reconstruction-error scoring by isolating the mid-band frequencies that diffusion models cannot faithfully reconstruct (Chu et al. 2025). Conceptually this mirrors curvature-based text detection, both ask whether the input sits where a generative

model expects its own samples to lie, and both inherit the same sensitivity to heavy editing and to distribution shift across generator families.

4.4 Low-Level Statistical Acoustic Cues (Audio)

At the statistical end of the audio spectrum, basic acoustic descriptors such as fundamental frequency (F0) and jitter have been considered, although they are less commonly adopted in isolation because such characteristics can be more readily mimicked by AI-generated audio (Xue et al. 2022). As in text and image, these low-level statistical cues are most useful as one signal among several rather than as a standalone detector; richer feature- and learning-based audio methods are treated in Sections 5 and 7. The reason these cues are weak in isolation is itself the cross-modal lesson: like text perplexity and image reconstruction error, F0/jitter statistics assume the generator has failed to match a low-order property of real data, an assumption modern TTS increasingly violates by design, since prosodic naturalness is an explicit training objective. The signal therefore attenuates exactly as the generator improves, which is why it survives only as one input among several.

Cross-Modal Takeaway

Model-based scoring is attractive precisely because it is generator-aware and training-light, and the curvature/reconstruction analogy shows it is genuinely cross-modal rather than a text-only trick; the family has now reached video as well, where training-free spatial–temporal likelihood scoring detects generated clips without any detector training (Ben Hayun et al. 2026). Its transferable failure mode is equally important: as generators close the gap to the reference distribution, and as benign post-processing perturbs the input, the very signal these methods rely on attenuates. This motivates the complementary, more robust signal families surveyed next.

5 Frequency-Domain Analysis

Frequency-domain analysis is arguably the clearest example of a detection technique that recurs across modalities. Generative pipelines leave characteristic signatures in the spectral domain: GAN upsampling produces atypical Fourier patterns in images, vocoders and synthesis models shape the time–frequency structure of audio, and re-encoding and synthesis leave compression and Fourier footprints in video. The shared strength is that these cues are often highly discriminative in controlled settings; the shared weakness, equally cross-modal, is that they can be attenuated or "washed" by post-processing, re-encoding, and adversarial laundering, so spectral cues are typically most effective when fused with complementary spatial, residual, or temporal signals. Figure 6 illustrates the underlying signature.

5.1 Spectral Signatures in Images

A strong line of work analyses spectral signatures introduced by generative pipelines, particularly GAN upsampling, which may create atypical frequency patterns compared to natural images (Durall et al. 2020; Dzanic et al. 2020). Diffusion models leave their

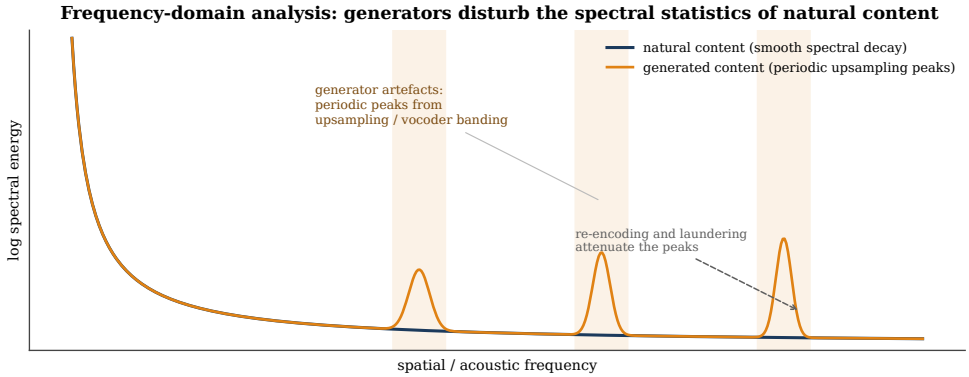


Fig. 6 The frequency-domain signature exploited across image, audio, and video detection. Natural content exhibits a smooth spectral decay, whereas generative pipelines superimpose periodic peaks (from upsampling in CNN generators, or banding in neural vocoders) on that envelope. The same mechanism explains the technique's shared weakness: re-encoding, resizing, and adversarial laundering attenuate exactly these peaks. The curves are schematic illustrations of the qualitative effect rather than measured spectra.

own spectral fingerprints, characterised in detail by Corvi et al. (2023), and Synthbuster exploits these periodic frequency artefacts to detect diffusion images (Bammey 2024). Recent methods make spectral cues more transferable: SPAI treats the spectral distribution of real images as a self-supervised, generator-invariant pattern and flags generated images as spectral outliers at any resolution (Karageorgiou et al. 2025), the Artefact Purification Network separates artefact features from semantic content via suspicious frequency-band proposals (Meng et al. 2024), and UGAD combines radial-integral and spatial-Fourier features for universal frequency-fingerprint detection across GAN and diffusion sources (Alam et al. 2024). Frequency-based detectors can be highly effective in controlled settings, but they may be vulnerable to post-processing and adversarial "washing" that attenuates spectral artefacts (Wesselkamp et al. 2022). For this reason, spectral cues are typically most useful when combined with complementary spatial or residual signals.

5.2 Time–Frequency Features in Audio

In audio, the frequency domain is the dominant representation for the classical detection pipeline: handcrafted acoustic features are extracted and fed to a relatively simple classifier (e. g., SVM, logistic regression, or an MLP). Constant-Q Cepstral Coefficients (CQCC) are derived via the Constant-Q transform and provide compact, informative representations for AI-generated audio (AGA) detection (Wang et al. 2017), while Mel-Frequency Cepstral Coefficients (MFCCs), designed to reflect human auditory perception, are even more widely used (Wijethunga et al. 2020). Spectrogram-based features such as Mel-spectrograms, spectral bandwidth, and spectral centroid encode richer time–frequency information and have also been explored (Jbara and Soud 2025). Music-specific deepfake detection adds structural features such as beats and rhythmic patterns, while otherwise reusing general audio features (Li et al. 2024b); the SONICS

benchmark extends this setting to fully AI-generated songs, showing that long-range spectro-temporal modelling is needed to catch holistic fakes whose vocals, lyrics, and instrumentals are all synthetic (Rahman et al. 2025).

Given the wide range of available features, there is no consensus on which are most suitable, and the effectiveness of combined feature sets varies across datasets. The deeper point is why no single feature wins: handcrafted descriptors localise different artefacts (constant-Q resolution favours high-frequency traces, Mel/linear spacing favours low-frequency ones), so feature choice implicitly bets on where a given generator leaks. This is the audio instance of a cross-modal pattern: frequency-domain detection works precisely when generator artefacts are spectrally concentrated, and fails by the same mechanism (benign re-encoding and resampling redistribute or erase those bands) as in the image and video cases. A further complication specific to speech is speaker variation: accent, dialect, and informal pronunciation shift the very spectral and prosodic distributions that these features measure, so a detector tuned on one accent group can misread natural variation in another as evidence of synthesis (or the reverse), which is both a robustness and a fairness concern when some accents are under-represented in training corpora. Nevertheless, these features represent an important foundation and suggest a useful design principle: AI-generated audio should not be trivially separable on explicit acoustic cues alone if the goal is realism, yet effective detectors should continue to account for such cues.

5.3 Frequency and Compression Cues in Video

Spectral cues (e. g., Fourier magnitude patterns) and compression footprints may expose generator-specific artefacts or manipulation traces in video. As in the image and audio cases, these signals can be informative but are susceptible to attenuation via re-encoding, resizing, filtering, and adversarial laundering; therefore, robustness-aware evaluation is essential. Two recent results sharpen this picture: forensic-oriented augmentation that suppresses semantic shortcuts in favour of low-level spectral fingerprints markedly improves generalisation to unseen video generators (Corvi et al. 2025), and even the detector's own resize-and-crop preprocessing can destroy the relevant traces, motivating native-resolution architectures that classify videos at their original spatial and temporal scale (Li et al. 2026b).

Cross-Modal Takeaway

Across images, audio, and video, frequency analysis exploits the same insight, that generators perturb the spectral structure of natural signals, and runs into the same limitation, namely fragility under post-processing that reshapes those spectra. The transferable design lesson is consistent: spectral cues are best treated as one branch of a multi-signal detector rather than a standalone test, a pattern we see realised in the hybrid spatial+frequency image architectures of Section 7 and the multimodal video systems of Section 8.

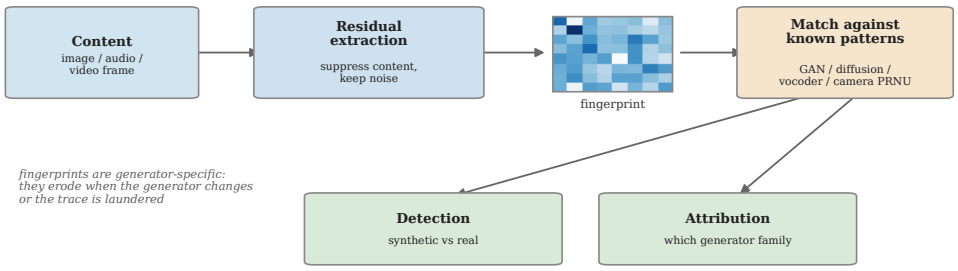


Fig. 7 The artefact and fingerprint forensics pipeline shared by image, audio, and video detection. Content-suppressing residual extraction isolates a noise-like fingerprint, which is matched against known generator, vocoder, or camera (PRNU) patterns; the match supports both detection and source attribution. The strength of this family, source attribution, is inseparable from its weakness: fingerprints are generator-specific, so they erode when the generator changes or when the trace is laundered by post-processing.

6 Artefact and Fingerprint Forensics

A second passive family infers synthetic origin from *artefacts*: low-level traces left by a specific generator or capture pipeline. These traces act as fingerprints: GAN architectures leave spectral and residual signatures, vocoders imprint consistent artefacts on synthesised speech, and camera pipelines leave sensor-noise patterns whose violation can betray a fake. The cross-modal strength is interpretability and the possibility of *attribution* (identifying the source family); the cross-modal weakness is that fingerprints are generator-specific and tend to degrade when the generative mechanism changes (e. g., GAN→diffusion) or when post-processing launders the trace. Figure 7 summarises the pipeline.

6.1 GAN Fingerprints and Model Attribution (Image)

GANs can embed model-specific artefacts ("fingerprints") arising from architectural choices and training procedures (Yu et al. 2019; Marra et al. 2019). This idea has deep roots in classical media forensics, where sensor pattern noise (PRNU) was shown to uniquely identify the camera that captured an image (Lukáš et al. 2006); learned generator fingerprints are the synthetic-media analogue, and the camera-pipeline idea itself has been modernised into a zero-shot test of whether an image exhibits the RAW–RGB alignment traces a physical sensor pipeline would leave (Wu et al. 2026a). Early learned detectors such as MesoNet already showed that compact CNNs can pick up mesoscopic facial-forgery traces (Afchar et al. 2018), and high-fidelity generators like StyleGAN2 sharpened the need for such artefact analysis by markedly reducing visible defects (Karras et al. 2020). Fingerprint-based detectors exploit these artefacts for detection and attribution. For example, FingerprintNet aims to learn *transferable* fingerprint representations across multiple GAN architectures rather than overfitting to a single generator, improving cross-architecture transferability (Jeong et al.

2022). Complementarily, gradient-based formulations such as LGrad probe generator-induced gradient traces as an alternative signal for attribution and transfer (Tan et al. 2023). A related insight targets the *up-sampling* operations common to CNN generators: Neighbouring Pixel Relationships (NPR) capture the local structural artefacts these operations leave, yielding strong generalisation across both GAN and diffusion sources (Tan et al. 2024b). Nonetheless, GAN-centric fingerprints may degrade on diffusion-generated images, whose generative mechanism produces qualitatively different traces (Rombach et al. 2022; Frank et al. 2020).

High-level physical consistency cues.

Beyond low-level traces, detectors may exploit violations of image-formation physics, such as inconsistent shadows, illumination direction, reflections, and geometry-consistency errors across objects. These cues are often more interpretable, but can be brittle under heavy editing and may require scene context or auxiliary estimators (e. g., lighting/shadow predictors), motivating hybrid designs that combine physical checks with learnt representations.

6.2 Vocoder Artefacts as Implicit Fingerprints (Audio)

The fingerprint idea transfers directly to audio. Beyond features of the speech signal itself, generators leave characteristic traces: vocoders typically introduce subtle but consistent artefacts into synthesised audio, and these *vocoder artefacts* have been shown to function as implicit watermarks for identifying generated audio (Cao 2025; Sun et al. 2023). While watermarking is normally used to assert ownership or copyright (Zong et al. 2025), the same artefacts can be exploited as discriminative features that point back to the underlying generation model, an audio analogue of GAN fingerprinting; this capability is now directly benchmarked by STOPA, which systematically varies acoustic models and vocoders to support open-set source tracing and attribution (Firc et al. 2025). The analytical significance is that this collapses a distinction the rest of the survey keeps separate: vocoder traces are passively present yet behave like an actively embedded watermark, so the same signal serves both detection and attribution. This dual role is what makes audio fingerprints more powerful than their image counterpart, but it inherits the identical fragility: the trace is tied to a specific vocoder family and to a near-unprocessed signal, so it erodes as vocoders evolve and as laundering is applied.

