The Deepfake Defense Stack: Why No Single Layer Works and How They Must Compose

Anonymous authors
Paper under double-blind review

Abstract

Every defense against AI-synthesized media, whether passive detection, invisible watermarking, or content provenance, has been shown to fail when deployed in isolation. Detectors suffer 45–50% accuracy degradation from laboratory to deployment and collapse on outputs from unseen generator architectures. Watermarks are removable by regeneration attacks and screenshot capture. Provenance metadata is stripped by most social media platforms. Yet no prior work has formally analyzed how these defenses *compose*: which attack classes each layer blocks, where cascade failures propagate, and what residual vulnerabilities survive the full stack.

We present the first composition analysis of deepfake defenses. Through a Defense Composition Matrix covering 58 detection methods, 23 proactive defense systems, and 7 adversarial attack classes, we map the interaction between three defense layers (detection, watermarking, provenance) and seven attack classes. We identify two attack classes that penetrate all three layers simultaneously, one that bypasses the primary trust layer, and three emergent composition patterns where stacking defenses creates vulnerabilities absent from any individual layer. We formulate a Detection Ceiling Conjecture arguing, with supporting evidence, that post-hoc detection faces an information-theoretic bound that provenance-based approaches do not share. Our composition analysis draws on 190 papers spanning generation, detection, proactive defense, adversarial attacks, benchmarks, and the societal impact of AI-synthesized media across the 2014–2026 period. We provide focused technical background for each defense layer (Sections 3–5) sufficient to support the composition analysis; readers seeking comprehensive coverage of individual layers should consult the dedicated surveys cited in Table 1. We identify eight open problems with falsifiable hypotheses and proposed experimental protocols.

1 Introduction

The AI-synthesized media ecosystem now generates content that is indistinguishable from authentic recordings under typical viewing conditions. An estimated 8 million synthetic media files circulated on social media in 2025 DeepStrike (2025); deepfake-enabled fraud reached an estimated $1.1 billion in 2025 alone ($2.19 billion cumulative as of April 2026) Surfshark (2026); and detection accuracy degrades by 45–50% from laboratory benchmarks to real-world deployment Chandra et al. (2025).

The response to this crisis has produced three classes of defense mechanisms. *Passive detection* builds classifiers that distinguish real from synthetic content based on pixel-level analysis. A second approach, *watermarking*, embeds imperceptible markers in AI-generated outputs to enable later identification. The third and most recent class, *content provenance*, takes a fundamentally different tack: rather than analyzing the content itself, it attaches cryptographic credentials at the point of capture to authenticate the origin of media.

Each class has been studied extensively in isolation. Detection methods spanning five paradigms have been surveyed by Pei et al. (2024), Croitoru et al. (2024), and at least 15 others since 2024. Watermarking has

Table 1: Positioning against related surveys. ✓= covered in depth; ∘ = mentioned but not primary focus; – = not addressed. No prior survey covers composition analysis or formal threat modeling across all defense layers.

Survey	Year	Gen.	Det.	Prov.	WM	Gov.	Cross-mod.	Compos.	Formal
Tolosana et al. (Info. Fusion)	2020	✓	✓	–	–	–	–	–	–
Mirsky & Lee (2021)	2021	✓	✓	–	–	–	–	–	–
Pei et al. (ACM CSUR)	2024	✓	✓	–	–	–	–	–	–
Croitoru et al. (arXiv)	2024	✓	✓	–	–	–	∘	–	–
Wang et al. (ACM CSUR)	2024	–	✓	–	–	–	–	–	–
Deng et al. (ACM CSUR)	2025	–	✓	∘	✓	–	–	–	–
Zhao et al. (IEEE S&P)	2025	–	–	∘	✓	–	–	–	✓
Zou et al. (arXiv)	2025	∘	✓	–	–	∘	–	–	–
Li et al. (ACM CSUR)	2025	–	✓	–	–	–	–	–	–
Gov. Info. Quarterly Survey	2025	∘	✓	✓	∘	✓	–	–	–
This work	**2026**	✓	✓	✓	✓	∘	✓	✓	✓

been systematized by Zhao et al. (2025) at IEEE S&P 2025. Proactive defenses (combining watermarking and disruption) are covered by Deng et al. (2025) in ACM Computing Surveys.

What no prior work has addressed is the question at the heart of deployment: *how do these defenses compose?* When detection fails against a new generator, does watermarking catch the content? When a watermark is stripped by a screenshot, does provenance survive? When an adversary targets all three layers simultaneously, which attack strategies succeed and which are blocked by the redundancy of the stack?

We provide the first composition analysis of deepfake defenses. Our contributions are:

1. A **Defense Composition Matrix** mapping seven attack classes against three defense layers, identifying which attacks each layer blocks, degrades, or is bypassed by (§6).
2. A **Cascade Failure Taxonomy** documenting how attacks that defeat one layer propagate through the stack, identifying two attack classes that penetrate all three layers and one that bypasses the primary trust layer (§6).
3. A **Detection Ceiling Conjecture** formalizing the information-theoretic bound on post-hoc detection that provenance does not share (§7).
4. A unified survey of generation mechanisms across six modalities, detection across five paradigms, and proactive defenses, organized as a comparative taxonomy by detectability properties (§3–5).
5. A **gap analysis** positioning this work against 25+ competing surveys published since 2024 (§2).
6. Eight **open problems** with falsifiable hypotheses and proposed experimental protocols (§11).

Deng et al. (2025) provide the closest existing work: a multi-layer taxonomy covering detection, disruption, and authentication. We build on their taxonomy to analyze *composition*: how layers interact, where cascade failures occur, and what the joint failure probability is when layers are stacked. Taxonomy organizes what exists; composition analysis tells you how to build a defense system.

2 Related Surveys and Positioning

At least 25 surveys published since 2024 cover portions of the terrain this paper addresses. Table 1 positions this work against the most relevant prior surveys across eight dimensions.

The key observation from Table 1 is that existing surveys cover at most four of the eight dimensions, and none addresses composition analysis or formal threat modeling across all three defense layers. The closest

work, Deng et al. (2025), provides a multi-layer taxonomy covering detection, disruption, and authentication, but does not analyze how these layers interact under adversarial pressure or where cascade failures occur. Zhao et al. (2025) provide formal threat models for watermarking specifically but do not extend this analysis to the full defense stack. Our contribution fills the composition and formal analysis columns.

2.1 Search Strategy and Scope

We surveyed papers published between 2014 and March 2026 across arXiv, Semantic Scholar, Google Scholar, IEEE Xplore, and ACM Digital Library using queries combining "deepfake" with "detection," "provenance," "watermark," "C2PA," "adversarial," and "foundation model." For generation and detection methods (Sections 3–4), we prioritized papers at top-tier venues (NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, IEEE TPAMI, ACM CSUR). For proactive defenses (§5), we included industry standards (C2PA specifications), commercial products (SynthID, VideoSeal), and government guidance (NSA/CISA). For harm quantification (§10), we drew on government reports, verified incident databases (Surfshark, Sumsub), and investigative journalism, noting the provenance class of each statistic. The resulting evidence base comprises 190 papers and reports. We do not claim exhaustive coverage; we prioritize representative landmark works across each defense layer and attack class to support the composition analysis.

3 Generation: A Comparative Taxonomy

We provide focused background on generation mechanisms, organized by the forensic signals each architecture produces and the defense layers each is vulnerable to. Comprehensive coverage of generation methods is available in Pei et al. (2024) and Croitoru et al. (2024); we summarize the properties relevant to our composition analysis.

