Preprints.org

Article

Not peer-reviewed version

Deepnude Forensics: Detection, Attribution, and Defense Against AI-Generated Intimate Image Abuse

Joas Antonio dos Santos *

Posted Date: 11 November 2026

Keywords: deep nude detection; deepfake detection; synthetic CSAM; deepnude; neural networks; deep learning; diffusion models; digital forensics; generative model attribution; model parsing; reverse engineering; nudify services; adversarial immunization; provenance; C2PA; SynthID; PhotoDNA; federated learning; law enforcement

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Deepnude Forensics: Detection, Attribution, and Defense Against AI-Generated Intimate Image Abuse

Joas Antonio dos Santos

Independent Security Research, Brazil; joas.santos@redteamleaders.com

Abstract

The commoditization of open-weight generative image models has collapsed the technical and economic barriers to producing non-consensual intimate imagery ("deepnudes") and, in its most severe form, synthetic Child Sexual Abuse Material (CSAM). Unlike photographic abuse material, synthetic content is unbounded in volume, detaches from any specific victim-identification signal, and systematically defeats the perceptual-hash infrastructure (PhotoDNA, Project Arachnid, INTERPOL ICSE) on which global child-protection operations depend. Building on our prior work on machine-learning deep nude detection, this paper first develops the theoretical foundations of deep nude detection techniques—decomposing the task into explicitness, manipulation, and subject-attribute factors with a decision- and information-theoretic account of multi-cue fusion—and then presents an end-to-end, defense-oriented framework with cybersecurity at its core that (i) formally characterizes the open "uncensored" generation ecosystem and the model-level mechanisms by which safety alignment is stripped, treated strictly as a threat model rather than a recipe; (ii) formalizes the reverse engineering of generative pipelines for attribution—tracing a synthetic artifact back to its originating model family through diffusion reconstruction error, latent and weight-space fingerprints, spectral forensics, membership inference, and foundation-model feature forensics; (iii) specifies a multi-stream detection architecture, with algorithms, designed to be trained without possessing illegal material, using a curated-benign public track with suspect material confined to accredited enclaves, benign adult/synthetic proxies, and a privacy-preserving federated protocol; (iv) analyzes preemptive defenses—adversarial immunization (PhotoGuard, Glaze, Nightshade), data poisoning, and provenance watermarking (C2PA, SynthID, Stable Signature)—together with their documented robustness failures under purification and adaptive attack; and (v) maps the whole to operational law-enforcement workflows (INTERPOL, NCMEC, Canadian Centre for Child Protection), to a multi-jurisdiction legal analysis, and to forensic-admissibility and chain-of-custody constraints. We contribute a layered risk taxonomy, a reproducible evaluation protocol and benchmark design with explicit ethical controls, a cybersecurity-hardening checklist for the model supply chain, an illustrative (benign) end-to-end walkthrough, and a prioritized research and policy agenda. Our central thesis is that durable defense requires shifting from post hoc content matching toward generative-pipeline forensics: detecting, attributing, and raising the cost of synthesis at the model and provenance layers. We position this work as a critical synthesis, a proposed reference architecture, and an evaluation protocol rather than an empirically validated system: the figures are illustrative of the protocol, not measured results, and we mark throughout what is established in prior work, what is proposed here, and what remains to be validated—particularly for the hashing, attribution, and privacy guarantees, whose scope we state conservatively.

Keywords: deep nude detection; deepfake detection; synthetic CSAM; deepnude; neural networks; deep learning; diffusion models; digital forensics; generative model attribution; model parsing; reverse engineering; nudify services; adversarial immunization; provenance; C2PA; SynthID; PhotoDNA; federated learning; law enforcement

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

1. Introduction

Generative image models have, within roughly three years, moved from closed research artifacts to freely downloadable open-weight checkpoints that run on consumer GPUs. The same democratization that enables creative, medical, and scientific applications has produced a parallel abuse economy. Two abuse classes dominate this paper. The first is *non-consensual intimate imagery* (NCII), colloquially "deepnudes," in which a clothed photograph of a real person—frequently a woman, a minor, or a public figure—is edited into a nude or sexual depiction. The second is *synthetic CSAM*, in which models generate sexual depictions of children with no corresponding photographic victim, or re-victimize known survivors by extrapolating from existing abuse imagery. These classes overlap: a deepnude produced from a photograph of a minor is simultaneously NCII and CSAM, and is treated as the latter under most legal systems.