6.3 Frame-Level and Physiological Artefacts (Video)

In video, frame-level cues capture local inconsistencies such as unnatural textures, blending boundaries, colour-statistics mismatches, and camera-pipeline inconsistencies. Many approaches treat detection as frame classification followed by temporal aggregation; these can achieve strong in-domain accuracy but often degrade under cross-dataset/cross-generator shift, motivating generalisation-oriented training and representation learning. A complementary, distinctly video-specific direction exploits *physiological* plausibility (e. g., remote photoplethysmography, rPPG) and biometric consistency over time; such cues can provide strong evidence when applicable, but

may be sensitive to illumination, motion, and video quality. A representative line of work leverages biological signals for synthetic video detection (Ciftci et al. 2020), and biologically grounded cues have recently been extended from subject physiology to viewer perception: natural videos trace straighter trajectories in perceptual representation space than generated ones, yielding a training-light geometric detector (Internò et al. 2025).

Cross-Modal Takeaway

Artefact forensics generalises the notion of a "fingerprint" across modalities (spectral/residual traces in images, vocoder traces in audio, and frame/pipeline inconsistencies in video) and uniquely supports *attribution*, not just binary detection. The transferable limitation is that fingerprints are tied to a particular generator family and to an unprocessed signal, so they erode under generator evolution and laundering. Notably, physiological cues in video illustrate the *boundary* of cross-modal transfer: some artefact signals are genuinely modality-specific and have no clean analogue elsewhere.

7 Transformer-Based Representation Learning

If frequency analysis is the clearest *passive* cross-modal technique, supervised representation learning, and the Transformer paradigm in particular, is the one that appears in *all four* modalities. Across text, image, audio, and video, the dominant recipe is to train a discriminative encoder on labelled real-vs-synthetic corpora, increasingly using attention-based architectures that capture long-range dependencies and subtle global inconsistencies. The cross-modal strength is strong in-domain accuracy; the cross-modal weakness, repeated in every modality, is poor generalisation to unseen generators, domains, or styles without carefully designed training mixtures and evaluation protocols. Figure 8 depicts both the shared recipe and this shared failure mode.

7.1 Encoder-Based Text Detectors and Domain Generalisation

Supervised text detectors train discriminative models (typically Transformer encoders) on labelled corpora of human vs machine text. These methods can achieve strong in-domain performance, but often generalise poorly when the generator, domain, or writing style changes (Guo et al. 2023a; Sadasivan et al. 2025). A common recipe fine-tunes an encoder (e. g., RoBERTa/DeBERTa-like architectures) on paired human/LLM corpora with careful control of prompts, length, and domain; an early and still widely used instance is OpenAI's RoBERTa-based GPT-2 output detector (Community 2024), and benchmarks such as HC3 facilitate systematic evaluation under controlled conditions (Guo et al. 2023a). Robustness-oriented variants such as ImBD first imitate machine stylistic preferences and then score with style-conditional probability curvature, explicitly targeting machine-revised text from the newest LLMs and evasion-oriented rewriting (Chen et al. 2025a), and RepreGuard shows that internal hidden-representation patterns of a surrogate LLM separate human from machine text more robustly than surface statistics (Chen et al. 2025c). However, these gains do not

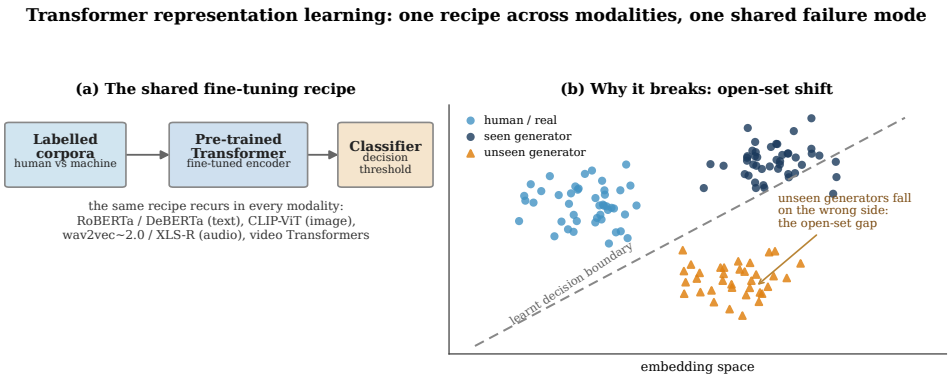


Fig. 8 Transformer-based representation learning as a cross-modal recipe. (a) A pre-trained Transformer encoder is fine-tuned on labelled human-versus-machine corpora and read out by a thresholded classifier; the same recipe recurs with RoBERTa- and DeBERTa-style encoders in text, CLIP-ViT backbones in image, wav2vec 2.0 and XLS-R front-ends in audio, and spatio-temporal Transformers in video. (b) The shared failure mode: samples from generators unseen during training fall in regions of the embedding space where the learnt decision boundary no longer separates real from synthetic, producing the open-set gap analysed in Section 10. The embedding plot is a schematic illustration.

automatically transfer: accuracy and calibration can deteriorate under topic shift, temporal drift, or unseen generators, making cross-domain testing and uncertainty-aware thresholds essential (Sadasivan et al. 2025).

Recent work explicitly targets domain and generator shift through dedicated domain-generalisation (DG) objectives. EAGLE learns generator-invariant representations via adversarial and contrastive signals, improving transfer to unseen domains and generators (Bhattacharjee et al. 2024), and training an LLM to rewrite its input, using edit distance as the detection signal, generalises across dozens of domains (Hao et al. 2025). Complementarily, MGT-Prism employs spectral alignment in the frequency domain to suppress spurious domain-specific cues and enhance cross-domain robustness (Liu et al. 2026), and disentangling generator-invariant semantics from generator-specific artefacts pushes cross-generator transfer further still (Pu et al. 2026). These approaches are highly relevant for deployment, yet still exhibit pronounced class-wise asymmetries (e. g., higher AI recall at the expense of human recall) under strong shift, underscoring that domain generalisation remains open and that robustness, calibration, and fairness must be evaluated under realistic, fixed operating-point conditions. Detection can also be framed as a representation-learning problem in which robust embeddings and calibrated scores matter as much as raw accuracy; practical systems often fuse multiple weak signals (likelihood + stylometry + classifier logits), though such fusion adds tuning complexity and can fail under severe post-editing or open-set generators (Sadasivan et al. 2025; Lu et al. 2024). A complementary direction examines whether strong generators can support detection via self-detection or model-based scoring; earlier evidence is encouraging under certain assumptions (Zellers et al. 2019), but gains are not guaranteed under open-set or adversarial regimes.

7.2 CNN, ViT, and Universal Detectors (Image)

Supervised image detectors train discriminative models on labelled real-vs-synthetic corpora. CNN classifiers learn discriminative features directly from pixel space (often aided by ImageNet pretraining); while strong in-domain, they suffer from limited cross-dataset generalisation and vulnerability to adversarial perturbations, motivating more robust training and cross-generator evaluation (Frank et al. 2020; Wesselkamp et al. 2022). Vision Transformers (ViTs) leverage global self-attention to model long-range dependencies and subtle inconsistencies missed by local convolutions (Liu et al. 2024), and recent designs increasingly adopt hybrid pipelines (e. g., spatial+frequency branches or multi-scale fusion) to improve transfer across GAN and diffusion generators (Liu et al. 2024). To reduce dependence on any single generator family, *universal* detectors combine complementary cues and train on diverse mixtures of generators and post-processing pipelines, improving open-set transfer (Alam et al. 2024; Tan et al. 2024a; Zhu et al. 2023a); Community Forensics takes this strategy to its logical extreme, training on 2.7 million images from 4 803 generators and showing that generalisation scales systematically with generator count (Park and Owens 2025). A particularly influential direction repurposes frozen vision-language features: UnivFD shows that CLIP-ViT features generalise across generator families with only a linear probe (Ojha et al. 2023), a paradigm extended by Cozzolino et al. (2024), by C2P-CLIP through category-common prompt injection (Tan et al. 2025), by CO-Spy through fusing semantic and pixel features (Cheng et al. 2025a), by few-shot adaptation to new generators (Wu et al. 2025b), and by Effort, which traces the failure of naive fine-tuning to feature-space rank collapse and preserves the pretrained subspace via orthogonal decomposition (Yan et al. 2025b); TAP demonstrates that patch-level tokens of general-purpose vision foundation models carry equally transferable forensic signal (Abdullah et al. 2026). Data-centric advances matter as much as architecture: B-Free shows that bias-free training data (self-conditioned reconstructions of real images) substantially improves generalisation and calibration (Guillaro et al. 2025), a data-alignment principle independently established with autoencoder-aligned real/fake pairs (Rajan et al. 2025), while generator-aware prototype learning (GAPL) confronts a *benefit-then-conflict* dilemma in which naively expanding source diversity eventually degrades performance (Qin et al. 2026). Self-supervised pre-training also matters: frozen self-supervised ViTs (e. g., DINO-pretrained) can outperform supervised backbones for deepfake detection under limited data (Nguyen et al. 2024).

7.3 Deep and Large-Model Audio Detectors

Audio detection has followed the same trajectory from handcrafted features (Section 5) toward end-to-end deep models, which take raw waveforms or Mel-spectrograms and generally yield improved performance because the entire feature pipeline is optimised for the detection task (Verma et al. 2025). When Mel-spectrograms are the input, CNN-based models (notably ResNet variants) are well suited to region-based patterns (Chen et al. 2020; Patel et al. 2023), though they are weaker at capturing temporal dependencies. Because raw audio is inherently sequential, RNNs and hybrid CNN–LSTM architectures jointly model sequential and spatial structure (Bhargavi

Table 6 Representative audio anti-spoofing countermeasures on the ASVspoof logical-access (LA) protocols, reported as equal error rate (EER, %; lower is better). Figures are drawn from the respective ASVspoof 2019 and 2021 LA evaluation sets (indicated per row) and are therefore not directly comparable across protocols; the trend, an order-of-magnitude EER reduction as the field moved from hand-crafted front-ends to self-supervised speech representations, parallels the encoder-fine-tuning trend in text and image detection. In-domain gains are large, but cross-dataset and codec-shifted evaluation remains the dominant open challenge.

Countermeasure	**Front-end / model**	**EER (%)**
CQCC–GMM (baseline) (Yamagishi et al. 2021)	Hand-crafted cepstral (2021 LA)	15.62
RawNet2 (Yamagishi et al. 2021)	Raw-waveform CNN (2019 LA)	0.99
AASIST (Jung et al. 2022)	Spectro-temporal graph attention (2019 LA)	0.83
wav2vec2-XLS-R + AASIST (Tak et al. 2022)	SSL front-end + graph attention (2021 LA)	0.82
XLSR-Mamba (Xiao and Das 2025)	SSL front-end + bidirectional Mamba (2021 LA)	0.93

et al. 2024; Chitale et al. 2024). Most recently, large pre-trained models (Whisper (Radford et al. 2023), wav2vec 2.0 (Baevski et al. 2020), and WaveNet (van den Oord et al. 2016)) have proven effective for deepfake audio detection, an effectiveness largely attributed to the Transformer architecture's capacity to model long-range temporal dependencies and capture subtle artefacts; using Whisper's ASR encoder as a detection front-end, for instance, measurably improves robustness (Kawa et al. 2023). A pivotal result showed that a self-supervised wav2vec 2.0 front-end, fine-tuned with data augmentation, dramatically lowers equal error rates on ASVspoof benchmarks (Tak et al. 2022), and back-end design on top of such foundation models has become a research axis of its own, with lightweight nested architectures such as Nes2Net exploiting the full high-dimensional SSL feature space (Liu et al. 2025b); the ASVspoof 2021 challenge established the spoofing/deepfake evaluation protocols these systems are measured on (Yamagishi et al. 2021; Liu et al. 2023), and the recent ASVspoof 5 challenge extends evaluation to crowdsourced speakers and adversarial attacks at scale (Wang et al. 2024b), with the Spoof-SUPERB benchmark systematically comparing twenty self-supervised front-ends for this task (Ali et al. 2026) and a complementary 2026 study reporting large cross-corpus differences across self-supervised backbones, with wav2vec 2.0 variants outperforming WavLM and Whisper (Lopez et al. 2026). Beyond standard self-attention, graph-attention methods such as AASIST (Jung et al. 2022), hybrid CNN–Transformer architectures (Gaikawad and Ghosh 2025), and large language model-based approaches such as DFALLM (Li et al. 2025b) have been explored, and state-space alternatives such as XLSR-Mamba match or exceed attention-based countermeasures at lower inference cost (Xiao and Das 2025).

Table 6 illustrates the resulting trajectory on the standard ASVspoof logical-access protocols, where equal error rate (EER) has fallen by roughly an order of magnitude as the field moved from hand-crafted front-ends to self-supervised representations, while also exposing the central failure mode: the same systems degrade sharply under cross-dataset and codec-shifted evaluation, mirroring the open-set fragility seen in the text and image modalities.