3.1 Generative Adversarial Networks

The GAN framework (Goodfellow et al., 2014) introduced the minimax game between generator G and discriminator D:

$$\min_G \max_D \; \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{1}$$

A critical property follows directly from this formulation: the generator is explicitly optimized to produce outputs indistinguishable from real data. Any forensic classifier faces an adversary that has been specifically trained to defeat exactly such classifiers. Forensic adversariality is therefore an architectural feature of the GAN paradigm.

The architecture matured through three phases. Karras et al.'s Progressive GAN (Karras et al., 2018) introduced layer-by-layer resolution scaling from 4×4 to 1024×1024, enabling the first photorealistic face synthesis. StyleGAN (Karras et al., 2019) introduced a mapping network $f : \mathcal{Z} \to \mathcal{W}$ that transforms a latent code into a disentangled intermediate space, with adaptive instance normalization injecting style at each synthesis layer. StyleGAN2 (Karras et al., 2020) replaced AdaIN with weight demodulation, reducing FID to 2.84 on FFHQ at 1024^2. StyleGAN3 (Karras et al., 2021) addressed "texture sticking" through formal application of the Nyquist-Shannon sampling theorem, achieving alias-free synthesis with continuous translation and rotation equivariance.

Autoencoder-based face swapping. The autoencoder paradigm, exemplified by DeepFaceLab (Petrov et al., 2023) (responsible for over 95% of deepfake videos with 35,000+ GitHub stars), uses a shared encoder and identity-specific decoders: $\hat{A}_B = D_A(E(f_B))$ where the target face f_B is encoded through the shared encoder and decoded through the source decoder. SimSwap (Chen et al., 2020) introduced ID Injection for single-stage generalized swapping without per-identity training. HifiFace incorporated 3D Morphable Model supervision for shape-aware identity transfer at 1024 px.

Face reenactment. Face2Face (Thies et al., 2016) introduced real-time RGB-only reenactment via dense photometric tracking at 30+ fps. The First Order Motion Model (Siarohin et al., 2019) provided training-free animation through self-supervised keypoint learning with local affine transformations. The accessibility

trajectory has been stark: what required PhD-level expertise in 2014 can now be accomplished with one-click mobile applications for approximately $10 per 50 videos.

Detectability properties. GANs leave systematic frequency-domain artifacts traceable to upsampling operations Frank et al. (2020), enabling spectral detection at 95%+ accuracy across multiple architectures. Face-swapping methods leave blending boundaries that Face X-Ray exploits. These artifacts are architecture-specific and do not transfer to diffusion models.

3.2 Diffusion Models and Text-to-Image Systems

Ho et al. (2020) established denoising diffusion probabilistic models. The forward process gradually adds Gaussian noise: $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$, and a neural network $\epsilon_\theta(x_t, t)$ learns to reverse this process. The breakthrough for practical synthesis was Rombach et al.'s Latent Diffusion (CVPR 2022), which moved the process into a pre-trained autoencoder's latent space, reducing computation by approximately 64× while adding cross-attention conditioning on text embeddings. Classifier-free guidance interpolates conditional and unconditional predictions with scale s: $\tilde{\epsilon}_\theta = \epsilon_\theta(\cdot, \varnothing) + s \cdot (\epsilon_\theta(\cdot, y) - \epsilon_\theta(\cdot, \varnothing))$.

This architecture underlies Stable Diffusion (open-source, August 2022, progressing from 512^2 through SDXL to DiT-backbone SD 3), DALL-E 2 and 3 (Betker et al., 2023), and Midjourney (V1 through V7, adding video generation). The transition from GAN to diffusion generation has profound implications for detection. The diffusion process produces images through iterative refinement from noise, a mechanism that avoids the upsampling bottlenecks responsible for GAN spectral fingerprints.

Detectability properties. Diffusion models eliminate the spectral fingerprints that detectors relied on. Ricker et al. (2024) confirmed approximately 15.2 percentage-point AUROC drops when GAN-trained detectors encounter diffusion outputs. DF40 Yan et al. (2024) found detectors at 85% AUC on face-swap content dropping to 47% on Stable Diffusion outputs. Detection now depends on semantic features rather than architectural artifacts, a shift that has driven the foundation model approaches discussed in §4.

3.3 Video Generation

Video synthesis reached a critical inflection in 2024–2025. Sora (OpenAI, 2024) (discontinuation announced 24 March 2026; app closure 26 April 2026, API closure September 2026) uses a Diffusion Transformer (DiT) architecture operating on spacetime patches compressed by a spatial-temporal VAE, generating up to 20 seconds at 1080p (December 2024). Sora 2 (September 2025) added synchronized audio, persistent world state, and "Characters" (personalized avatars from short recordings) alongside a $1 billion Disney partnership. Google's Veo 2 (December 2024) and Veo 3 (May 2025) became the first models to natively generate synchronized dialogue, sound effects, and ambient audio alongside video. Runway released Gen-3 Alpha through Gen-4.5, with Gen-4.5 (December 2025) achieving top ranking on the Artificial Analysis Text-to-Video Leaderboard. Kuaishou's Kling AI iterated through 20+ versions, reaching Kling 2.6 with simultaneous audio-visual generation serving over 6 million users.

The barrier to entry is now $20 per month: photorealistic video with native audio, persistent characters, and arbitrary text-directed control, produced from a text prompt.

Detectability properties. Temporal inconsistency, the signal exploited by recurrent detectors (FTCN, XcepTemporal), is progressively eliminated as video generators develop explicit temporal modeling. Frame-level detection via foundation models (GenD) averages per-frame softmax probabilities, discarding inter-frame information. Recent CVPR 2025 methods (DFD-FCG (Han et al., 2025), Spatiotemporal Adapter (Yan et al., 2025b)) begin to address temporal detection, but none has been evaluated on production-quality outputs from Sora 2 or Veo 3.

3.4 Voice Cloning and Lip Synchronization

Voice synthesis has progressed from autoregressive models (WaveNet, 2016; Tacotron 2, 2018, MOS 4.53 vs. 4.58 for real speech) to zero-shot cloning. Microsoft's VALL-E (Wang et al., 2023) treats text-to-speech as conditional language modeling over discrete neural codec tokens, enabling voice cloning from 3 seconds

of enrollment audio, trained on 60,000 hours of LibriLight. VALL-E 2 (Chen et al., 2024) achieved what researchers described as "human parity" on LibriSpeech and VCTK benchmarks, though the system was withheld from public release due to misuse risks.

ElevenLabs (The Recursive, 2026) reached an $11 billion valuation in February 2026 ($500M Series D) with approximately $330M annual recurring revenue, offering voice cloning from ∼30 seconds in 70+ languages. The company served 41% of Fortune 500 companies as of early 2024. Its technology has been directly implicated in deepfake fraud incidents, including the Biden robocall of January 2024.

Wav2Lip (Prajwal et al., 2020) solved speaker-independent lip synchronization using a frozen pre-trained SyncNet discriminator, coupling arbitrary audio to any target face without per-subject training. Combined with voice cloning, this creates a complete audio-visual fabrication pipeline requiring only seconds of reference material.

3.5 Neural 3D Synthesis

Mildenhall et al.'s NeRF (Mildenhall et al., 2020) represents scenes as continuous volumetric functions $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (c, \sigma)$ rendered via differentiable volume integration. NeRF-based talking-head synthesis (AD-NeRF, ICCV 2021) extended this to audio-driven facial animation with full 3D consistency. Kerbl et al.'s 3D Gaussian Splatting (Kerbl et al., 2023) replaced implicit representations with explicit anisotropic 3D Gaussians and tile-based GPU rasterization, achieving real-time rendering at $\geq$100 fps at 1080p. This has been rapidly applied to talking-head synthesis: TalkingGaussian (ECCV 2024), GaussianTalker, and SyncGaussian (IJCAI 2025) all achieve real-time audio-driven avatars, enabling deepfake video calls that respond dynamically to conversation.