1.1. The Structural Break in Child Protection

For two decades, scaled interdiction has relied on *perceptual hashing*. A victim-identified image is fingerprinted once (e.g., Microsoft PhotoDNA), and the robust hash is matched across platforms and against curated corpora such as INTERPOL's International Child Sexual Exploitation (ICSE) database and the Canadian Centre for Child Protection's Project Arachnid [6,7]. PhotoDNA computes a signature that survives resizing and minor edits, enabling matches far beyond exact cryptographic hashing. This architecture is, however, intrinsically *retrospective*: it can only match what has been seen, verified, and hashed before.

Synthetic generation breaks every assumption of that architecture. It produces an effectively infinite stream of novel pixels, each with a novel hash, each outside every known-bad corpus. The volume is unbounded and the marginal cost approaches zero. The Internet Watch Foundation and multiple national centers have reported the operational reality of this shift, with investigators triaging unprecedented volumes of AI-generated material intermixed with photographic abuse [8]. The National Center for Missing & Exploited Children's CyberTipline has reported a corresponding surge in generative-AI-related submissions [9].

Compounding this, Thiel et al. at the Stanford Internet Observatory demonstrated that LAION-5B—the web-scraped corpus behind widely distributed text-to-image models—itself contained validated CSAM, so the capability to synthesize such content is partly *baked into* pretrained weights rather than requiring deliberate fine-tuning [2]. The dataset was withdrawn "out of an abundance of caution," but derived checkpoints remain in circulation. Carlini et al. separately showed that diffusion models *memorize* and can regurgitate training images [36], raising the direct prospect of re-victimization through reproduction of a known survivor's imagery.

1.2. From Content Matching to Pipeline Forensics

This paper extends our earlier study on detecting deepnude imagery with machine learning [1] from a narrow classifier into a full *generative-pipeline forensics* program. The governing intuition is that, while a synthetic artifact's pixels are novel, the *process* that produced it leaves systematic, recoverable traces: spectral signatures from up-sampling, reconstruction behaviour under diffusion inversion, latent and weight-space fingerprints, and (where cooperating generators are used) provenance marks. Forensics should therefore target the pipeline, not merely the pixels.

This reframing also changes the question an investigator asks. Hash matching answers "have I seen this exact image before?"; pipeline forensics answers "was this generated, by what, and who is implicated?"—the questions that actually drive triage, de-confliction, and prosecution when every artifact is novel. It is the same conceptual move that camera forensics made decades ago, from matching known photographs to fingerprinting the device that produced an unknown one (§6). We develop the theory of that move for the generative setting and connect it, end to end, to how child-protection operations actually run.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Our framing is deliberately asymmetric and defensive. We treat the abuse pipeline as an adversary to be *characterized, attributed, detected, and cost-inflated*, and we are explicit throughout about what we withhold. We never provide operational means to produce illegal content.

1.3. Contributions

1. **Formal problem statement and risk taxonomy** of the uncensored ecosystem: a layered model of how open model hubs, fine-tunes, adapters, and inference tooling lower abuse cost, framed as a threat model (Sections 2, 5).
2. **Reverse engineering for attribution**: a formal treatment of tracing synthetic artifacts to originating model families via diffusion reconstruction error, spectral/latent/weight fingerprints, membership inference, and foundation-model feature forensics, with algorithms (Section 6).
3. **A no-possession detection architecture**: a multi-stream detector and training methodology, with pseudocode and a differentially private federated protocol, that never requires the investigator to hold illegal material (Section 8).
4. **Preemptive-defense analysis**: a critical, math-level review of adversarial immunization, poisoning, and provenance watermarking, drawing on robustness failures *documented in the cited literature* (not measured here) under purification and adaptive attack (Sections 12–13).
5. **Operational, legal, and policy integration**: mapping to INTERPOL/NCMEC/C3P workflows, a multi-jurisdiction legal table, forensic admissibility, a cybersecurity-hardening checklist, an illustrative benign walkthrough, and a prioritized agenda for authorities (Sections 14–22).
6. **Theoretical foundations and a research methodology**: a decomposition of deep nude detection into explicitness/manipulation/subject factors, a decision- and information-theoretic account of why calibrated multi-cue fusion is necessary and why the signal shrinks as generators improve, and a reusable research scaffold—research questions, validity analysis, benchmark governance, and a capability-maturity model (Sections 7, 20).