7.4 Spatio-Temporal and Video Transformers

Because video fakes often manifest through *temporal* inconsistencies (flickering, jitter, identity drift, inconsistent motion), modern detectors model spatio-temporal dynamics: 3D CNNs jointly learn appearance and short-term motion and remain a strong baseline; sequential models (e. g., CNN+RNN/TCN) capture longer-range dependencies; and Video Transformers extend attention across space and time to capture long-range inconsistencies, at the cost of higher compute and latency. Universal detectors such as UNITE extend attention beyond face-cropped inputs to background manipulations and fully synthetic text-to-video content within a single model (Kundu et al. 2025), while vulnerability-aware multi-task designs predict where a video is spatially and temporally most likely to betray forgery artefacts (Nguyen et al. 2025). This progression, from convolutional to recurrent to attention-based encoders, mirrors the audio trajectory above and the text/image encoder evolution, reinforcing that the Transformer is a genuinely cross-modal decision paradigm. Two further frontiers extend this paradigm. For video, parameter-efficient adaptation of pre-trained vision encoders (e. g., tuning only LayerNorm parameters) achieves strong cross-benchmark generalisation (Yermakov et al. 2026), as do plug-and-play spatiotemporal adapters combined with video-level blending of temporal-artefact pseudo-fakes (Yan et al. 2025c), and multimodal large language models (MLLMs) are increasingly used not only to classify but to *explain*: AIGI-Holmes generates human-verifiable explanations for image detection (Zhou et al. 2025), FakeShield jointly detects and localises forgeries (Xu et al. 2025b), SIDA extends the detection–localisation–explanation triad to social-media imagery with the 300 k-image SID-Set (Huang et al. 2025b), FakeVLM couples detection with fine-grained artefact-level explanation on the FakeClue corpus (Wen et al. 2025b), and BusterX++ targets unified cross-modal image-and-video detection (Wen et al. 2025a), while Veritas equips the MLLM with explicit reasoning patterns such as planning and self-reflection, improving transfer to in-the-wild forgeries on its accompanying HydraFake benchmark (Tan et al. 2026). These directions connect representation learning to the consistency and provenance signals of Sections 8 and 9.

Cross-Modal Takeaway

Supervised representation learning is the most broadly shared technique in the survey: the same fine-tuned-encoder recipe, the same shift toward attention-based and large pre-trained backbones, and the same open-set generalisation gap recur in every modality. Crucially, the *mitigations* also rhyme: generator-invariant DG objectives in text, universal detectors via diverse training mixtures in image, and source-balancing in audio are three modality-specific instances of one cross-modal strategy, which we revisit in Section 10.

8 Consistency and Semantic Checks

A distinct family moves beyond low-level traces to ask whether content is internally and contextually *coherent*. Generators that are locally convincing often fail to maintain

global consistency: discourse coherence in text, physically plausible illumination in images, acoustic-environment plausibility in audio, and, most directly cross-modal of all, synchronisation between lip motion and speech in video. This family is where the survey's cross-modal premise is most literal, since some cues require *two* modalities at once.

8.1 Linguistic and Discourse Consistency (Text)

Stylometric methods use linguistics- and style-based features (e. g., POS distributions, lexical diversity, discourse-coherence cues) to distinguish human and machine text. They can be lightweight and interpretable, but are sensitive to genre, topic, author style, and language proficiency, and may encode undesirable biases in real-world settings (Sadasivan et al. 2025). Consequently, stylometry is most effective as one consistency signal fused with others rather than as a standalone detector. Notably, humans who frequently use LLMs exploit exactly such lexical and discourse cues to out-detect most automatic systems, even under paraphrase attacks, suggesting the interpretable stylometric signal is richer than current feature sets capture (Russell et al. 2025).

8.2 Physical-Consistency Cues (Image and Video)

For images and video, a complementary consistency check exploits violations of image-formation physics (inconsistent shadows, illumination direction, reflections, and geometry across objects and frames. As noted in Section 6, these cues are interpretable but brittle under heavy editing and often require scene context or auxiliary estimators, motivating hybrid designs that pair physical checks with learnt representations. In video, the same idea extends temporally: physically implausible motion or lighting that drifts across frames is itself a consistency violation, a principle recently formalised by physics-driven detectors that test whether a video's spatiotemporal dynamics obey the probability-flow conservation laws real footage satisfies (Zhang et al. 2025).

8.3 Within-Modality Audio Consistency

Even within the audio modality, auxiliary high-level cues provide signal. Speaker-identity verification (Datta et al. 2025b) and emotion-related cues (Li et al. 2025a) have been explored as complementary features for detecting manipulated audio. These are essentially task-driven consistency checks whose value depends on the scenario: security applications emphasise speaker-identity consistency, whereas political-speech analysis emphasises content consistency and factual correctness. Conceptually, these are the audio analogue of the linguistic- and physical-consistency checks used in text and video: rather than scoring low-level artefacts, they ask whether high-level attributes, speaker identity, emotional trajectory, remain internally coherent. The transferable limitation is that such cues presuppose a genuine reference (an enrolled speaker, an expected affect) and so degrade toward chance in the open, reference-free setting that characterises real deployment.

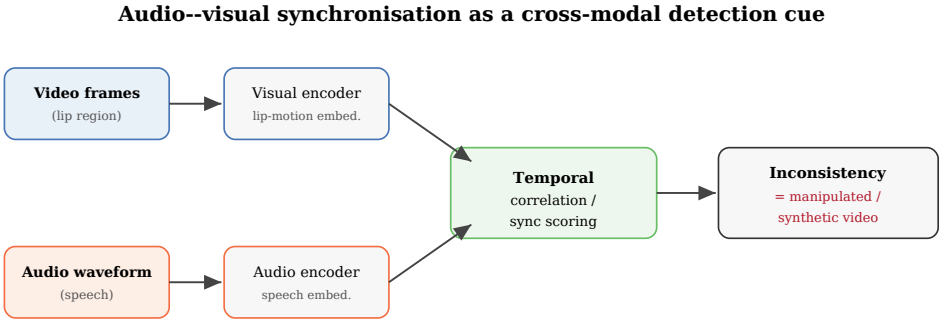


Fig. 9 Audio–visual synchronisation as a detection cue. The detector compares visual lip-motion embeddings with audio embeddings over time; inconsistencies (misalignment or atypical correspondence patterns) can indicate synthetic or manipulated video.

8.4 Audio–Visual Synchronisation (Video, Cross-Modal)

The most explicitly cross-modal consistency cue asks whether lip dynamics align with the phonetic content and timing of speech. Lip-centric detectors exploit the observation that many manipulations struggle to preserve fine-grained mouth-motion consistency over time; a representative and widely cited approach focuses on lip motion as a robust forgery cue (Haliassos et al. 2021), recent work analyses mouth inconsistencies with a vision temporal transformer specifically for lip-syncing deepfakes (Datta et al. 2025a), and X-AVDT reads speech–motion misalignment directly from the audio–visual cross-attention maps of generative talking-head models (Kim et al. 2026). Audio generation algorithms likewise struggle to achieve full cross-modal coherence, so inconsistencies between audio and visual streams have been widely exploited for deepfake detection (Gu et al. 2020; Zheng et al. 2025), though such approaches require visual information, which is absent in audio-only scenarios; reported gains can moreover ride on dataset shortcuts such as leading silence, motivating audio–visual detectors trained only on real data (Smeu et al. 2025). For diffusion-generated video specifically, learning a multi-modal forgery representation from a large multimodal model captures forgery traces beyond facial regions and across diverse semantics (Song et al. 2024). Beyond lip–speech alignment, cross-modal consistency can exploit acoustic–visual plausibility, such as whether reverberation and room characteristics match the depicted environment and motion. Figure 9 sketches the canonical audio–visual synchronisation detector.

8.5 End-to-End Multimodal Architectures (Video)

Recent systems increasingly adopt end-to-end multimodal encoders (visual backbone + audio backbone + fusion) trained for robust classification and, when available, attribution/localisation. Contrastive objectives and large-scale pretraining can improve transfer under generator shift, but robustness to re-encoding and partial edits remains a key challenge.

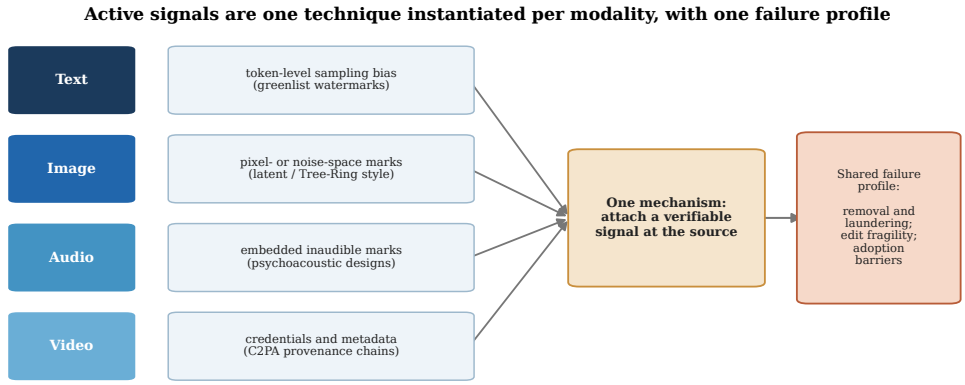


Fig. 10 Active signals are a single cross-modal technique instantiated once per modality: token-level sampling bias in text, pixel- or noise-space marks in images, embedded inaudible marks in audio, and credential chains for video. Because every instantiation attaches a verifiable signal at the source, all four inherit the same failure profile, namely removal and laundering attacks, fragility under heavy editing, and dependence on generator- or platform-side adoption.

Cross-Modal Takeaway

Consistency checks span a spectrum from *within-modality* (discourse coherence, physical plausibility, speaker identity) to *across-modality* (lip–speech synchronisation, acoustic–scene plausibility). They tend to be more interpretable and harder to fake than low-level cues, but they require richer context or auxiliary modalities and can be brittle under heavy editing. Audio–visual synchronisation is the clearest realisation of the survey's thesis that exploiting relationships *between* modalities yields signals unavailable to any single-modality detector, a direction we argue is central to future work (Section 11).

9 Active Signals: Watermarking and Provenance

Because passive detection is inherently brittle, *active* approaches embed verifiable signals at generation or capture time. This family is strikingly uniform across modalities: token-level watermarks in text, imperceptible pixel watermarks in images, embedded marks in audio, and cryptographic provenance/credentials for video all share one mechanism: attach a detectable signal at the source, and, crucially, one cross-modal *failure profile*: vulnerability to removal and laundering, degradation under heavy editing, and dependence on generator- or platform-side adoption. The shared failure profile is itself a cross-modal finding worth stating explicitly, and it has now been formalised in a systematisation of watermarking security notions and threat models spanning text, image, and audio (Zhao et al. 2025); Figure 10 summarises this one-mechanism, one-failure-profile structure.

9.1 Token-Level Watermarking (Text)

Token-level watermarking biases sampling so that outputs carry a detectable statistical signature; a representative framework is the "greenlist" watermarking approach (Kirchenbauer et al. 2023), recently extended from left-to-right decoding to any-order diffusion language models by enforcing the greenlist bias in expectation over undetermined context (Gloaguen et al. 2026). This line has reached production scale: SynthID-Text integrates watermarking with speculative sampling to preserve quality while enabling detection, validated across millions of live responses (Dathathri et al. 2024); beyond zero-bit detection, error-correction-based multi-bit schemes embed attributable identifiers with provable robustness to bounded edits (Qu et al. 2025). While watermarking can improve auditability in principle, it faces well-known challenges: degradation under paraphrasing/translation and heavy edits, removability via rewriting or re-generation, and adoption barriers because it requires generator-side cooperation (Kirchenbauer et al. 2023; Lu et al. 2024); indeed, n-gram watermarks can be systematically removed, and their green–red partitions stolen, through query-based probing (Chen et al. 2025b). These failure modes mirror those of watermarking in other modalities, where removal or laundering attacks are a persistent concern. Operationally, reliability concerns have led to rollbacks: OpenAI discontinued its AI Text Classifier after reporting limited reliability across use cases (OpenAI 2023), reinforcing that deployed detectors must be assessed for calibration, robustness, and fairness under realistic shift, not mean accuracy alone (Sadasivan et al. 2025).

9.2 Image Watermarking and Provenance

Image watermarking inserts imperceptible marks that can be detected later, enabling reliable identification when the signal survives common transformations (Fernandez et al. 2023; Atallah et al. 2001); recent work recasts watermarking as segmentation, localising multiple distinct marks within a single spliced or partially edited image (Sander et al. 2025). Diffusion-specific schemes embed the mark in the initial noise vector: Tree-Ring watermarks structure the signal in Fourier space for invariance to crops, rotations, and flips (Wen et al. 2023), RingID refines this for robust multi-key identification (Ci et al. 2024), and pseudorandom error-correcting codes push the approach to cryptographic undetectability, making watermarked and unwatermarked outputs provably indistinguishable without the key (Gunn et al. 2025). Such semantic watermarks are not unassailable: they can be forged onto arbitrary images, or stripped, in a purely black-box setting using an unrelated diffusion model (Müller et al. 2025). The robustness of such schemes is itself now systematically benchmarked: WAVES stress-tests image watermarks against distortion, regeneration, and adversarial attacks (An et al. 2024), and dedicated attacks demonstrate that watermark-based detection can be evaded (Jiang et al. 2023); UnMarker shows that a single scheme-agnostic spectral attack removes every tested defensive image watermark without query access (Kassis and Hengartner 2025). However, active approaches require generator-side adoption and remain vulnerable to removal, laundering, or

incompatibility across platforms (Fernandez et al. 2023). As with text, image provenance shifts the problem from *inference* to *verification*, reliable only when the signal is preserved end-to-end.