3.6 AI-Generated Text

Large language models represent a parallel but converging axis of AI-synthesized content. GPT, Claude, Gemini, and their successors generate text that is difficult for humans to distinguish from human-authored content. Wu et al. documented a 474% rise in AI-generated misinformation sites between January 2022 and May 2023. Detection methods span watermarking (Kirchenbauer et al.'s token-level statistical signatures), zero-shot statistical approaches (DetectGPT using log-probability curvature), and supervised neural classifiers. The convergence of text and media synthesis is critical: modern multimodal AI systems generate coordinated text, images, audio, and video from unified prompts, producing compound synthetic content that no single-modality detector can fully evaluate.

3.7 Comparative Detectability Taxonomy

Table 2 maps each generation architecture to the forensic signals it produces, organized by the defense layer best positioned to detect it.

The taxonomy reveals a critical trend: as generation technology advances from left to right in Table 2, the number of exploitable forensic signals decreases. Modern diffusion-based systems leave only weak semantic signals and metadata traces. This pattern motivates our composition analysis (§6): if detection signals are diminishing, the defense stack must rely increasingly on provenance and watermarking layers that operate independently of pixel content.

4 Detection: Five Paradigms

We summarize detection across five paradigms, focusing on each paradigm's robustness under the attack classes defined in §6. Comprehensive surveys of detection methods are provided by Pei et al. (2024), Deng et al. (2025), and Zou et al. (2025); we extract the properties relevant to the composition analysis.

Table 2: Generation architectures and their detectability properties across six forensic dimensions. ✓= signal present and exploitable; ∼ = signal weak or inconsistent; – = signal absent.

Architecture	*Freq. artifacts*	*Temporal*	*Biological*	*Blending*	*Semantic*	*Metadata*
GAN (face swap)	✓	∼	∼	✓	∼	✓
GAN (full synth.)	✓	–	–	–	∼	✓
Diffusion (image)	–	–	–	–	∼	✓
Diffusion (video)	–	∼	∼	–	∼	✓
Voice cloning	∼	–	–	–	✓	∼
Lip sync	–	✓	✓	∼	✓	∼
NeRF/3DGS	–	∼	–	–	∼	∼
LLM text	–	–	–	–	∼	∼

4.1 Temporal and Recurrent Architectures

Guéra and Delp (AVSS 2018) established temporal consistency as a discriminating signal by feeding CNN frame features into an LSTM for video-level classification. Chintha et al. (2020) combined XceptionNet with bidirectional LSTM/GRU layers and entropy-based cost functions for joint audio-visual detection, achieving 100% accuracy on FaceForensics++ and strong cross-domain generalization. FTCN (Zheng et al., 2021) reduced all spatial kernels to 1×1 to force exclusive reliance on temporal features, achieving state-of-the-art cross-dataset generalization simultaneously across four benchmarks.

Robustness under attack. Temporal signals are vulnerable to A1 (generator evasion): video generators with explicit temporal modeling (Sora, Veo) are progressively eliminating inter-frame inconsistencies. They are robust to A3 (compression) at moderate levels but degrade under heavy recompression.

4.2 Spatial CNN and Attention Methods

Frame-level spatial analysis forms the backbone of the field. XceptionNet (Chollet, 2017) (benchmarked by Rössler et al., ICCV 2019) achieves 99.26% accuracy on raw FF++ data through depthwise separable convolutions that efficiently capture fine-grained spatial artifacts. However, accuracy falls to 95.73% at moderate compression (c23) and 81.00% at heavy compression (c40), establishing that compression degrades detection systematically.

Face X-Ray (Li et al., 2020a) introduced a fundamentally different approach: detecting the universal blending boundary present in most face manipulations using an HRNet model trained *without any images from known manipulation methods*. This manipulation-agnostic design achieves 98.52% AUC on FF++ with 74.2% cross-dataset AUC on Celeb-DF. The key observation driving Face X-Ray is that most face manipulation methods share the common step of blending an altered face into an existing background image, and there exist intrinsic image discrepancies across blending boundaries regardless of the manipulation technique.

Multi-Attentional Detection (Zhao et al., 2021) recast detection as fine-grained classification using multiple spatial attention heads, a textural feature enhancement block, and attention-guided data augmentation, achieving ∼99% AUC on FF++ (c23).

Wang et al. (2020) demonstrated a remarkable property: a ResNet-50 trained only on ProGAN outputs generalizes at 92%+ AUC to 11 unseen CNN generators, suggesting shared low-level "CNN fingerprints" caused by upsampling artifacts. This property does not extend to diffusion models.

Robustness under attack. Spatial CNN detectors are highly vulnerable to A1 (28 pp cross- dataset drop from FF++ to Celeb-DF), A2 (Carlini & Farid (2020) reduce AUC to 0.0005 with white-box attacks), and A3 (18 pp drop under heavy compression).

4.3 Biological Signal and Frequency Methods

A third paradigm exploits physiological or physical signals that synthetic generators fail to reproduce. FakeCatcher (Ciftci et al., 2020) uses remote photoplethysmography (rPPG), the subtle skin color changes from cardiac blood flow in the 0.7–4 Hz band, as authenticity descriptors, achieving 94.65% on FF++ and commercialized by Intel claiming 96% accuracy. However, a 2025 Frontiers in Imaging study demonstrated that modern generators successfully reproduce rPPG signals, substantially undermining the assumption that biological signals are inherently unforgeable.

Li et al. (2018) exploited the observation that early deepfake training data rarely contained closed eyes, enabling a blink-frequency detector achieving 99% AUC on early datasets. This signal was rapidly nullified as generators incorporated natural blink patterns, illustrating the vulnerability of artifact-specific detection to generator adaptation.

Frank et al. (2020) (ICML 2020) demonstrated that all GAN architectures exhibit characteristic spectral artifacts in high-frequency DCT bands, traceable to nearest-neighbor or bilinear upsampling. A linear classifier trained on these spectral coefficients achieves over 95% accuracy across multiple GAN architectures. This represented the field's most productive forensic signal until the diffusion transition.

LipForensics (Haliassos et al., 2021) takes a semantic rather than artifact-based approach, pre-training a spatio-temporal network on lipreading across 500,000 utterances, then freezing the feature extractor and fine-tuning only the temporal classifier. Because the pre-training captures physiologically natural motion patterns, the method degrades more gracefully as generation quality improves.

Robustness under attack. Frequency signals are eliminated by the diffusion transition (A1; Ricker et al. (2024) confirmed 15.2 pp AUROC drop). Biological signals (rPPG, blink patterns) are being systematically reproduced by modern generators. Semantic approaches (LipForensics) are more resilient but still degrade on high-quality generators. All three signal types are vulnerable to A3 (compression strips high-frequency information that all three exploit).

4.4 Vision Transformer and Hybrid Architectures

Coccomini et al. (2022) combined EfficientNet-B0 with a Vision Transformer, achieving AUC 0.951 on DFDC. ViT-based temporal attention over frame embeddings outperforms both pure-CNN and pure-RNN approaches on challenging data.

4.5 Foundation Model Approaches

The most promising recent direction adapts large-scale pre-trained vision or vision-language models for detection. The turning point came when Ojha et al. (2023) showed that a linear classifier on top of unmodified CLIP representations could generalize to unseen generators, a result that suggested pre-trained vision encoders already capture the distinction between real and synthetic content at a coarse level.