1.4. Scope, Ethics, and Responsible Disclosure

This work is defensive. We do *not* name specific abuse-enabling checkpoints, provide prompts, fine-tuning recipes, filter-bypass procedures, or any instruction whose primary utility is producing illegal content. Where we discuss safety-removal mechanisms, we do so at the level of *category and consequence*, as is standard in the red-teaming literature [35], solely to justify detector and policy design. All empirical protocols in Section 19 are specified so that no researcher must possess CSAM: evaluation uses (a) adult, consensual, or fully synthetic benign imagery as structural proxies, and (b) government-operated enclaves where lawful. Handling of real material is restricted to accredited entities (NCMEC, INTERPOL National Central Bureaus, national units) under their legal authorities. The paper's organization is summarized in Fig. 1.

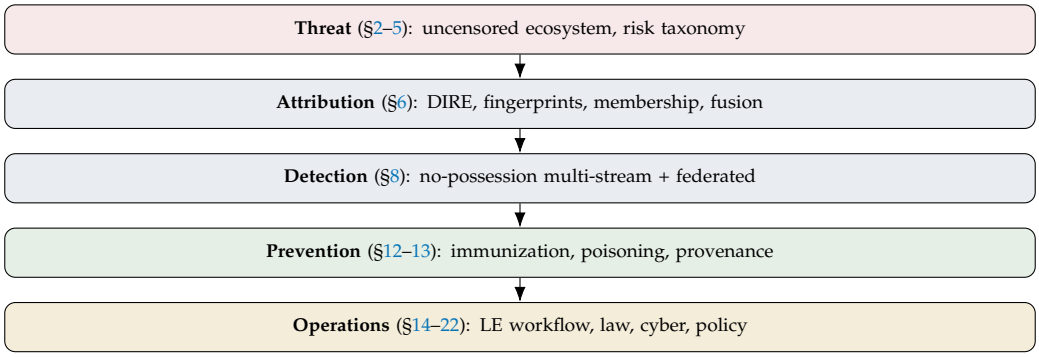


Figure 1. Paper structure as a defense-in-depth stack: characterize the threat, attribute and detect artifacts, prevent synthesis and misuse, and integrate into operations, law, and policy.

2. Formal Problem Statement

We fix notation used throughout.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Definition 1 (Generative artifact). *A generative artifact is an image $x \in [0,1]^{H\times W\times 3}$ produced by a sampling process $x = \mathcal{G}_m(z, c)$, where $\mathcal{G}_m$ is a generator with parameters indexing a* model family *$m \in \mathcal{M}$ (base model, fine-tune, or adapter composition), z a latent seed, and c a conditioning signal (text prompt, mask, reference image).*

Definition 2 (Forensic tasks). *Given an artifact x, we define three tasks: (1)* Detection *$D(x) \in \{\mathcal{R}, \mathcal{G}\}$: is x a real photograph or model-generated? (2)* Attribution *$A(x) \in \mathcal{M} \cup \{\varnothing\}$: which model family produced x, or "unknown" ($\varnothing$) under open-set conditions. (3)* Risk classification *$C(x) \in [0,1]$: probability that x depicts sexualized content involving a minor (the contraband region), estimated* only *within accredited enclaves (Section 8).*

Definition 3 (Laundering channel). *A laundering channel T is a stochastic transformation (JPEG at quality q, resize, crop, blur, screenshot re-capture, diffusion purification [33]) applied to evade detection/attribution. A forensic method is T-*robust *if its decision is stable in distribution over T.*

Property 1 (Hash-stack limitation). *Let K be the known-bad corpus, $\mathrm{ph}(\cdot)$ a perceptual hash, and τ a fixed matching radius. Retrospective matching flags x iff $\min_{k\in K} \mathrm{dist}(\mathrm{ph}(x), \mathrm{ph}(k)) \le \tau$. By construction this decision depends only on* proximity to previously seen material*: a novel artifact $x = \mathcal{G}_m(z, c)$ not perceptually within τ of any $k \in K$ is not flagged, regardless of whether it is harmful. We make no claim that fresh synthesis is* always *far from K—memorization, re-victimizing conditioning, and near-duplicate generation are explicit counterexamples (§6)—only the converse:* a match to known content certifies nothing about the harmfulness of unmatched content. *Hence coverage of* novel *harmful synthesis cannot come from membership in K and must be sought in D, A, C.*

This limitation motivates the program: because known-content matching neither detects novel synthesis nor certifies the remainder as safe, the defensive signal must come from the *generative process*, captured by D and A, and from *provenance*—with K retained only for the retrospective re-share case. This same asymmetry (a non-match is not a safety certificate) drives the data-handling design of §8.