9.3 Audio Watermarking

In audio, embedded watermarks have been proposed for deepfake-speech detection (e. g., AudioMarkNet (Zong et al. 2025)), and cross-attention designs with psychoacoustic-aligned losses now deliver simultaneous detection and source attribution that survive aggressive audio editing (Liu et al. 2025c). As discussed in Section 6, vocoder artefacts can also act as *implicit* watermarks (Sun et al. 2023), blurring the line between passive forensics and active provenance in this modality. The same adoption and removal constraints apply.

9.4 Provenance and Credentials (Video)

For video, generation/editing pipelines can attach verifiable information (cryptographic signatures, metadata binding, or robust watermark-like signals) enabling downstream verification when the chain is preserved. Standardisation efforts such as C2PA aim to provide interoperable provenance and content credentials across tools and platforms (Coalition for Content Provenance and Authenticity (C2PA) 2024), an approach increasingly advocated for safeguarding trust in the digital news ecosystem (European Broadcasting Union 2024), and open neural video watermarking frameworks such as Meta's Video Seal embed marks robust to codecs and re-encoding (Fernandez et al. 2024), while in-generation schemes such as VideoShield embed the mark directly in the diffusion noise of the video generator itself and additionally localise spatial and temporal tampering (Hu et al. 2025b). The robustness of video watermarks is now benchmarked directly by VideoMarkBench (Jiang et al. 2025), and industry deployments increasingly layer multiple signals, combining watermarking with C2PA metadata and public verification, as in recent provenance tooling (OpenAI 2026).

Cross-Modal Takeaway

Active signals offer the highest reliability *when present and preserved*, and they generalise cleanly across modalities at the level of both mechanism and limitation. The recurring obstacles (removal/laundering, edit fragility, and adoption incentives) are the same everywhere, which is precisely why the field increasingly favours *hybrid* stacks: use active provenance when available and fall back to passive forensic evidence otherwise (Section 10).

10 Cross-Modal Synthesis and Open Challenges

Having surveyed detection technique by technique, we now read the cross-modal picture back from the taxonomy of Section 2 and Table 3. Four challenges recur in *every* modality, with modality-specific surfaces but a shared mechanism: the laboratory-to-deployment gap, open-set and cross-domain generalisation, adversarial robustness, and

Table 7 Cross-modal transferability of the six detection principles surveyed in Sections 4–9. For each technique family, the table states the principle shared across modalities, the modalities in which the survey documents substantive instantiations, and the dominant failure mode that limits transfer.

Technique family	Shared principle	Transfers to	Where it breaks
Statistical / model-based	Typicality under a reference generator (likelihood, curvature, reconstruction error)	Text, image, audio; emerging in video	Signal attenuates as generators approach natural data; fragile to post-editing
Frequency-domain	Generators disturb the spectral statistics of natural content	Image, audio, video	Re-encoding and laundering erase the spectral peaks; no text analogue
Artefact / fingerprint	Generator-specific residual traces support detection and attribution	Image, audio, video	Erodes across generator families and under laundering; no text analogue
Transformer representation	Fine-tuned pre-trained encoders separate real from synthetic	Text, image, audio, video	Open-set collapse on unseen generators, domains, and post-processing
Consistency / semantic	High-level coherence is harder to fake than low-level statistics	All four; strongest audio–visual	Requires context or a reference; brittle under partial, localised edits
Active watermark / provenance	A verifiable signal attached at the source	Text, image, audio, video	Removal and laundering; fragile to heavy edits; requires generator-side adoption

fairness. Treating these as one cross-modal phenomenon each, rather than four parallel discussions, is the central synthesis of this survey. As an entry point, Figure 11 distils the strengths and weaknesses of the six technique families into a common qualitative profile. Table 7 states the same synthesis in propositional form: for each family, the shared principle, the modalities it transfers to, and the point at which modality-specific assumptions break.

10.1 The Laboratory-to-Deployment Gap

The single most consistent finding across modalities is that strong laboratory performance does not survive deployment. While controlled accuracy can approach 99%, realistic conditions (domain shift, re-encoding, compression, paraphrasing, and lightweight human edits) routinely degrade performance (Jisc 2025). This pattern holds for likelihood-based text detectors, spectral image detectors, vocoder-cue audio detectors, and frame-level video detectors alike, as Figure 12 makes visually explicit. Table 8 compiles reported in-domain and cross-domain results study by study; rather than a single dramatic figure, we report the within-study degradation and flag where metrics are not directly comparable. The cross-modal regularity is qualitative but robust: image detectors that reach 99.9% in-domain fall to a 72.1% cross-generator average, and below 55% on the most dissimilar generators (Zhu et al. 2023b), a gap AIGIBench confirms systematically for state-of-the-art detectors under social-media-style degradations (Li et al. 2025d); text detectors advertised at $\geq$99%

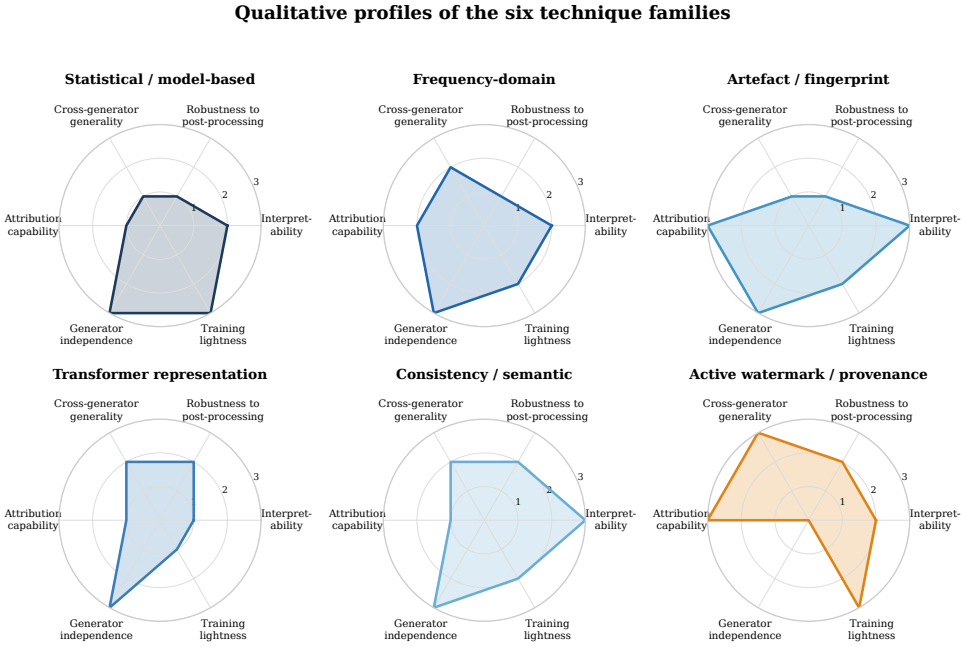


Fig. 11 Qualitative profiles of the six technique families of Sections 4–9, rated 0–3 on six criteria distilled from the per-family cross-modal takeaways: interpretability, robustness to post-processing, cross-generator generality, attribution capability, generator independence (no generator-side cooperation required), and training lightness. The passive families trade robustness for independence in different ways, whereas active watermarking and provenance invert the trade-off: strong on robustness, generality, and attribution, but the only family that requires generator-side adoption. The ratings are a qualitative synthesis of the surveyed literature, not measured scores.

lose 36–41 points under adversarial perturbations such as homoglyphs and synonym swaps (Dugan et al. 2024); and audio detectors with sub-1% in-domain EER degrade by an order of magnitude once evaluation leaves studio-quality, in-distribution conditions, a collapse quantified systematically by VoiceWukong, where detection performance degrades sharply across 38 everyday manipulations for both academic and commercial detectors (Yan et al. 2025d). Self-supervised speech front-ends can reach sub-1% EER on ASVspoof 2021 LA when fine-tuned with data augmentation (Tak et al. 2022), yet without such augmentation the same class of features degrades from 1.47% EER on the 2019 LA benchmark to 10.97% on the codec-shifted 2021 LA evaluation (Wang and Yamagishi 2022); on ASVspoof 5, the AASIST baseline that was near-perfect on earlier studio corpora records a 29.1% EER, and even the best challenge systems reach below 5% EER only in the open condition that permits external self-supervised pre-training (Wang et al. 2024b).

10.2 Open-Set and Cross-Domain Generalisation

Open-set collapse, that is, failure on generators or domains unseen in training, is arguably the strongest genuinely cross-modal failure mode. Supervised text detectors

Table 8 The laboratory-to-deployment gap across modalities. Each row reports a representative detector's *in-domain* performance and its *cross-domain / open-set* performance from the same study (or a directly comparable follow-up evaluation), with the resulting degradation. The upper block lists standard detectors, for which the gap is large; the lower block lists recent *robustness-oriented* methods, which are explicitly designed to narrow it. The final row contrasts the open and closed training conditions of the same challenge rather than an in-domain/cross-domain pair. Metrics differ across modalities (accuracy or AUC for text, image, and video; equal error rate, EER, for audio, where *lower* is better), so figures are not directly comparable *across* rows; the consistent signal is the *direction and magnitude* of within-study degradation.

Study / detector	Modality	In-domain	Cross-domain / open-set	Degradation
GenImage, ResNet-50 (Zhu et al. 2023b)	Image	99.9% acc. (train/test SD v1.4)	72.1% avg. acc. over 8 test generators (train on SD v1.4); as low as 52.0% (BigGAN), 53.5% (ADM)	∼28 pts avg.; up to ∼48 pts
UnivFD (CLIP linear) (Ojha et al. 2023)	Image	100.0% acc. on ProGAN (in-domain)	81.4% mean acc. over unseen GAN, diffusion, and autoregressive generators; ∼82% on unseen diffusion families	∼19 pts under cross-generator shift
RAID, metric/neural detectors (Dugan et al. 2024)	Text	Vendor-reported ≥99% acc.	Five detectors lose 40.6% acc. on average under a homoglyph attack; Binoculars −36.1% under synonym swaps	36–41 pts under adversarial / unseen shift
W2V-XLSR front-end (Wang and Yamagishi 2022)	Audio	1.47% EER (ASVspoof 2019 LA)	10.97% EER on the codec-shifted ASVspoof 2021 LA evaluation with the same front-end	∼7× relative EER increase
AASIST baseline (Wang et al. 2024b)	Audio	0.83% EER on ASVspoof 2019 LA (Jung et al. 2022)	29.1% EER on ASVspoof 5 (crowdsourced speech, codec compression, adversarial attacks)	>1 order of magnitude (EER)
Xception, FF++→Celeb-DF (Li et al. 2020)	Video	99.7% AUC (FF++, in-domain)	48–65% AUC on Celeb-DF (unseen manipulations)	∼35–50 pts AUC
DeBERTa-v3 + FeatAttn (Mady et al. 2026b)	Text	99.5% bal. acc. (HC3 PLUS, in-domain)	85.9% bal. acc. on the multi-domain, multi-generator M4 benchmark under a fixed decision threshold	∼14 pts under cross-dataset shift
TDT (Temporal Tomography) (West et al. 2025)	Text	0.855 overall AUROC on RAID across seen and unseen generators, 6.9% above the best baseline	+14.1% AUROC over the best baseline under paraphrasing attacks (HART Level 2)	Retains its advantage under adversarial shift
GAPL (Qin et al. 2026)	Image	98.2% acc. on GenImage SD v1.4	90.4% mean acc. across six GAN and diffusion benchmarks (55 test subsets)	Mitigates "benefit-then-conflict" as source diversity grows
Top ASVspoof 5 systems (Wang et al. 2024b)	Audio	2.59% EER for the best open-condition system (external SSL pre-training permitted)	8.61% EER for the best closed-condition system, without external pre-training	>3× EER without SSL pre-training