This result catalyzed a wave of adaptation strategies. CLIPping the Deception (Khan & Dang-Nguyen, 2024) adapts CLIP via parameter-efficient prompt tuning, outperforming prior methods by +5.01% mAP across 21 datasets while using less than one-third of training data. GenD (Yermakov et al., 2025) tunes only the Layer Normalization parameters of CLIP ViT-L/14 (0.03% of weights), enforces L2 normalization, and applies uniformity and alignment losses, achieving state-of-the-art average cross-dataset AUROC across 14 benchmarks spanning six years of deepfake evolution. D^3 (Yang et al., 2025b) scales multi-generator training with a parallel discrepancy branch. EFFORT (Yan et al., 2025a) decomposes weight matrices via SVD, freezing principal components and fine-tuning only the residual orthogonal subspace.

Table 3: Detection performance summary (AUC %). FF++ = FaceForensics++ HQ. Cross = FF++ → Celeb-DF v2.

Paradigm	Method	FF++	Cross	Venue
Spatial CNN	XceptionNet	95.7	71.5	ICCV'19
Spatial CNN	Face X-Ray	98.8	74.2	CVPR'20
Temporal	FTCN	98.8	86.9	ICCV'21
Biological	LipForensics	97.1	82.4	CVPR'21
ViT Hybrid	ENet+ViT	95.1	64.8	ICIAP'22
Foundation	CLIPping	96.7	88.1	ICMR'24
Foundation	GenD (CLIP)	96.0	92.8	2025

At the multimodal LLM frontier, several 2025–2026 methods integrate large language models with detection. FakeVLM (Wen et al., 2025) trains on 100,000+ images with natural-language artifact annotations, producing both a classification and a human-readable explanation of the observed artifacts. VIGIL (Li et al., 2026) takes a forensic-inspired approach: it decomposes the face into anatomical regions, independently examines each region for evidence of manipulation, and synthesizes a verdict from the accumulated part-level findings. ForensicZip (Lai et al., 2026) tackles the computational bottleneck of processing high-resolution content through multimodal LLMs, achieving a 2.97× speedup by selectively retaining tokens that carry forensic rather than semantic information.

A companion analysis provides a detailed analysis of why foundation models generalize (testing three competing hypotheses) and where they fail.

Robustness under attack. Foundation models show improved robustness to A1 (new generators) compared to CNN detectors, likely because they exploit semantic rather than artifact-specific features. Their robustness to A2 (adversarial perturbation) has not been systematically evaluated; CLIP is known to be vulnerable to typographic and embedding-space attacks, and VLM-based detectors face additional prompt injection risks. We flag adversarial evaluation of foundation-model detectors as the most urgent open experimental question.

4.6 Cross-Paradigm Summary

Table 3 summarizes representative performance across paradigms.

The trend is clear: cross-dataset performance improves monotonically from spatial CNN methods (71.5%) through temporal (86.9%) and semantic (82.4%) approaches to foundation model methods (92.8%). The generalization gap has narrowed from 28 percentage points (XceptionNet) to 3.2 points (GenD), though the challenge of entirely new generation paradigms remains open.

5 Proactive Defenses: Provenance and Watermarking

We survey three classes of proactive defense, each analyzed for its vulnerability profile under the attack classes defined in §6.

5.1 C2PA Content Provenance

The Coalition for Content Provenance and Authenticity (C2PA, founded February 2021 by Adobe, Arm, BBC, Intel, Microsoft, and Truepic) has published specification v2.2 (May 2025). Content Credentials are cryptographically signed JUMBF structures using X.509 PKI, CBOR encoding, SHA-256 content hashes, and RFC 3161 timestamping. Any tampering with the content breaks the cryptographic signature. The Content Authenticity Initiative (CAI), C2PA's implementation arm, surpassed 5,000 members by mid-2025,

including TikTok, Meta, Google, OpenAI, Cloudflare (∼20% of web traffic), and major news organizations (AP, AFP, Reuters, Washington Post).

C2PA's critical advantage is architecture-agnostic authentication: it makes no assumptions about how content was generated and is not subject to obsolescence as generation methods evolve. This is the property that makes it robust to attack classes A1–A4 in the composition matrix.

TikTok implemented C2PA Content Credentials at scale (May 2024) and removed 51,618 synthetic media videos in H2 2025. Meta labels AI content using C2PA metadata detection, but its Oversight Board ruled in March 2026 that the system "falls short," finding only ∼30% of AI content correctly labeled.

Vulnerability profile. C2PA is robust to attacks A1–A4 (its cryptographic integrity is independent of pixel content, watermark status, and generator type). It is vulnerable to A5 (metadata stripping: platforms that re-encode content strip the JUMBF manifest), A6 (analog hole), and A7 (infrastructure compromise: CA key theft or fraudulent device attestation). The critical adoption gap: most intermediary platforms still strip C2PA metadata during upload.

5.2 Hardware-Anchored Provenance

Qualcomm's Snapdragon (Truepic, 2025) 8 Elite Gen 5 (September 2025) embeds Truepic's Secure Media Library in the mobile Trusted Execution Environment, bringing hardware-certified provenance to billions of devices via Samsung Galaxy S26 and Xiaomi 17. Hardware-origin credentials are structurally superior to software-added ones: they are embedded at capture time within a tamper-resistant boundary before any software manipulation is possible.

5.3 AI Output Watermarking

SynthID (Google DeepMind, 2023) has watermarked over 10 billion items across images (Imagen), video (Veo), audio (Lyria), and text (Gemini). OpenAI adds C2PA metadata to all DALL-E 3 outputs with ∼98% detection accuracy at <0.5% false positive rate. VideoSeal (Meta AI, 2024) (December 2024) provides open-source video watermarking, later expanded to the comprehensive Meta Seal framework covering all modalities. Kirchenbauer et al. (2024) introduced token-level statistical watermarking for LLM text at ICML 2023, modifying logit distributions during generation to embed detectable signatures.

Vulnerability profile. Watermarks are robust to A1 (new generators do not affect embedded marks) and partially robust to A3 (moderate compression). They are vulnerable to A4: Zhao et al. (2024) (NeurIPS 2024) proved that regeneration attacks remove 98% of invisible watermarks while maintaining PSNR >30. LightShed Foerster et al. (2025) (USENIX Security 2025) demonstrated generalizable removal of pixel-level protections from Glaze, Nightshade, and similar tools with 99.98% success in ∼0.014 seconds per image. The NeurIPS 2024 "No Free Lunch" paper Pang et al. (2024) proved fundamental trade-offs between robustness and spoofability: robust watermarks enable "piggyback spoofing attacks" embedding legitimate marks into harmful content. The simplest attack (a screenshot) removes all metadata-based watermarks instantly. These results indicate that watermarking is a necessary but insufficient layer requiring complementary mechanisms.

5.4 Consent Infrastructure

HaveIBeenTrained (Spawning.ai, 2025) processed over 1 billion opt-outs. Nightshade and Glaze poison training data and mask artistic style, but LightShed's circumvention of both tools indicates that perturbation-based protection faces the same arms-race dynamics as detection.

6 The Defense Composition Matrix

The individual defense mechanisms surveyed in Sections 4 and 5 are well understood in isolation. What has not been analyzed is how they interact when deployed as layers of a unified defense stack. We present this composition analysis here.