Table 1 contrasts photographic and synthetic abuse material along the dimensions that determine which defense applies.

Table 1. Why the retrospective hashing stack degrades against synthesis.

Property	Photographic CSAM	Synthetic CSAM
Known-bad corpus	Finite, curated	Unbounded, novel
Perceptual hash hit	High (re-shares)	≈ 0 (new pixels)
Victim ID signal	Present	Absent / spoofed
Marginal cost	High	Near zero
Attribution target	Camera / source	Model family / user
Primary defense	Hash match	Pipeline forensics

3. Background and Threat Model

3.1. Generative Pipelines

We summarize the three pipeline families that matter for forensics, with the minimal mathematics needed later.

GANs. A generator G and discriminator D play a minimax game [10],

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Early deepnude tools (the original "DeepNude" application and successors) were conditional image-to-image GANs performing clothed→nude translation. GANs leave strong, model-specific spectral fingerprints from their transposed-convolution up-samplers [18,22].

Diffusion / latent diffusion. Denoising diffusion probabilistic models learn to reverse a Gaussian noising process [11]. The forward process is

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\big), \tag{2}$$

and a network ϵ_θ is trained with the simplified objective

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t}\big[\, \|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 \,\big]. \tag{3}$$

Latent diffusion (Stable Diffusion) runs the process in a VAE latent space $E(\cdot)$ for efficiency [12].

Score-based view. The discrete chain has a continuous-time limit as a stochastic differential equation whose reverse-time solution is governed by the score $\nabla_x \log p_t(x)$, which the network ϵ_θ estimates up to scale. Sampling steers this trajectory with *classifier-free guidance*, interpolating conditional and unconditional score estimates,

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \varnothing) + w\,\big(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)\big), \tag{4}$$

where the guidance weight w amplifies adherence to the conditioning c. This matters forensically for two reasons: the guidance scale leaves a recoverable statistical imprint (over-guided samples are over-saturated and over-smoothed), and the unconditional branch is where "safety" suppression, when present, is weakest.

Manipulation operators. The abuse-relevant capability is not generation *ex nihilo* but a small set of conditional operators applied to a real photograph: inpainting $x' = \mathrm{Inpaint}(x, M, c)$ regenerates a masked region while preserving surrounding identity cues; image-to-image $x' = \mathrm{I2I}(x, c, s)$ re-renders the whole image at strength s; and identity transfer grafts a face (§3.2). Inpainting is the dominant modern vector for both NCII and synthetic CSAM precisely because it minimizes the edited footprint—and therefore the detectable artifact—while maximizing perceived authenticity, the exact property the localization theory of §7.5 exploits in reverse.

Personalization / adapters. Low-Rank Adaptation (LoRA) [13] and DreamBooth [14] adapt a base model to a specific identity or style from a handful of images, enabling targeted NCII against a named victim at near-zero cost. Forensically, adapters are both a threat vector and an *attribution handle*: a LoRA imprints a recoverable signature.

3.2. Modern Deepfake and Personalization Advances

The threat surface is not static; generation quality, speed, and controllability have advanced sharply, and each advance degrades a detection assumption. We summarize the axes that matter forensically, strictly as a threat characterization.