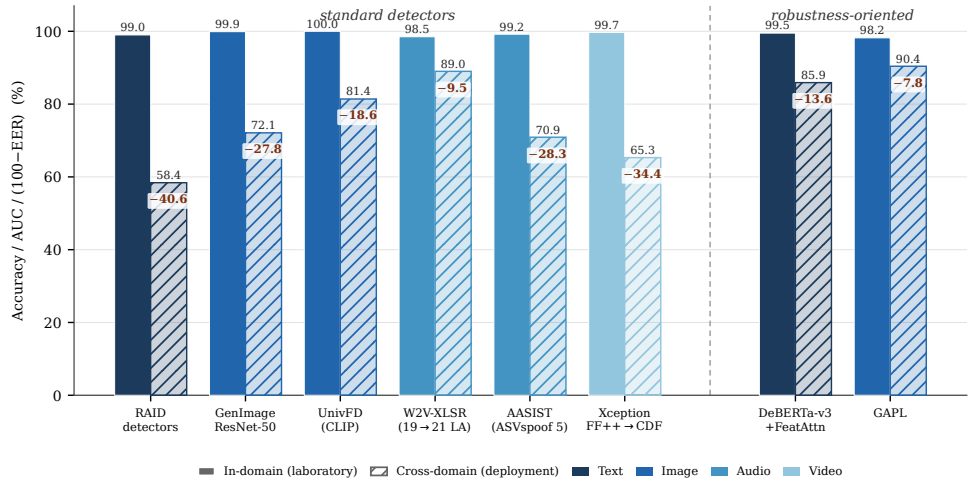


Fig. 12 The laboratory-to-deployment gap is a cross-modal regularity. For each of eight representative studies (colour-coded by modality), the solid bar gives in-domain performance and the hatched bar gives cross-domain / open-set performance, with the absolute drop in points shown on the bar. The six *standard detectors* (left) lose 10–41 points under realistic shift (unseen generators, adversarial edits, codec change, or unseen manipulations), whereas the two recent *robustness-oriented* methods (right) are explicitly designed to narrow the gap and show visibly smaller drops. Values are the verified figures of Table 8; audio is shown as 100 − EER for visual comparability, and metrics are not directly comparable across studies.

degrade under unseen generators and topic/temporal drift (Guo et al. 2023a; Sadasivan et al. 2025), an effect that is especially visible under a deployment-realistic protocol in which a single decision threshold is calibrated once and held fixed across all test distributions (Dugan et al. 2024; Mady et al. 2026b); image detectors that excel in-domain collapse under cross-generator and cross-dataset shift (Frank et al. 2020; Zhu et al. 2023a); audio detectors overfit to specific datasets given diverse generator architectures and watermarking strategies (Gu et al. 2025); and video detectors face the same problem under platform processing and partial edits. Strikingly, the *mitigations* are modality-specific instances of one strategy, namely reducing dependence on any single generator: generator-invariant DG objectives in text (EAGLE (Bhattacharjee et al. 2024), MGT-Prism (Liu et al. 2026)) and attention-based fusion of linguistic features that improves cross-dataset transfer (Mady et al. 2026b), universal detectors trained on diverse generator mixtures in image (Alam et al. 2024; Zhu et al. 2023a), and source-balancing across generators in audio (Li et al. 2024c). This convergence is exactly the kind of transferable insight the cross-modal lens is meant to surface. Crucially, this open-set fragility is increasingly measured by deployment-oriented *benchmarks* rather than curated test sets: RAID and DetectRL stress-test text detectors under domain, generator, and adversarial shift (Dugan et al. 2024; Wu et al. 2024); GenImage provides a million-scale image benchmark spanning GAN and diffusion sources (Zhu et al. 2023b); LOKI evaluates large multimodal models as cross-modal detectors (Ye et al. 2025), Forensics-Bench extends this probing to 112 forgery types, exposing systematic blind spots of large vision-language models (Wang et al. 2025), and TriDF

shows that hallucinated evidence, rather than perception alone, is the failure mode limiting such detectors across image, video, and audio (Jiang-Lin et al. 2026); and unified image-and-video benchmarks such as Ivy-Fake and the multimodal video–audio dataset MVAD push evaluation toward genuinely cross-modal settings (Jiang et al. 2026; Hu et al. 2025a), with Omni-Fake unifying image, audio, video, and talking-head detection at million-sample scale under an explicit out-of-distribution split (Li et al. 2026a). The video and audio–visual side rests on a lineage of face-manipulation datasets, including FaceForensics++ (Rössler et al. 2019), the large-scale DeepFake Detection Challenge (DFDC) corpus (Dolhansky et al. 2020), FakeAVCeleb's audio–visual deepfakes (Khalid et al. 2021), and the recent Celeb-DF++ benchmark designed explicitly to probe generalisation (Li et al. 2025c), alongside emerging large-scale AI-generated-video benchmarks such as GenVidBench (Ni et al. 2026) and the in-the-wild Deepfake-Eval-2024 collection (Chandra et al. 2026). Reliable comparison across all of these also demands sound evaluation methodology: threshold-free metrics built on the ROC curve (Fawcett 2006) and attention to detector *calibration*, since modern neural networks are frequently overconfident (Guo et al. 2017).

10.3 Adversarial Robustness

As detection increasingly relies on generator-specific artefacts, successive generations of generators suppress those cues, producing an adversarial arms race analogous to GAN training dynamics (Goodfellow et al. 2014). The pattern is cross-modal: paraphrasing and rewriting loops evade text detectors (Lu et al. 2024; Sadasivan et al. 2025), and detector-guided adversarial paraphrasing now humanises AI text in a single universal, quality-preserving pass (Cheng et al. 2025b); adaptive perturbations and spectral "washing" evade image detectors (Wesselkamp et al. 2022), though higher-order frequency statistics that such attacks leave largely untouched can restore much of the lost robustness without adversarial training (Shahjahan et al. 2026); adversarial attacks against audio detectors have emerged (Rabhi et al. 2024), and the arms race now extends to *proactive* defences, with purification attacks such as De-AntiFake stripping the protective perturbations intended to prevent voice cloning at the source (Fan et al. 2025); and re-encoding, partial manipulation, and feature-targeting attacks threaten video detectors. The shared lesson is that robustness stress-tests must be treated as a first-class part of evaluation rather than an optional add-on; community-scale efforts such as the NTIRE 2026 challenge on robust in-the-wild AI-generated image detection (Gushchin et al. 2026) mark a welcome institutionalisation of exactly this practice.

10.4 Fairness and Bias Across Modalities

Fairness failures, too, are best understood as a single phenomenon surfacing differently per modality, with a shared root in skewed training distributions. In text, Liang et al. report that seven widely used GPT detectors misclassify non-native English TOEFL essays as AI-generated with an average false-positive rate of 61.22%, while achieving near-perfect accuracy on native-speaker essays; rewriting prompts that increase linguistic diversity reduce this rate to 11.77%, and conversely, prompting native essays

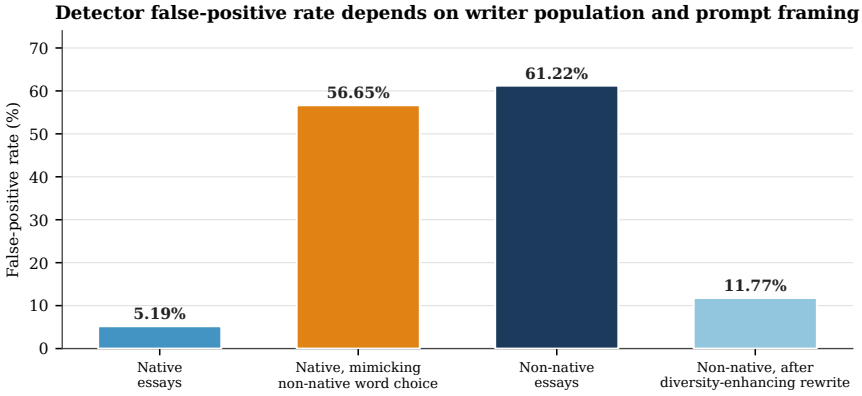


Fig. 13 False-positive rates of seven widely used GPT detectors on TOEFL and US student essays, averaged over detectors, as reported by Liang et al. (2023). Non-native essays are flagged as AI-generated at 61.22%, against 5.19% for native essays; prompting native writers to mimic non-native word choice raises their false-positive rate to 56.65%, while diversity-enhancing rewriting of non-native essays lowers it to 11.77%. The disparity is therefore driven by measurable properties of linguistic diversity, not by authorship, which is why fairness-aware evaluation must be a first-class metric.

to mimic non-native word choice increases false positives from 5.19% to 56.65% (Liang et al. 2023). A 2026 audit of sixteen deployed detectors on real student essays extends this finding beyond native-speaker status, documenting elevated false-positive rates for English-language learners and economically disadvantaged writers (Stowe et al. 2026). Figure 13 makes the magnitude of the disparity concrete.

The audio analogue is accent and dialect variation, which can disadvantage speakers under-represented in training data. The common cause, namely English-centric and demographically skewed corpora, implies a common remedy: fairness-aware evaluation across languages, proficiency levels, accents, and demographic groups should be a first-class metric in every modality, not a text-specific afterthought.

10.5 Toward Hybrid Verification Stacks

The synthesis points consistently to one architectural recommendation. Because passive methods are brittle and active methods are adoption-limited, the most promising direction is a *hybrid verification stack*: use active provenance when available (for high-confidence, traceable decisions), and fall back to passive forensic evidence otherwise, with explicit uncertainty handling and abstention for high-stakes settings. This recommendation is itself cross-modal, holding equally for text, image, audio, and video, which is why we state it once here rather than repeating it per modality. Figure 14 sketches the resulting architecture.

11 Future Directions

Rather than enumerating open topics, we distil the evidence of Sections 4–10 into six concrete research directions. Each is stated in the same form: what the surveyed

Fig. 14 The hybrid verification stack recommended in this survey. Incoming media are first checked for active signals (content credentials and watermark verification); a verified signal yields high-confidence provenance. When the signal is absent, passive forensic evidence (statistical scoring, frequency analysis, artefact and fingerprint cues, and learnt representations) is consulted, backed by consistency checks (linguistic, physical, and audio–visual) for ambiguous cases; all passive evidence is combined by calibrated, uncertainty-aware fusion, and high-stakes decisions may abstain rather than guess.

literature establishes, where current systems fail, and what future systems therefore need.

Open-set and generator-invariant detection.

The literature establishes that generalisation improves with generator diversity in training, with invariance-oriented objectives, and with fine-tuning schemes that preserve pre-trained structure (Sections 7 and 10). What fails is the transfer of these gains to genuinely unseen conditions: fixed decision thresholds collapse under distribution shift, and curated benchmarks overstate robustness relative to human-indistinguishable in-the-wild content. Future systems therefore need training mixtures that scale with the generator population, invariance objectives evaluated under frozen thresholds, and benchmark suites whose test generators are disjoint from training by construction.

Cross-modal consistency as a first-class signal.

Audio–visual synchronisation and physical-plausibility checks are among the most robust cues surveyed, precisely because generators optimise each modality separately (Section 8). What fails is their integration: most systems still fuse modalities by concatenation or generic cross-attention, dataset shortcuts can masquerade as consistency signals, and reference-free deployment blunts identity- and context-based checks. Future systems need consistency modelling as a primary training objective, spanning temporal, semantic, and physical coherence, together with nonverbal cues

such as prosody, gesture, and gaze that are harder to control than pixel-level appearance; explanation-oriented multimodal models should be systematically assessed and extended for this purpose rather than used off the shelf (Chen et al. 2025d).

Hybrid provenance–forensics systems.

Active signals now work at production scale, and cryptographically undetectable watermarks exist in principle (Section 9). What fails is reliance on any single scheme: removal, forgery, and purification attacks defeat individual watermarks, and adoption remains voluntary and uneven. Future systems need the layered stack of Figure 14, in which verified provenance yields high-confidence decisions and passive forensic evidence provides a calibrated fallback, together with interoperability standards and incentives that make source-side participation the default.

Adversarially adaptive evaluation.

Every passive family surveyed has a demonstrated adaptive attack, from detector-guided paraphrasing in text to spectral laundering in images and purification of protective perturbations in audio (Section 10). What fails is the evaluation cadence: static benchmarks age quickly, and robustness results are rarely reported against attacks published after the detector. Future evaluation needs standing, versioned, community red-team challenges in every modality, with attack-aware reporting treated as a first-class result rather than an appendix.

Uncertainty, abstention, and deployment-aware calibration.

The deployment evidence of Section 10 shows that modern detectors are frequently overconfident, that threshold-free metrics flatter systems whose operating points drift, and that reliability concerns have already forced deployed detectors to be withdrawn. What fails is the decision layer rather than the representation: scores are consumed as if calibrated when they are not. Future systems need abstention as a designed outcome for high-stakes decisions, calibration and fixed-operating-point reporting as standard evaluation practice, and lightweight architectures whose latency, privacy, and energy budgets fit real-time and edge deployment; they must also serve detection's growing infrastructural role, where calibrated, auditable scores filter training corpora and have been shown to prevent language-model collapse (Drayson et al. 2025).

Fair and multilingual detection.