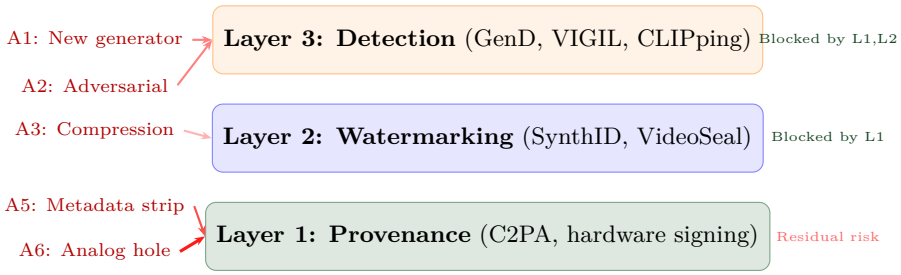

Figure 1: The three-layer defense stack and attack classes. Attacks enter from the left; each layer blocks, degrades, or is bypassed. Only A5 and A6 penetrate all three layers. Table 4 quantifies these interactions.

6.1 Attack Classes

We organize adversarial strategies against AI-synthesized media defenses into six classes, ordered by increasing sophistication:

A1: Generator evasion. Each new generation architecture produces outputs outside the training distribution of existing detectors. The GAN-to-diffusion transition eliminated the spectral fingerprints on which an entire class of detectors relied Ricker et al. (2024); Frank et al. (2020). This is not an adversarial attack in the traditional sense; it is the natural consequence of generator evolution.

A2: Adversarial perturbation. Imperceptible pixel-level modifications that flip detector predictions. Carlini & Farid (2020) reduced a forensic classifier from AUC 0.95 to 0.0005 with white-box attacks and to 0.22 in the black-box setting, perturbing as little as 1% of image area.

A3: Compression and social media laundering. Platforms apply proprietary recompression during upload, stripping high-frequency artifacts that detection methods exploit. XceptionNet accuracy drops from 99.26% (raw) to 81.00% under heavy compression (c40).

A4: Watermark removal and spoofing. Zhao et al. (2024) proved that regeneration attacks remove 98% of invisible watermarks while maintaining PSNR >30. LightShed Foerster et al. (2025) removes pixel-level protections from Glaze and Nightshade with 99.98% success in 0.014 seconds per image. The NeurIPS 2024 "No Free Lunch" paper Pang et al. (2024) proved fundamental trade-offs: robust watermarks enable "piggyback spoofing attacks" embedding legitimate marks into harmful content.

A5: Metadata stripping. A screenshot removes all embedded metadata (C2PA credentials, EXIF data, invisible watermarks encoded in metadata fields) instantly. Platform re-encoding strips C2PA manifests unless the platform explicitly preserves them (as of March 2026, TikTok and Adobe tools do; most others do not).

A6: The analog hole. Re-recording content from a screen with a camera produces a new capture that carries no forensic connection to the original digital content. This is an irreducible attack that no digital defense can fully address, because it operates at the boundary between the digital and physical domains.

A7: Infrastructure compromise. Compromise of a C2PA certificate authority's signing key, a hardware TEE attestation root, or a watermark embedding key allows an adversary to produce fabricated content with valid credentials. This attack class does not exploit weaknesses in the defense *mechanism* but in the *infrastructure* on which the mechanism depends. We include it because the paper's conclusion that provenance is the most resilient layer holds only under the assumption that provenance infrastructure is uncompromised; this assumption deserves explicit treatment.

6.2 The Composition Matrix

Table 4 maps the interaction between the seven attack classes and the three defense layers. We define three operational outcomes: **Blocked (B)**: the defense layer maintains >90% of its laboratory performance against this attack class, based on published evaluations. **Degraded (D)**: performance drops by 15–50% but the

Table 4: Defense Composition Matrix. B = attack blocked by this layer; D = layer degraded but partially effective; X = layer bypassed entirely. Cells highlighted in red indicate that the attack bypasses *all three layers* simultaneously.

Attack Class	**Detection**	**Watermark**	**Provenance**	**Composition**
A1: Generator evasion	X	B	B	2 of 3 block
A2: Adversarial perturb.	X	D	B	1 block, 1 degrade
A3: Compression	D	D	B	1 block, 2 degrade
A4: Watermark removal	D†	X	B	1 block, 1 degrade
A5: Metadata strip	D	X	X	**1 degrade, 2 bypass**
A6: Analog hole	X	X	X	**All 3 bypassed**
A7: Infra. compromise	B	D	X	**Primary layer bypassed**

layer retains partial discriminative value. **Bypassed (X)**: performance drops to near-chance (<60% AUC) or the attack renders the layer structurally inoperative. We assign ratings based on the empirical evidence reviewed in Sections 4–5; specific supporting citations are noted where the assignment is non-obvious.

6.3 Cascade Failure Analysis

The matrix reveals several structural properties of the defense stack.

Provenance is the most resilient layer. It survives attacks A1 through A4, because cryptographic signatures are independent of both pixel content and watermark integrity. Only metadata stripping (A5) and the analog hole (A6) defeat it.

Detection and watermarking have complementary vulnerabilities. Attacks that defeat detection (A1: new generators, A2: adversarial perturbations) do not defeat watermarking, because watermarks are embedded before the attack occurs. Attacks that defeat watermarking (A4: removal) do not defeat detection, because removal artifacts may themselves be detectable. This complementarity means that a two-layer stack (detection + watermarking) is substantially more robust than either layer alone.

Two attack classes defeat the entire stack. Metadata stripping (A5) bypasses watermarks and provenance, leaving only degraded detection. The analog hole (A6) defeats all three layers. These represent the irreducible residual vulnerabilities of any digital defense architecture.

Compression (A3) is a universal degrader. It does not defeat any layer outright but degrades both detection and watermarking. Social media platforms that apply aggressive recompression effectively weaken two of three defense layers simultaneously.

6.4 Emergent Composition Effects

We identify three cases where stacking defenses creates effects that are absent from individual layers.

E1: Authenticated contradictions. When content carries valid C2PA provenance *and* a watermark indicating AI generation, the two signals contradict each other: provenance says "captured by a real camera" while the watermark says "generated by AI." This occurs when a real photograph is post-processed by an AI system that embeds a generation watermark. The "Authenticated Contradictions" paper Nemecek et al. (2026) shows that this conflict is not hypothetical; it arises naturally in editing workflows and creates a trust ambiguity that individual layers do not.

E2: False provenance amplification (hypothesis). If an adversary compromises a C2PA signing key (e.g., through a supply-chain attack on a camera manufacturer), they can produce fabricated content with valid provenance. We hypothesize that in a stacked system where users have learned to trust provenance as the primary signal, the damage from compromised provenance is amplified relative to a system without

provenance, because the presence of credentials may train users to lower their guard. This hypothesis requires empirical validation through user studies comparing trust calibration in provenance-present versus provenance-absent environments; we include it as an open question rather than a demonstrated finding.

E3: Detection-watermark feedback loops (hypothesis). If detection methods use the presence or absence of watermarks as an input feature, an adversary who learns to embed legitimate watermarks into fabricated content (the "piggyback spoofing attack" from Pang et al. (2024)) could exploit this dependency to fool both the watermark verifier and the watermark-aware detector simultaneously. We are not aware of deployed detection systems that explicitly condition on watermark status, but the architectural possibility exists as watermark-aware detection pipelines are proposed. We flag this as a design risk for future systems rather than a demonstrated vulnerability.

7 The Detection Ceiling

We argue that passive detection faces a structural limit that provenance-based approaches do not share. We state this as a conjecture rather than a theorem, because a complete proof would require formalizing the space of all possible generators, which we leave for future work.

Theorem (TV-AUC Bound). For any binary classifier D distinguishing samples from distributions p_G and p_{real} under equal priors:

$$\text{Acc}_{\text{bal}}(D, G) \leq \frac{1}{2} + \frac{1}{2}\,\text{TV}(p_G, p_{\text{real}}) \tag{2}$$

This is a standard result connecting total variation distance to optimal binary classification (the Neyman-Pearson lemma and Scheffé's identity under equal priors).

Conjecture 1 (Detection Ceiling)**.** *As generators improve, $TV(p_G, p_{real}) \to 0$, tightening the bound toward chance (balanced accuracy $= 0.5$). ($AUC = 0.5$). Content provenance is not subject to this bound because it operates on cryptographic metadata, not pixel distributions.*

The theorem is established; the conjecture is the empirical claim that TV distance is decreasing as generators evolve. This follows from the data processing inequality and the relationship between total variation distance and optimal binary classification (Le Cam's two-point method): no classifier, regardless of architecture, can exceed $\frac{1}{2} + \frac{1}{2}$ TV in AUC when distinguishing samples from two distributions.

The conceptual argument that detection difficulty scales with distributional proximity was established for deepfake forensics by Agarwal & Varshney (2019), who derived Neyman-Pearson and Bayesian error bounds for GAN-based synthesis. Our contribution extends this to the three-layer composition framework, contrasting detection with provenance. We note that Equation 2 is a standard result in hypothesis testing, not a novel claim. Our contribution is the observation that this bound has practical implications for the deepfake defense ecosystem that the field has not confronted: as generator quality improves, the total variation distance between generated and real content decreases, and no amount of detector engineering can overcome this information-theoretic limit. The empirical evidence is consistent: cross-dataset AUC has declined from 95%+ (GAN detectors on GAN content) to 47% (GAN detectors on diffusion content), tracking the decrease in TV distance as generators improve.

The distinction between information-theoretic limits and detector-specific limitations is critical. The 47% AUC of GAN-trained detectors on diffusion content may reflect GAN-specific feature overfitting rather than a genuine convergence of p_G and p_{real}. Foundation model detectors achieve 92.8% cross-dataset AUC, suggesting that the current ceiling is architectural, not information-theoretic. However, as generators continue to improve (Sora 2, Veo 3), the TV distance will eventually tighten the bound even for foundation models. The question is not whether the ceiling will be reached, but when.

Supporting evidence. The empirical record is consistent with this conjecture. Detection accuracy has degraded systematically as generators have improved: GAN-trained detectors achieve 95%+ on GAN outputs but 47% on diffusion outputs Yan et al. (2024); Deepfake-Eval-2024 Chandra et al. (2025) documents 45–50% accuracy degradation from laboratory to in-the-wild deployment across 44 hours of real-world content. the scaling laws of Wang et al. (2025) show that detection error follows a power law $1 - \text{AUC} = A \cdot N^{-\alpha}$ with

respect to training diversity, but with no guarantee that the law extends to qualitatively new generation paradigms.

Why provenance escapes the ceiling. Provenance operates on a fundamentally different signal. A C2PA credential is a cryptographic assertion about the content's origin ("this was captured by device X at time T") that is independent of the content's statistical properties. Whether the content is a natural photograph or a perfect synthetic replica, the credential either exists and verifies or it does not. The detection ceiling arises from the convergence of generator output distributions to the real data distribution; provenance is unaffected by this convergence because it does not depend on the distributions at all.

Limitations of the conjecture. We emphasize two caveats. First, the conjecture applies to detectors operating on pixel content alone; detectors that incorporate metadata, provenance, or out-of-band signals are not constrained. Second, the conjecture describes an asymptotic limit, not a statement about current systems. Current generators remain far from perfectly matching the real data distribution, and detection is both useful and effective against the generators deployed today. The conjecture's practical implication is about investment priorities: as generators continue to improve, the returns to detection research will diminish while the returns to provenance infrastructure will not.

8 Benchmarks and Evaluation

The empirical foundation of detection research rests on several benchmark datasets with distinct properties. FaceForensics++ (Rössler et al., 2019) provides 1,000 videos manipulated via four methods at three compression levels; XceptionNet accuracy ranges from 99.26% (raw) to 81.00% (c40). Celeb-DF v2 (Li et al., 2020b) provides substantially higher visual quality; most detectors drop to 50–65% AUC, establishing it as the standard cross-dataset test. DFDC (Dolhansky et al., 2020) (Meta, $10M) provides 128,154 clips from 3,426 actors; the competition winner achieved only 65.18% on the black-box test. DF40 (NeurIPS 2024) covers 40 deepfake approaches including diffusion models, addressing the critical gap that pre-2023 datasets contain no diffusion-model deepfakes. Deepfake-Eval-2024 Chandra et al. (2025) provides the first large-scale in-the-wild benchmark: 44 hours of video, 56.5 hours of audio, and 1,975 images from 88 websites in 52 languages, documenting 45–50% accuracy degradation for deployed systems.

Yan et al. (2023) introduced DeepfakeBench at NeurIPS 2023, unifying 15 detection methods across 5 datasets in a standardized pipeline, addressing the critical problem of inconsistent evaluation protocols that made cross-method comparison unreliable. The AI-Face benchmark (Yang et al., 2025a) provides the first million-scale demographically annotated dataset across 37 generation methods, enabling systematic fairness evaluation. VIGIL's OmniFake benchmark (2026) introduces a hierarchical 5-level evaluation from in-domain to in-the-wild social media data, progressively tested up to content from the latest generators (Nano Banana, Veo 3, Sora 2).

A critical gap persists: no benchmark simultaneously covers video, audio, text, and multimodal synthetic content at production quality, creating a blind spot for evaluating the comprehensive detection pipelines needed for the composed defense stack described in §6.

9 The Generalization Crisis

The composition analysis in §6 is motivated by the systematic failure of detection as a standalone defense. We summarize the six structural limitations.

Cross-dataset collapse. XceptionNet drops from 99.26% on FF++ to 71.5% on Celeb-DF (28 pp). Deepfake-Eval-2024 Chandra et al. (2025) documents 45–50% degradation across 44 hours of in-the-wild content. Many off-the-shelf models produce AUC near 0.50.

Adversarial attacks. Carlini & Farid (2020) reduced a 0.95-AUC detector to 0.0005 (white-box) and 0.22 (black-box) with imperceptible perturbations altering as little as 1% of image area.

Compression. Social media platforms apply proprietary recompression that strips the high-frequency artifacts detection methods exploit. A deepfake detectable on one platform may evade detection after reprocessing by another.

The diffusion barrier. GAN-trained frequency detectors suffer approximately 15.2 pp AUROC drops on diffusion outputs Ricker et al. (2024). DF40 found detectors at 85% AUC on face-swap datasets dropping to 47% on Stable Diffusion outputs Yan et al. (2024).

Demographic bias. Lin et al. (2024) documented maximum FPR gaps of 20.6 across intersectional subgroups. Ju et al. (2024) found Black men misclassified as deepfake at 39.1% versus white women at 15.6%.

Human performance. Diel et al.'s meta-analysis of 56 papers (86,155 participants) found average accuracy of 55.54% (95% CI: 48.87–62.10%), not significantly above chance. Without being informed a deepfake might be present, detection

10 Societal Context

10.1 Scale of Harm

The scale of harm is now documented with specificity. Deepfake-enabled fraud drained an estimated $1.1 billion in 2025, with cumulative losses reaching $2.19 billion as of April 2026 Surfshark (2026). Over 80% of losses occurred on social media, tripling from the prior year. Deloitte (2024) projects AI-facilitated fraud losses reaching $40 billion by 2027 at a 32% compound annual growth rate. The Arup incident (World Economic Forum, 2025) (January 2024) saw a finance worker authorize $25.6M in transfers to deepfake participants in a multi-person video call. Non-consensual intimate imagery constitutes 96–98% of deepfake videos DeepStrike (2025), targeting women in over 99% of cases. Graphika (2023) reported 24 million unique monthly visitors to 34 "nudify" service websites in September 2023. Recorded Future (2024) documented 82 political deepfakes across 38 countries from July 2023 to July 2024.