Most detection research remains tied to English, and to standard, edited English in particular. This raises two related open problems. The first is intra-language variation: dialect, sociolect, slang, and, in speech, accent. Because detectors are trained predominantly on standard varieties, non-standard language tends to fall outside the learned "human" distribution, which inflates false positives for exactly the speakers and writers who are already under-represented. This is visible in text, where widely used detectors misclassify non-native English essays as machine-generated at a much higher rate than native essays (Liang et al. 2023), and the same mechanism is expected in audio, where accented or dialectal speech under-represented in anti-spoofing corpora can be scored less reliably. Slang and rapidly evolving informal registers compound

the problem, since they shift faster than training sets are refreshed. The second problem is cross-lingual transfer. Multilingual benchmarks such as M4 (Wang et al. 2024c), MULTITuDE (Macko et al. 2023), and the recent DetectRL-X (Wu et al. 2026b) show that detectors trained on one language transfer only partially to others, with accuracy dropping most for languages typologically distant from the training set. Several factors plausibly make some languages harder than others: tokeniser coverage (languages with poor sub-word segmentation yield noisier likelihood signals), the volume of pre-training and detector-training data available per language (low-resource languages offer weaker human baselines), morphological richness (highly inflected or agglutinative languages produce sparser n-gram statistics), and script (non-Latin or mixed-script text can degrade features calibrated on English). A practical consequence is that zero-shot likelihood-based methods, which depend on a well-calibrated reference model, tend to degrade faster off-English than fine-tuned classifiers trained directly on the target language. We therefore see genuinely multilingual and variety-aware evaluation, rather than English-only accuracy, as a prerequisite for claims of general detection capability, and as a concrete direction in which cross-modal insights (for example, treating accent in audio and dialect in text as the same distribution-shift problem) can transfer.

12 Conclusion

Generative AI has rapidly shifted synthetic media from a niche concern to a mainstream security and trust challenge across text, images, audio, and video. In this survey, we synthesised recent progress (2022–2026) in AI-generated content detection through a unified, cross-modal lens. Rather than treating each modality in isolation, we organised the literature *by detection technique*: statistical and model-based scoring, frequency-domain analysis, artefact and fingerprint forensics, transformer-based representation learning, consistency and semantic checks, and active watermarking and provenance, and showed, for each, where the technique transfers across modalities and where it breaks. A four-axis taxonomy connects *passive* forensic cues, *active* provenance signals, and *hybrid* pipelines that combine both.

Across modalities, three broad methodological paradigms recur. Passive detection leverages content-only traces such as likelihood/curvature statistics in text, spectral and residual artefacts in images, vocoder- and channel-related cues in audio, and spatio-temporal inconsistencies in video; these offer broad backward compatibility but are frequently brittle under distribution shift, post-processing, and partial edits (Sadasivan et al. 2025; Wesselkamp et al. 2022). Supervised discriminative detectors (Transformer-based encoders, CNN/ViT image models, and spatio-temporal video networks) reach strong in-domain performance, yet often struggle to generalise to unseen generators or domains without carefully designed evaluation protocols and diverse training mixtures (Guo et al. 2023a; Zhu et al. 2023a; Zhou et al. 2024). Active approaches including watermarking and content provenance/credentials promise higher reliability when signals survive platform transformations and are widely adopted, but remain constrained by deployment incentives, interoperability, and removal/spoofing risks (Kirchenbauer et al. 2023; Fernandez et al. 2023).

A central conclusion is that the field's dominant bottleneck is no longer purely model capacity, but **reliability under realistic conditions**. Re-encoding, resizing, cropping, re-generation, and lightweight human edits can substantially degrade detector performance and calibration, widening the laboratory-to-deployment gap. Robustness and uncertainty-awareness must therefore be treated as first-class design objectives. In practice, the most promising direction is a hybrid verification stack: use active provenance when available, and fall back to passive forensic evidence otherwise, with explicit handling of uncertainty and abstention for high-stakes settings.

Looking forward, we expect progress to be driven by: (i) evaluation regimes that reflect open-set generators, cross-domain drift, and platform post-processing; (ii) multimodal consistency modelling that couples temporal, semantic, and cross-modal signals rather than treating modalities independently; (iii) adversarially informed training and testing that anticipates adaptive evasion; and (iv) fairness-aware assessment that reduces systematic disparities across languages, writing styles, and accents (Sadasivan et al. 2025; Zhou et al. 2024). Ultimately, sustaining trust in digital media will require aligning technical detection advances with deployment realities, provenance adoption, and transparent operational practices. We hope this survey helps consolidate the current landscape, clarifies robust research directions, and supports both researchers and practitioners in the continuing arms race between generation and detection.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing interests. The authors declare no competing interests.

Data availability. No new data were generated or analysed in support of this review. The datasets discussed are third-party resources cited in the text; two corpora released by the authors are available at the repositories cited in Section 3.

Code availability. Not applicable; no new software was developed for this review.

Ethics approval and consent to participate. Not applicable; this article reviews previously published literature and involves no studies with human participants or animals performed by any of the authors.

Consent for publication. Not applicable.

References

Abdullah A, Ebert N, Wasenmüller O (2026) TAP into the patch tokens: Leveraging vision foundation model features for AI-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 379–389

Afchar D, Nozick V, Yamagishi J, et al (2018) MesoNet: A compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp 1–7, https://doi.org/10.1109/WIFS.2018.8630761

Ajder H, Patrini G, Cavalli F, et al (2019) The state of deepfakes: Landscape, threats, and impact. In: Deeptrace Report, URL https://regmedia.co.uk/2019/10/07/deepfake_report.pdf

Alam I, Muneer MS, Woo SS (2024) UGAD: Universal generative AI detector utilizing frequency fingerprints. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), pp 4332–4340, https://doi.org/10.1145/3627673.3680085

Ali H, Adupa NS, Subramani S, et al (2026) A SUPERB-style benchmark of self-supervised speech models for audio deepfake detection. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 16512–16516, https://doi.org/10.1109/ICASSP55912.2026.11460421

Almutairi Z, Elgibreen H (2022) A review of modern audio deepfake detection methods: Challenges and future directions. Algorithms 15(5):155. https://doi.org/10.3390/a15050155

An B, Ding M, Rabbani T, et al (2024) WAVES: Benchmarking the robustness of image watermarks. In: Proceedings of the 41st International Conference on Machine Learning (ICML)

Atallah MJ, Raskin V, Crogan M, et al (2001) Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In: Information Hiding: 4th International Workshop (IH 2001), Lecture Notes in Computer Science, vol 2137. Springer, pp 185–200

Baevski A, Zhou Y, Mohamed A, et al (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems

Bammey Q (2024) Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing 5:1–9

Bao G, Zhao Y, Teng Z, et al (2024) Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR)

Ben Hayun O, Betser R, Levi MY, et al (2026) Training-free detection of generated videos via spatial-temporal likelihoods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Bhargavi KD, Deepa V, Harshitha I (2024) Audio classification through mfcc features using rnn algorithm. In: 2024 4th Asian Conference on Innovation in Technology (ASIANCON), IEEE, pp 1–5

Bhattacharjee A, Moraffah R, Garland J, et al (2024) Eagle: A domain generalization framework for AI-generated text detection. arXiv:2403.15690

Bloomberg Intelligence (2023) Generative AI to become a $1.3 trillion market by 2032, research finds. Bloomberg L.P. press release, URL https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

Cable News Network (CNN) (2024) Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. CNN Business, URL https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

Cai Z, Kuckreja K, Ghosh S, et al (2025) AV-Deepfake1M++: A large-scale audio-visual deepfake benchmark with real-world perturbations. In: Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM). ACM, pp 13686–13691, https://doi.org/10.1145/3746027.3761979

Cao J, Li Q, Zhang Z, et al (2025a) Secure and robust watermarking for AI-generated images: A comprehensive survey. arXiv preprint arXiv:2510.02384

Cao L (2025) Watermarking for ai content detection: A review on text, visual, and audio modalities. arXiv preprint arXiv:2504.03765

Cao Y, Li S, Liu Y, et al (2025b) A survey of AI-generated content (AIGC). ACM Computing Surveys 57(5):1–38. https://doi.org/10.1145/3704262

Chandra NA, Lee H, Murtfeldt R, et al (2026) Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp 10668–10678

Chen B, Zeng J, Yang J, et al (2024) DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Proceedings of the 41st International Conference on Machine Learning (ICML), pp 7621–7639

Chen J, Zhu X, Liu T, et al (2025a) Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 39. AAAI Press, pp 23559–23567, https://doi.org/10.1609/aaai.v39i22.34525

Chen R, Wu Y, Guo J, et al (2025b) De-mark: Watermark removal in large language models. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). PMLR

Chen T, Kumar A, Nagarsheth P, et al (2020) Generalization of audio deepfake detection. In: Odyssey, pp 132–137

Chen X, Wu J, Yang S, et al (2025c) RepreGuard: Detecting LLM-generated text by revealing hidden representation patterns. Transactions of the Association for Computational Linguistics 13:1812–1831. https://doi.org/10.1162/tacl.a.61

Chen Y, Yan Z, Cheng G, et al (2025d) X^2-DFD: A framework for explainable and extendable deepfake detection. In: Advances in Neural Information Processing Systems (NeurIPS)

Cheng S, Lyu L, Wang Z, et al (2025a) CO-Spy: Combining semantic and pixel features to detect synthetic images by AI. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13455–13465

Cheng Y, Sadasivan VS, Saberi M, et al (2025b) Adversarial paraphrasing: A universal attack for humanizing AI-generated text. In: Advances in Neural Information Processing Systems (NeurIPS)

Chitale M, Dhawale A, Dubey M, et al (2024) A hybrid cnn-lstm approach for deepfake audio detection. In: 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), IEEE, pp 1–6

Chu B, Xu X, Wang X, et al (2025) FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12830–12839

Ci H, Yang P, Song Y, et al (2024) RingID: Rethinking tree-ring watermarking for enhanced multi-key identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 338–354

Ciftci UA, Demir I, Yin L (2020) How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), https://doi.org/10.48550/arXiv.2008.11363, URL https://arxiv.org/abs/2008.11363

Coalition for Content Provenance and Authenticity (C2PA) (2024) C2pa specification. https://c2pa.org/specifications/

Community O (2024) Roberta base openai detector

Corvi R, Cozzolino D, Zingarini G, et al (2023) On the detection of synthetic images generated by diffusion models. In: ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1–5

Corvi R, Cozzolino D, Prashnani E, et al (2025) Seeing what matters: Generalizable AI-generated video detection with forensic-oriented augmentation. In: Advances in Neural Information Processing Systems (NeurIPS)

Cozzolino D, Poggi G, Corvi R, et al (2024) Raising the bar of AI-generated image detection with CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 4356–4366

Croitoru FA, Hondru V, Ionescu RT, et al (2023) Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(9):10850–10869

Dathathri S, See A, Ghaisas S, et al (2024) Scalable watermarking for identifying large language model outputs. Nature 634(8035):818–823. https://doi.org/10.1038/s41586-024-08025-4

Datta SK, Jia S, Lyu S (2025a) Detecting lip-syncing deepfakes: Vision temporal transformer for analyzing mouth inconsistencies. arXiv preprint arXiv:2504.01470

Datta SK, Ranga T, Sun C, et al (2025b) Pia: Deepfake detection using phoneme-temporal and identity-dynamic analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1596–1606

Deloitte Center for Financial Services (2024) Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. Tech. rep., Deloitte Insights, URL https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html

Dolhansky B, Bitton J, Pflaum B, et al (2020) The deepfake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397

Drayson G, Yilmaz E, Lampos V (2025) Machine-generated text detection prevents language model collapse. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 29657–29673, https://doi.org/10.18653/v1/2025.emnlp-main.1506

Dugan L, Hwang A, Trhlik F, et al (2024) RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)

Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7890–7899

Dzanic T, Shah K, Witherden F (2020) Fourier spectrum discrepancies in deep network generated images. In: Advances in Neural Information Processing Systems (NeurIPS), pp 3022–3032

Epstein DC, Jain I, Wang O, et al (2023) Online detection of AI-generated images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, pp 382–392, https://doi.org/10.1109/ICCVW60793.

2023.00045

European Broadcasting Union (2024) Content provenance and disinformation: Ensuring trust in the digital news ecosystem. European Broadcasting Union (EBU), Technical Report, URL https://www.ebu.ch/publications/content-provenance-and-disinformation, accessed: 2026-02-01

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024, transparency obligations for synthetic content (Article 50) apply from 2 August 2026

Fan W, Chen K, Liu C, et al (2025) De-AntiFake: Rethinking the protective perturbations against voice cloning attacks. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), vol 267. PMLR, pp 15700–15718

Fariello S, Fenza G, Forte F, et al (2025) Distinguishing human from machine: A review of advances and challenges in AI-generated text detection. International Journal of Interactive Multimedia and Artificial Intelligence 9(3):6–18. https://doi.org/10.9781/ijimai.2024.12.002

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feng H, Chen W, Chen J, et al (2026) CoCoVideo: The high-quality commercial-model-based contrastive benchmark for AI-generated video detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 11346–11356

Fernandez P, Couairon G, Jégou H, et al (2023) The stable signature: Rooting watermarks in latent diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 22466–22477

Fernandez P, Elsahar H, Yalniz IZ, et al (2024) Video seal: Open and efficient video watermarking. arXiv preprint arXiv:2412.09492

Firc A, Chhibber M, Mishra J, et al (2025) STOPA: A dataset of systematic variation of deepfake audio for open-set source tracing and attribution. In: Proceedings of Interspeech 2025. ISCA, pp 1553–1557, https://doi.org/10.21437/Interspeech.2025-2065

Frank J, Schönherr L (2021) WaveFake: A data set to facilitate audio deepfake detection. In: Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks

Frank J, Eisenhofer T, Schönherr L, et al (2020) Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the 37th International Conference

on Machine Learning (ICML), pp 3247–3258

Fröhling L, Zubiaga A (2021) Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and grover. PeerJ Computer Science 7:e443. https://doi.org/10.7717/peerj-cs.443

Gaikawad M, Ghosh S (2025) A robust and lightweight cnn-transformer model for audio deepfake detection in indian languages. In: 2025 7th International Conference on Signal Processing, Computing and Control (ISPCC), IEEE, pp 382–387

Gehrmann S, Strobelt H, Rush AM (2019) GLTR: Statistical detection and visualization of generated text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp 111–116

Gloaguen T, Staab R, Jovanović N, et al (2026) Watermarking diffusion language models. In: Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)

Goodfellow I, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial nets. In: Advances in Neural Information Processing Systems

Gu H, Yi J, Wang C, et al (2025) Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp 11736–11745

Gu Y, Zhao X, Gong C, et al (2020) Deepfake video detection using audio-visual consistency. In: International Workshop on Digital Watermarking, Springer, pp 168–180

Guillaro F, Zingarini G, Usman B, et al (2025) A bias-free training paradigm for more general AI-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 18685–18694

Gunn S, Zhao X, Song D (2025) An undetectable watermark for generative image models. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Guo B, Zhang X, Wang Z, et al (2023a) How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597 https://doi.org/10.48550/arXiv.2301.07597, URL https://arxiv.org/abs/2301.07597

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp 1321–1330

Guo H, Wang X, Lyu S (2023b) Detection of real-time deepfakes in video conferencing with active probing and corneal reflection. In: Proceedings of the IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1–5, https://doi.org/10.1109/ICASSP49357.2023.10094720

Gushchin A, Abud K, Shumitskaya E, et al (2026) NTIRE 2026 challenge on robust AI-generated image detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops

Haliassos A, Vougioukas K, Petridis S, et al (2021) Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5037–5047, https://doi.org/10.1109/CVPR46437.2021.00500

Hao W, Li R, Zhao W, et al (2025) Learning to rewrite: Generalized LLM-generated text detection. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Association for Computational Linguistics, pp 6421–6434, https://doi.org/10.18653/v1/2025.acl-long.322

He Q, Peng C, Liu D, et al (2024) GazeForensics: Deepfake detection via gaze-guided spatial inconsistency learning. Neural Networks 180:106636. https://doi.org/10.1016/j.neunet.2024.106636

Hu M, Diao Y, Miao C, et al (2025a) MVAD: A benchmark dataset for multimodal AI-generated video-audio detection. arXiv preprint arXiv:2512.00336

Hu R, Zhang J, Li Y, et al (2025b) VideoShield: Regulating diffusion-based video generation models via watermarking. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Huang W, Gu Y, Wang Z, et al (2025a) SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Association for Computational Linguistics, pp 9985–9998, https://doi.org/10.18653/v1/2025.acl-long.493

Huang Z, Hu J, Li X, et al (2025b) SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 28831–28841

Internò C, Geirhos R, Olhofer M, et al (2025) AI-generated video detection via perceptual straightening. In: Advances in Neural Information Processing Systems (NeurIPS)

Jawahar G, Abdul-Mageed M, Lakshmanan LVS (2020) Automatic detection of machine generated text: A critical survey. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp 2296–2309, https://doi.org/10.18653/v1/2020.coling-main.208

Jbara WA, Soud JH (2025) Deepfake audio detection via mfcc features and mel-spectrogram using deep learning. In: AIP Conference Proceedings, AIP Publishing LLC, p 030027

Jeong Y, Kim D, Min S, et al (2022) FingerprintNet: Synthesized fingerprints for generated image detection. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp 76–94

Jiang C, Dong W, Zhang Z, et al (2026) Ivy-Fake: A unified explainable framework and benchmark for image and video AIGC detection. In: Proceedings of the 2026 International Conference on Multimedia Retrieval (ICMR). ACM, pp 2438–2447, https://doi.org/10.1145/3805622.3810615

Jiang Z, Zhang J, Gong NZ (2023) Evading watermark based detection of AI-generated content. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp 1168–1181

Jiang Z, Guo M, Li K, et al (2025) VideoMarkBench: Benchmarking robustness of video watermarking. arXiv preprint arXiv:2505.21620

Jiang-Lin JY, Huang KY, Zou L, et al (2026) TriDF: Evaluating perception, detection, and hallucination for interpretable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Jisc (2025) Ai detection and assessment - an update for 2025. National Centre for AI URL https://nationalcentreforai.jiscinvolve.org/wp/2025/06/24/ai-detection-assessment-2025/

Jung Jw, Heo HS, Tak H, et al (2022) AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6367–6371, https://doi.org/10.1109/ICASSP43922.2022.9747766

Karageorgiou D, Papadopoulos S, Kompatsiaris I, et al (2025) Any-resolution AI-generated image detection by spectral learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 18706–18717

Karras T, Laine S, Aittala M, et al (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8110–8119, https://doi.org/10.1109/CVPR42600.2020.00813

Kassis A, Hengartner U (2025) UnMarker: A universal attack on defensive image watermarking. In: Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP). IEEE, pp 2602–2620, https://doi.org/10.1109/SP61157.2025.00005

Kaur A, Noori Hoshyar A, Saikrishna V, et al (2024) Deepfake video detection: Challenges and opportunities. Artificial Intelligence Review 57(6):1–47

Kawa P, Plata M, Czuba M, et al (2023) Improved deepfake detection using whisper features. In: Proc. INTERSPEECH 2023, pp 4009–4013

Khalid H, Tariq S, Kim M, et al (2021) FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In: Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks

Kietzmann J, Lee L, McCarthy A, et al (2020) Deepfakes: Trick or treat? Business Horizons 63(2):135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Kim Y, Yun K, Hong S, et al (2026) X-AVDT: Audio-visual cross-attention for robust deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4403–4414

Kirchenbauer J, Geiping J, Wen Y, et al (2023) A watermark for large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML), vol 202. PMLR, pp 17061–17084

Kundu R, Xiong H, Mohanty V, et al (2025) Towards a universal synthetic video detector: From face or background manipulations to fully AI-generated content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 28050–28060, https://doi.org/10.1109/CVPR52734.2025.02612

Li T, Huang Z, Wen H, et al (2026a) Omni-fake: Benchmarking unified multimodal social media deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 30299–30311

Li Y, Yang X, Sun P, et al (2020) Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3204–3213, https://doi.org/10.1109/CVPR42600.2020.00327

Li Y, Li Q, Cui L, et al (2024a) MAGE: Machine-generated text detection in the wild. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics, pp 36–53, https://doi.org/10.18653/v1/2024.acl-long.3

Li Y, Milling M, Specia L, et al (2024b) From audio deepfake detection to ai-generated music detection – a pathway and overview. URL https://arxiv.org/abs/2412.00571, arXiv:2412.00571

Li Y, Zhang M, Ren M, et al (2024c) Cross-domain audio deepfake detection: Dataset and analysis. In: Proceedings of the 2024 Conference on Empirical Methods in

Natural Language Processing, pp 4977–4983

Li Y, Lyu C, Wang L, et al (2025a) The affective bridge: Unifying feature representations for speech deepfake detection. URL https://arxiv.org/abs/2512.11241, arXiv:2512.11241

Li Y, Wang L, Wang Y, et al (2025b) Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components. URL https://arxiv.org/abs/2512.08403, arXiv:2512.08403

Li Y, Zhu D, Cui X, et al (2025c) Celeb-df++: A large-scale challenging video deepfake benchmark for generalizable forensics. arXiv preprint arXiv:2507.18015

Li Z, Yan J, He Z, et al (2025d) Is artificial intelligence generated image detection a solved problem? In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track

Li Z, Jiang C, Zhao H, et al (2026b) Preserving forgery artifacts: AI-generated video detection at native scale. In: Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)

Liang W, Yuksekgonul M, Mao Y, et al (2023) GPT detectors are biased against non-native English writers. Patterns 4(7):100779. https://doi.org/10.1016/j.patter.2023.100779

Liang W, Zhang Y, Wu Z, et al (2024) Mapping the increasing use of LLMs in scientific papers. In: Proceedings of the First Conference on Language Modeling (COLM)

Lin L, Gupta N, Zhang Y, et al (2024) Detecting multimedia generated by large AI models: A survey. arXiv preprint arXiv:2402.00045

Lin L, Santosh, Wu M, et al (2025) AI-Face: A million-scale demographically annotated AI-generated face dataset and fairness benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:2406.00783

Liu H, Tan Z, Tan C, et al (2024) Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10770–10780

Liu R, Huang H, Xiao X, et al (2025a) Zero-shot detection of LLM-generated text via implicit reward model. In: Advances in Neural Information Processing Systems (NeurIPS)

Liu S, Liu X, Li C, et al (2026) MGT-Prism: Enhancing domain generalization for machine-generated text detection via spectral alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 40. AAAI Press, pp 32132–32140,

https://doi.org/10.1609/aaai.v40i38.40485

Liu T, Truong DT, Das RK, et al (2025b) Nes2Net: A lightweight nested architecture for foundation model driven speech anti-spoofing. IEEE Transactions on Information Forensics and Security 20:12005–12018. https://doi.org/10.1109/TIFS.2025.3626963

Liu X, Wang X, Sahidullah M, et al (2023) ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31:2507–2522. https://doi.org/10.1109/TASLP.2023.3285283

Liu Y, Lu L, Jin J, et al (2025c) XAttnMark: Learning robust audio watermarking with cross-attention. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), vol 267. PMLR, pp 38987–39015

Lopez JA, Stemmer G, Cordourier Maruri H (2026) Generalizable detection of audio deepfakes. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 19057–19061, https://doi.org/10.1109/ICASSP55912.2026.11463470

Lu N, Liu S, He R, et al (2024) Large language models can be guided to evade AI-generated text detection. Transactions on Machine Learning Research

Lukáš J, Fridrich J, Goljan M (2006) Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security 1(2):205–214. https://doi.org/10.1109/TIFS.2006.873602

Luo Y, Du J, Yan K, et al (2024) LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17006–17015

Macko D, Moro R, Uchendu A, et al (2023) MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 9960–9987, https://doi.org/10.18653/v1/2023.emnlp-main.616

Mady M (2025) HC3 Gemini 2.0 Flash responses. Hugging Face dataset, https://huggingface.co/datasets/mohamedmady/HC3-Gemini-Flash-Responses

Mady M, Reschke J, Schuller BW (2026a) Academic text dataset: A large-scale academic-domain benchmark for AI-generated text detection. Hugging Face dataset, https://huggingface.co/datasets/mohamedmady/Academic-Text-arxiv-gpt-gemini

Mady M, Reschke J, Schuller BW (2026b) Feature-augmented transformers for robust AI-text detection across domains and generators. arXiv preprint arXiv:2605.03969

Marra F, Gragnaniello D, Cozzolino D, et al (2019) Do GANs leave artificial fingerprints? In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, pp 506–511

Meng Z, Peng B, Dong J, et al (2024) Artifact feature purification for cross-domain detection of AI-generated images. Computer Vision and Image Understanding

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: A survey. ACM Computing Surveys 54(1):1–41. https://doi.org/10.1145/3425780

Mitchell E, Lee Y, Khazatsky A, et al (2023) DetectGPT: Zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning (ICML), pp 24950–24962

Müller A, Lukovnikov D, Thietke J, et al (2025) Black-box forgery attacks on semantic watermarks for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 20937–20946, https://doi.org/10.1109/CVPR52734.2025.01950

Nguyen D, Astrid M, Kacem A, et al (2025) Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 10786–10796, https://doi.org/10.1109/ICCV51701.2025.01004

Nguyen HH, Yamagishi J, Echizen I (2024) Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. In: Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), pp 1–10

Nguyen-Le HH, Tran VT, Nguyen DT, et al (2026) Deepfake detection across image, video, and audio: A comprehensive survey with empirical evaluation of generalization and robustness. Artificial Intelligence Review 59(10):203. https://doi.org/10.1007/s10462-026-11608-4

Ni Z, Yan Q, Huang M, et al (2026) GenVidBench: A 6-million benchmark for AI-generated video detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 40. AAAI Press, pp 15582–15590, https://doi.org/10.1609/aaai.v40i18.38587

Ojha U, Li Y, Lee YJ (2023) Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 24480–24489

van den Oord A, Dieleman S, Zen H, et al (2016) WaveNet: A generative model for raw audio. In: Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW). ISCA, p 125

OpenAI (2023) New ai classifier for indicating ai-written text. OpenAI blog, URL https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/, accessed 2026-02-01

OpenAI (2026) Advancing content provenance for a safer, more transparent AI ecosystem. https://openai.com/index/advancing-content-provenance/

Park J, Owens A (2025) Community forensics: Using thousands of generators to train fake image detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8245–8257

Patel Y, Tanwar S, Bhattacharya P, et al (2023) An improved dense cnn architecture for deepfake image detection. IEEE Access 11:22081–22095

Pei G, Zhang J, Hu M, et al (2026) Deepfake generation and detection: A benchmark and survey. ACM Computing Surveys 58(11):273:1–273:41. https://doi.org/10.1145/3801962