The Deepfake-as-a-Service economy. A qualitatively new dimension of the threat is the emergence of DFaaS, a commoditized criminal market paralleling Ransomware-as-a-Service. Group-IB (2025) collected over 300 Telegram and dark web posts advertising DFaaS tooling between 2022 and September 2025. A synthetic identity kit (AI-generated face, cloned voice, documentation) sells for approximately $5; Dark LLM subscriptions for social engineering scripts cost ∼$30/month; and deepfake commissions range from $10 to $50. Because production costs have collapsed below $5 per synthetic identity, only systemic defenses (the composed stack described in §6) can match the threat's scale.

The agentic AI multiplier. Agentic AI can orchestrate multi-step fraud campaigns: generating synthetic identities, cloning executive voices, initiating video calls with deepfake participants, and executing financial transfers without sustained human involvement. Group-IB's 2025 threat intelligence found that while fully autonomous AI-driven cybercrime has not yet materialized at scale, hybrid human-AI operations are already reshaping fraud pipelines.

10.2 The Detection Paradox

A second-order harm is the *liar's dividend*: the ability of any actor to dismiss authentic evidence as AI-generated, exploiting the mere existence of synthesis capability. We have formalized this elsewhere, showing through a game-theoretic model that improving detector accuracy can paradoxically increase the payoff from false accusations of fabrication. The mechanism is the interaction between detector accuracy and challenge credibility: a more accurate detector makes the claim "this could be a deepfake" more persuasive. Provenance shifts this equilibrium by making the "claim it's fake" strategy dominated when cryptographic credentials exist.

10.3 Regulatory Context

The EU AI Act (European Parliament, 2024) (August 2024) requires AI output disclosure and machine-readable marking (Article 50, enforceable August 2026). The TAKE IT DOWN Act (Skadden, 2025)

(May 2025) criminalizes non-consensual intimate deepfakes in the U.S. China's Deep Synthesis Provisions (CAC, 2023) (January 2023) mandate visible labeling and real-identity verification. 47 U.S. states have enacted deepfake legislation, with 174 total laws (82% passed in 2024–2025). The regulatory trajectory supports the composition thesis: multiple layers of technical defense are necessary because no single regulatory framework achieves global coverage.

Dual-use acknowledgment. We acknowledge that the Defense Composition Matrix and cascade failure analysis in §6 could be read as an instructional blueprint for bypassing current defense systems. We have deliberately focused on architectural patterns rather than implementation-specific attack code, and the vulnerabilities we document (metadata stripping, the analog hole, watermark removal) are already well-known in the adversarial ML and security communities. We believe the benefit of systematizing these failure modes for defense designers outweighs the marginal information gain for adversaries, who already exploit these weaknesses. The composition thesis itself is a defense-forward contribution: it argues for building layered systems that cover each other's blind spots, which is actionable only for defenders.

11 Open Problems

Our analysis identifies eight open problems, each stated as a falsifiable hypothesis.

OP1: Architecture-agnostic detection. *Hypothesis:* Foundation model representations encode generator-invariant facial properties that enable detection across all current architectures. *Protocol:* Evaluate GenD and CLIPping on outputs from five generators released after their training data was collected. If cross-generator AUROC exceeds 85%, the hypothesis is supported.

OP2: Composition optimality. *Hypothesis:* The optimal ordering of defense layers is provenance → watermark → detection, because provenance handles the broadest attack set and detection serves as backstop. *Protocol:* Deploy three orderings on a platform and measure false acceptance rates under the seven attack classes.

OP3: Watermark-provenance integration. *Hypothesis:* Integrating watermark status into the C2PA manifest (rather than treating them as independent systems) eliminates the "authenticated contradiction" vulnerability (E1). *Protocol:* Implement integrated and independent versions; measure user trust in the contradiction scenario.

OP4: Calibrated detection for legal proceedings. *Hypothesis:* Foundation model detectors with metric learning objectives produce better-calibrated confidence estimates than cross-entropy-trained models. *Protocol:* Evaluate expected calibration error across methods in the Adaptation Efficiency Frontier.

OP5: Real-time detection at platform scale. *Hypothesis:* Knowledge distillation from GenD-scale models to mobile-deployable architectures preserves >90% of cross-dataset AUROC at >100× throughput. *Protocol:* Distill and benchmark on the DFDC test set under latency and throughput constraints.

OP6: Cross-modal joint detection. *Hypothesis:* Audio-visual joint detection using foundation models outperforms modality-specific detectors because cross-modal inconsistencies provide a generator-invariant signal. *Protocol:* Evaluate on AVFakeBench and CharadesDF.

OP7: Closing the analog hole. *Hypothesis:* Imperceptible screen-embedded signals (e.g., modulated backlight patterns) can survive re-recording with a camera, providing partial defense against A6. *Protocol:* Embed and recover signals across five re-recording conditions.

OP8: Demographic fairness in composed systems. *Hypothesis:* Composing detection with provenance reduces demographic bias because provenance decisions are identity-independent. *Protocol:* Evaluate FPR disparity across demographic subgroups for detection-only versus detection+provenance systems.

12 Deployment Guide

The composition analysis yields a practical deployment decision tree for system designers:

1. **Does your platform control the capture device?** If yes: deploy hardware-anchored C2PA signing (Layer 1). This blocks A1–A4 at the point of capture. If no: proceed to step 2.

2. **Does your platform generate AI content?** If yes: embed watermarks (SynthID, VideoSeal) in all outputs (Layer 2). This addresses A1 and partially A3. Also embed C2PA manifests marking the content as AI-generated.

3. **Can you preserve metadata through your pipeline?** If yes: implement C2PA manifest preservation during upload, transcoding, and delivery. This is the single highest-leverage engineering investment. If no: metadata stripping (A5) will defeat Layers 1 and 2; detection (Layer 3) becomes your primary defense.

4. **Deploy detection as backstop.** Use foundation-model detectors (e.g., GenD at 0.03% parameter tuning) for content lacking provenance or watermarks. Set asymmetric thresholds based on the deployment context: FPR $< 0.001\%$ for high-stakes forensic analysis; FPR $\approx 1\%$ for social media triage.

5. **Accept residual risk from A5 and A6.** Metadata stripping and the analog hole cannot be eliminated by digital defenses alone. Invest in media literacy and pre-bunking strategies for these residual cases.

13 Conclusion

The central finding of this survey is that no single defense mechanism can address the AI-synthesized media challenge at the scale and sophistication it has reached. The Defense Composition Matrix reveals that of seven attack classes, only two (metadata stripping and the analog hole) defeat all three defense layers simultaneously, while a third (infrastructure compromise) bypasses the primary trust layer. The remaining four are blocked by at least one layer, meaning that a properly composed stack provides substantially greater resilience than any individual component.

Provenance emerges as the most resilient layer: it survives attacks A1 through A4, because cryptographic signatures are independent of pixel content. Detection, despite its dramatic improvements from foundation model approaches (cross-dataset AUROC improving from 71.5% in 2019 to 92.8% in 2025), faces a structural ceiling as generators approach the real data distribution. Watermarking occupies a complementary position, catching content that evades detection but vulnerable to removal attacks that detection can identify.