Pu X, Cheng Z, Yuan L, et al (2026) Breaking the generator barrier: Disentangled representation for generalizable AI-text detection. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). Association for Computational Linguistics, pp 2586–2598, https://doi.org/10.18653/v1/2026.acl-long.120

Qin Z, Ji Y, Tao R, et al (2026) Scaling up AI-generated image detection with generator-aware prototypes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Qu W, Zheng W, Tao T, et al (2025) Provably robust multi-bit watermarking for AI-generated text. In: Proceedings of the 34th USENIX Security Symposium (USENIX Security). USENIX Association, pp 201–220

Quaremba G, Black E, Vrandečić D, et al (2026) TSM-Bench: Detecting LLM-generated text in real-world Wikipedia editing practices. In: Proceedings of the 14th International Conference on Learning Representations (ICLR)

Rabhi M, Bakiras S, Di Pietro R (2024) Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications 250:123941

Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners. Tech. rep., OpenAI

Radford A, Kim JW, Xu T, et al (2023) Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning (ICML), vol 202. PMLR, pp 28492–28518

Rahman MA, Paul B, Sarker NH, et al (2023) ArtiFact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In: 2023 IEEE International Conference on Image Processing (ICIP). IEEE, pp 2200–2204, https://doi.org/10.1109/ICIP49359.2023.10222083

Rahman MA, Hakim ZIA, Sarker NH, et al (2025) SONICS: Synthetic or not – identifying counterfeit songs. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Rajan AS, Ojha U, Schloesser J, et al (2025) Aligned datasets improve detection of latent diffusion-generated images. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Ricker J, Lukovnikov D, Fischer A (2024) AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9130–9140

Rombach R, Blattmann A, Lorenz D, et al (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10684–10695

Rössler A, Cozzolino D, Verdoliva L, et al (2019) FaceForensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 1–11

Russell J, Karpinska M, Iyyer M (2025) People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Association for Computational Linguistics, pp 5342–5373, https://doi.org/10.18653/v1/2025.acl-long.267

Sadasivan VS, Kumar A, Balasubramanian S, et al (2025) Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. Transactions on Machine Learning Research

Sander T, Fernandez P, Durmus AO, et al (2025) Watermark anything with localized messages. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Shahjahan ATM, Mannan M, Ben Hamza A, et al (2026) On improving robustness of deepfake image detectors. In: Proceedings of the 35th USENIX Security Symposium (USENIX Security). USENIX Association

Shen J, Pang R, Weiss RJ, et al (2018) Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp

4779–4783, https://doi.org/10.1109/ICASSP.2018.8461368

Sisman B, Yamagishi J, King S, et al (2021) An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29:132–157. https://doi.org/10.1109/TASLP.2020.3038524

Smeu S, Boldisor DA, Oneata D, et al (2025) Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 18815–18825

Song X, Guo X, Zhang J, et al (2024) On learning multi-modal forgery representation for diffusion generated video detection. In: Advances in Neural Information Processing Systems (NeurIPS), pp 122054–122077

Stowe K, Afanaseva S, Raimundo RC, et al (2026) Identifying bias in machine-generated text detection. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). Association for Computational Linguistics, pp 2383–2395, https://doi.org/10.18653/v1/2026.acl-long.109

Su J, Zhuo TY, Wang D, et al (2023a) DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp 12395–12412, https://doi.org/10.18653/v1/2023.findings-emnlp.827

Su Z, Wu X, Zhou W, et al (2023b) HC3 Plus: A semantic-invariant human ChatGPT comparison corpus. arXiv preprint arXiv:2309.02731

Sumsub (2023) Identity fraud report 2023. Tech. rep., Sumsub, URL https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/

Sun C, Jia S, Hou S, et al (2023) Ai-synthesized voice detection using neural vocoder artifacts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 904–912

Sun Q, Li Y, Alturki E, et al (2026) Towards friendly AI: A comprehensive review and new perspectives on human-AI alignment. AI and Ethics 6(2):193. https://doi.org/10.1007/s43681-026-01070-x

Tak H, Todisco M, Wang X, et al (2022) Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In: The Speaker and Language Recognition Workshop (Odyssey 2022), pp 112–119, https://doi.org/10.21437/Odyssey.2022-16

Tan C, Zhao Y, Wei S, et al (2023) Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12105–12114

Tan C, Zhao Y, Wei S, et al (2024a) Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5052–5060

Tan C, Zhao Y, Wei S, et al (2024b) Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 28130–28139

Tan C, Tao R, Liu H, et al (2025) C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in AI-generated image detection. In: Proceedings of the AAAI Conference on Artificial Intelligence

Tan H, Lan J, Tan Z, et al (2026) Veritas: Generalizable deepfake detection via pattern-aware reasoning. In: Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)

Tan X, Qin T, Soong F, et al (2021) A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561

Tolosana R, Vera-Rodriguez R, Fierrez J, et al (2020) Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion 64:131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Verdoliva L (2020) Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing 14(5):910–932. https://doi.org/10.1109/JSTSP.2020.3002101

Verma K, Mittal D, Samanta S, et al (2025) Deepfake audio detection: A comparative study of advanced deep learning models. IEEE Access

Wang J, Lv C, Li X, et al (2025) Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4233–4245

Wang T, Zhang Y, Qi S, et al (2024a) Security and privacy on generative data in AIGC: A survey. ACM Computing Surveys 57(4):1–34. https://doi.org/10.1145/3703626

Wang X, Yamagishi J (2022) Investigating self-supervised front ends for speech spoofing countermeasures. In: Proceedings of the Speaker and Language Recognition Workshop (Odyssey). ISCA, pp 100–106, https://doi.org/10.21437/Odyssey.2022-14

Wang X, Xiao Y, Zhu X (2017) Feature selection based on cqccs for automatic speaker verification spoofing. In: Interspeech, pp 32–36

Wang X, Delgado H, Tak H, et al (2024b) ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. In: The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), pp 1–8, https://doi.org/10.21437/ASVspoof.2024-1

Wang X, Delgado H, Tak H, et al (2026) ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. Computer Speech & Language 95:101825. https://doi.org/10.1016/j.csl.2025.101825

Wang Y, Mansurov J, Ivanov P, et al (2024c) M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics, pp 1369–1407, https://doi.org/10.18653/v1/2024.eacl-long.83

Wang Z, Cheng Z, Xiong J, et al (2024d) A timely survey on vision transformer for deepfake detection. arXiv preprint arXiv:2405.08463

Wen H, Li T, Huang Z, et al (2025a) BusterX++: Towards unified cross-modal AI-generated content detection and explanation with MLLM. arXiv preprint arXiv:2507.14632

Wen S, Ye J, Feng P, et al (2025b) Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In: Advances in Neural Information Processing Systems (NeurIPS)

Wen Y, Kirchenbauer J, Geiping J, et al (2023) Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In: Advances in Neural Information Processing Systems (NeurIPS)

Wesselkamp V, Rieck K, Arp D, et al (2022) Misleading deep-fake detection with GAN fingerprints. In: 2022 IEEE Security and Privacy Workshops (SPW). IEEE, pp 59–65

West A, Weng Y, Zhu M, et al (2025) AI-generated text is non-stationary: Detection via temporal tomography. arXiv preprint arXiv:2508.01754

Westerlund M (2019) The emergence of deepfake technology: A review. Technology Innovation Management Review 9(11):40–53. URL https://timreview.ca/article/1282

Wijethunga R, Matheesha D, Al Noman A, et al (2020) Deepfake audio detection: a deep learning based solution for group conversations. In: 2020 2nd International

conference on advancements in computing (ICAC), IEEE, pp 192–197

Wu H, Li F, Tu Z, et al (2026a) Zero-shot detection of AI-generated image via RAW-RGB alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Wu J, Zhan R, Wong DF, et al (2024) DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track

Wu J, Yang S, Zhan R, et al (2025a) A survey on LLM-generated text detection: Necessity, methods, and future directions. Computational Linguistics 51(1):275–338. https://doi.org/10.1162/coli_a_00549

Wu J, Liu Y, Zhu C, et al (2026b) DetectRL-X: Towards reliable multilingual and real-world LLM-generated text detection. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026). Association for Computational Linguistics, pp 38247–38294, https://doi.org/10.18653/v1/2026.acl-long.1773

Wu S, Liu J, Li J, et al (2025b) Few-shot learner generalizes across AI-generated image detection. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), vol 267. PMLR, pp 67449–67460

Wu W, Zhou W, Zhang W, et al (2022) Capturing the lighting inconsistency for deepfake detection. In: Proceedings of the International Conference on Artificial Intelligence and Security (ICAIS). Springer, pp 637–647, https://doi.org/10.1007/978-3-031-06788-4_52

Xiao Y, Das RK (2025) XLSR-mamba: A dual-column bidirectional state space model for spoofing attack detection. IEEE Signal Processing Letters 32:1276–1280. https://doi.org/10.1109/LSP.2025.3547861

Xie Y, Lu Y, Fu R, et al (2025) The Codecfake dataset and countermeasures for the universally detection of deepfake audio. IEEE Transactions on Audio, Speech and Language Processing 33:386–400. https://doi.org/10.1109/TASLPRO.2025.3525966

Xu Q, Mu W, Li J, et al (2026) Advancements in AI-generated content forensics: A systematic literature review. ACM Computing Surveys 58(3):60:1–60:36. https://doi.org/10.1145/3760526

Xu Y, Wang Y, Bi Y, et al (2025a) Training-free LLM-generated text detection by mining token probability sequences. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Xu Z, Zhang X, Li R, et al (2025b) FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In: International Conference

on Learning Representations (ICLR)

Xue J, Fan C, Lv Z, et al (2022) Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In: Proceedings of the 1st international workshop on deepfake detection for audio multimedia, pp 19–26

Yamagishi J, Wang X, Todisco M, et al (2021) ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In: Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof). ISCA, pp 47–54, https://doi.org/10.21437/ASVspoof.2021-8

Yan S, Li O, Cai J, et al (2025a) A sanity check for AI-generated image detection. In: Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)

Yan Z, Yao T, Chen S, et al (2024) DF40: Toward next-generation deepfake detection. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, arXiv:2406.13495

Yan Z, Wang J, Jin P, et al (2025b) Orthogonal subspace decomposition for generalizable AI-generated image detection. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), vol 267. PMLR, pp 70268–70288

Yan Z, Zhao Y, Chen S, et al (2025c) Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12615–12625, https://doi.org/10.1109/CVPR52734.2025.01177

Yan Z, Zhao Y, Wang H (2025d) VoiceWukong: Benchmarking deepfake voice detection. In: Proceedings of the 34th USENIX Security Symposium (USENIX Security). USENIX Association, pp 4561–4580

Yang X, Cheng W, Wu Y, et al (2024) DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. In: Proceedings of the 12th International Conference on Learning Representations (ICLR)

Yang Z, Feng Z, Huo R, et al (2025) The imitation game revisited: A comprehensive survey on recent advances in AI-generated text detection. Expert Systems with Applications 272:126694. https://doi.org/10.1016/j.eswa.2025.126694

Ye J, Zhou B, Huang Z, et al (2025) LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In: International Conference on Learning Representations (ICLR)

Ye P, Ren H, Li Z, et al (2026) Securing large language models: A survey of watermarking and fingerprinting techniques. ACM Computing Surveys 58(7):187:1–187:35. https://doi.org/10.1145/3773028

Yermakov A, Cech J, Matas J, et al (2026) Deepfake detection that generalizes across benchmarks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Yi J, Wang C, Tao J, et al (2023) Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970

Yu N, Davis L, Fritz M (2019) Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 7556–7566

Zellers R, Holtzman A, Rashkin H, et al (2019) Defending against neural fake news. In: Advances in Neural Information Processing Systems (NeurIPS)

Zeng C, Tang S, Chen Y, et al (2025) Human texts are outliers: Detecting LLM-generated texts via out-of-distribution detection. In: Advances in Neural Information Processing Systems (NeurIPS)

Zhang S, Lian Z, Yang J, et al (2025) Physics-driven spatiotemporal modeling for AI-generated video detection. In: Advances in Neural Information Processing Systems (NeurIPS)

Zhao X, Gunn S, Christ M, et al (2025) SoK: Watermarking for AI-generated content. In: Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP). IEEE, pp 2621–2639, https://doi.org/10.1109/SP61157.2025.00178

Zheng X, Zhang D, Sun C (2025) Frequency-enhanced multi-modal consistency learning for audio-visual deepfake detection. In: Proceedings of the 7th ACM International Conference on Multimedia in Asia, pp 1–7

Zhou P, Wang L, Liu Z, et al (2024) A survey on generative ai and llm for video generation, understanding, and streaming. arXiv preprint arXiv:2404.16038

Zhou Z, Luo Y, Wu Y, et al (2025) AIGI-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 18746–18758

Zhu M, Chen H, Huang M, et al (2023a) GenDet: Towards good generalizations for AI-generated image detection. arXiv preprint arXiv:2312.08880

Zhu M, Chen H, Yan Q, et al (2023b) GenImage: A million-scale benchmark for detecting AI-generated image. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track

Zong W, Chow YW, Susilo W, et al (2025) AudioMarkNet: Audio watermarking for deepfake speech detection. In: 34th USENIX Security Symposium (USENIX

Security 25), pp 4663–4682