The practical implication is architectural. A researcher designing a platform's content authenticity system should use the Defense Composition Matrix to select layers that cover each other's blind spots. The eight open problems we identify, each with a falsifiable hypothesis and proposed experimental protocol, chart the path toward a defense stack whose residual vulnerabilities are minimized.

The window for building this infrastructure is finite. As AI-generated content grows from a minority to a potentially dominant fraction of online media, establishing provenance before that inversion is a categorically different engineering challenge than attempting to authenticate the majority of internet content retroactively.

References

Sakshi Agarwal and Lav R. Varshney. Limits of deepfake detection: A robust estimation viewpoint. *arXiv:1905.03493*, 2019.

James Betker et al. Improving image generation with better captions. *OpenAI Technical Report*, 2023.

CAC. Deep synthesis provisions. 2023.

Nicholas Carlini and Hany Farid. Evading deepfake-image detectors. *CVPR Workshops*, 2020.

R. Chandra et al. Deepfake-eval-2024. *arXiv preprint arXiv:2503.02857*, 2025.

Kai Chen et al. VALL-E 2: Neural codec language models are human parity zero-shot TTS synthesizers. *arXiv:2406.05370*, 2024.

Renwang Chen et al. SimSwap: An efficient framework for high fidelity face swapping. *Proc. ACM MM*, 2020.

Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE JSTSP*, 14(5):1024–1037, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. CVPR*, 2017.

Umur Aybars Ciftci et al. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE TPAMI*, 2020.

Davide Alessandro Coccomini et al. Combining EfficientNet and vision transformers for video deepfake detection. In *Proc. ICIAP*, 2022.

Florinel-Alin Croitoru et al. Deepfake media in the generative AI era. *arXiv:2411.19537*, 2024.

DeepStrike. Deepfake statistics 2025. *DeepStrike Research*, 2025.

Deloitte. Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. 2024.

Jingyi Deng et al. Defenses against AI-generated visual media. *ACM Computing Surveys*, 2025.

Brian Dolhansky et al. The DeepFake Detection Challenge dataset. *arXiv:2006.07397*, 2020.

European Parliament. Regulation (EU) 2024/1689 (AI act). 2024.

M. Foerster et al. Lightshed. *USENIX Security*, 2025.

Joel Frank et al. Leveraging frequency analysis. *ICML*, 2020.

Ian J. Goodfellow et al. Generative adversarial nets. In *NeurIPS*, 2014.

Google DeepMind. SynthID: Identifying AI-generated content. 2023.

Graphika. A revealing picture: Tracking the growth of AI nudification. 2023.

Group-IB. From deepfakes to dark LLMs: 5 use-cases of how AI is powering cybercrime. 2025.

Alexandros Haliassos et al. Lips don't lie. In *Proc. CVPR*, 2021.

Chao Han et al. Facial component guided adaptation for foundation model. *Proc. CVPR*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Yan Ju et al. Improving fairness in deepfake detection. In *Proc. WACV*, 2024.

Tero Karras et al. Progressive growing of GANs. In *ICLR*, 2018.

Tero Karras et al. A style-based generator architecture for GANs. In *Proc. CVPR*, 2019.

Tero Karras et al. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

Tero Karras et al. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

Bernhard Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG (SIGGRAPH)*, 2023.

Syed Awais Khan and Duc-Tien Dang-Nguyen. CLIPping the Deception. *Proc. ICMR*, 2024.

John Kirchenbauer et al. A watermark for large language models. *Proc. ICLR*, 2024.

Yingxin Lai et al. ForensicZip. *arXiv:2603.12208*, 2026.

Lingzhi Li et al. Face X-Ray for more general face forgery detection. In *Proc. CVPR*, 2020a.

Xinghan Li et al. VIGIL: Part-grounded structured reasoning. *arXiv:2603.21526*, 2026.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. *IEEE WIFS*, 2018.

Yuezun Li et al. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proc. CVPR*, 2020b.

Zhiyuan Lin et al. Preserving fairness generalization in deepfake detection. In *Proc. CVPR*, 2024.

Meta AI. Video seal. 2024.

Ben Mildenhall et al. NeRF: Representing scenes as neural radiance fields. In *Proc. ECCV*, 2020.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 2021.

Alexander Nemecek, Hengzhi He, Guang Cheng, and Erman Ayday. Authenticated contradictions from desynchronized provenance and watermarking. *arXiv:2603.02378*, 2026.

Utkarsh Ojha et al. Towards universal fake image detectors. In *Proc. CVPR*, 2023.

OpenAI. Video generation models as world simulators. 2024.

W. Pang et al. No free lunch in LLM watermarking. *NeurIPS*, 2024.

Gan Pei et al. Deepfake generation and detection: Benchmark and survey. *ACM Computing Surveys*, 2024.

Ivan Petrov et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 2023.

K. R. Prajwal et al. A lip sync expert is all you need. In *Proc. ACM MM*, 2020.

Recorded Future. Political deepfakes in global elections. 2024.

Jonas Ricker et al. Towards detection of diffusion model deepfakes. *VISAPP*, 2024.

Andreas Rössler et al. FaceForensics++. In *Proc. ICCV*, 2019.

Aliaksandr Siarohin et al. First order motion model for image animation. In *NeurIPS*, 2019.

Skadden. TAKE IT DOWN act. 2025.

Spawning.ai. Creator consent layer for AI. 2025.

Surfshark. AI drives deepfake losses to $1.1 billion. *Surfshark Research*, 2026.

The Recursive. ElevenLabs raises $500m at $11b valuation. 2026.

Justus Thies et al. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proc. CVPR*, 2016.

Truepic. Qualcomm embeds truepic in snapdragon 8 elite. 2025.

Chengyi Wang et al. Neural codec language models are zero-shot TTS synthesizers. *arXiv:2301.02111*, 2023.

Sheng-Yu Wang et al. CNN-generated images are surprisingly easy to spot. *Proc. CVPR*, 2020.

Wenhao Wang et al. Scaling laws for deepfake detection. *arXiv:2510.16320*, 2025.

Siwei Wen et al. Spot the fake: FakeVLM. *Proc. NeurIPS*, 2025.

World Economic Forum. Cybercrime: Lessons from a $25m deepfake attack. 2025.

Zhiyuan Yan et al. DeepfakeBench: A comprehensive benchmark for deepfake detection. *NeurIPS D&B*, 2023.

Zhiyuan Yan et al. DF40: Next-generation deepfake detection. *NeurIPS D&B*, 2024.

Zhiyuan Yan et al. EFFORT: Orthogonal subspace decomposition. *Proc. ICML*, 2025a.

Zhiyuan Yan et al. Spatiotemporal adapter tuning. *Proc. CVPR*, 2025b.

Haoyuan Yang et al. AI-Face: Million-scale demographically annotated AI-generated face dataset. *Proc. CVPR*, 2025a.

Haoyuan Yang et al. D^3: Scaling up deepfake detection. In *Proc. CVPR*, 2025b.

Andrii Yermakov et al. Deepfake detection that generalizes across benchmarks. *arXiv:2508.06248*, 2025.

Hanqing Zhao et al. Multi-attentional deepfake detection. *Proc. CVPR*, 2021.

Xuandong Zhao et al. Regeneration attacks remove invisible watermarks. *NeurIPS*, 2024.

Xuandong Zhao et al. SoK: Watermarking for AI-generated content. *IEEE S&P*, 2025.

Yinglin Zheng et al. Exploring temporal coherence. In *Proc. ICCV*, 2021.

Yueying Zou, Peipei Li, et al. Survey on AI-generated media detection: From non-MLLM to MLLM. *arXiv:2502.05240*, 2025.