Face swapping and reenactment. Identity-transfer pipelines graft a source identity onto a target image or video. Subject-agnostic one-shot swappers (e.g., the SimSwap family [43]) and widely redistributed inswapper-style models produce stable swaps without per-identity training. Reenactment—driving a still portrait with another person's motion—originated with first-order motion models [44] and now reaches real-time, expression-faithful animation. Audio-driven talking-head synthesis (Wav2Lip lip-sync [45]) adds a voice-conditioned video vector. Forensic consequence: swaps and reenactments blend *real* background pixels with synthetic facial regions, so whole-image synthesis detectors underperform—the global synthesis prior is diluted by genuine pixels—and *localized* blending-boundary analysis becomes necessary (Section 10). Crucially for child protection, a swap can place a *minor's* face on an adult body (or the reverse), decoupling the age cue from the body cue and defeating any single-signal age classifier.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Identity-preserving adapters. Zero-shot identity injection—IP-Adapter [47], InstantID [48]—transfers a face from one or a few reference images into a fresh diffusion generation *without* fine-tuning, via a decoupled cross-attention branch that conditions on a face embedding, collapsing the cost of victim-specific NCII from a LoRA-training step to a single forward pass. This weakens LoRA-weight attribution (Section 6.6) because no persistent adapter is shipped; the signature migrates into the conditioning path and the cross-attention statistics instead.

Fast samplers and new objectives. Few-step samplers—adversarial diffusion distillation [49] and consistency distillation—cut sampling to one–four steps, and rectified-flow / flow-matching backbones [50] replace the curved DDPM denoising trajectory with near-straight transport. Both shift the *spectral and reconstruction* fingerprints that DIRE-class and frequency detectors rely on: a one-step generator barely touches the iterative denoiser that DIRE inverts through, and flow matching changes the residual noise statistics, so detectors trained only on 50-step DDPM outputs can silently fail on distilled or flow-matched content.

Video diffusion. Image-to-video and text-to-video diffusion (Stable Video Diffusion [46] and successors) extend synthesis to temporally coherent clips, raising the prospect of synthetic abuse *video* that defeats both frame-level hashing and per-frame image detectors lacking temporal modeling. Table 2 maps each advance to the detection assumption it breaks.

Table 2. Generation advances and the detection assumption each breaks.

Advance	Capability	Assumption broken
Face swap / reenact. [43,44]	Local identity graft onto real media	Whole-image synthesis cue
Identity adapters [47,48]	Zero-shot victim injection	Persistent LoRA weight signature
Fast samplers [49]	1–4 step generation	Fixed spectral/reconstruction print
Flow matching [50]	New noise trajectory	DDPM-trained DIRE prior
Video diffusion [46]	Coherent synthetic clips	Per-frame-only detection

3.3. Generative Paradigms Compared for Forensics

The four generative paradigms that matter leave distinct forensic signatures, and a detector's reach is bounded by which it was built against (Table 3). **GANs** learn an implicit sampler through the minimax game; their transposed-convolution up-samplers leave the sharpest, most periodic spectral fingerprint, which is why GAN-era detection was comparatively easy and why GAN attribution is mature [22]. **VAEs / latent autoencoders** underpin latent diffusion; their decoder leaves a softer periodic grid and defines the manifold that ARE exploits (§6.7). **Diffusion models** learn an explicit (score) model sampled iteratively; their trace is weaker and content-adaptive, recovered best by reconstruction (DIRE) rather than by a fixed spectral peak. **Autoregressive and masked-token image models** generate in discrete latent space and leave block/tokenization artifacts distinct from both GAN and diffusion. The forensic consequence is that *no single cue family is complete*: a spectral detector tuned on GANs underperforms on diffusion, and a DDPM-reconstruction detector underperforms on flow-matched or autoregressive output—the formal justification for the multi-cue fusion of §8.3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Table 3. Generative paradigms and their dominant forensic signature.

Paradigm	Mechanism	Dominant signature
GAN [10]	Implicit minimax sampler	Sharp periodic spectrum
VAE/LDM [12]	Latent autoencoding	Soft grid; on-manifold (ARE)
Diffusion [11]	Iterative score sampling	Low recon. error (DIRE)
Autoregressive	Discrete-token prediction	Tokenization/block trace

3.4. Adversary Model

We assume an adversary who: downloads open-weight checkpoints and community fine-tunes; runs *local* inference, so no server-side moderation applies; uses inpainting and personalization for either victim-specific NCII or victimless-appearing synthetic CSAM; and launders outputs through the channel T to defeat both hashing and fingerprinting. The adversary's *cost drivers* are compute, model access, data for personalization, and the risk of attribution. Each is a defensive lever. We also consider an *adaptive* adversary who knows our detectors and applies gradient-based or purification-based evasion (Section 12); clean-only accuracy is treated as insufficient reporting throughout.

Table 4. Adversary capability tiers and dominant defensive lever.

Tier	Capability	Dominant lever
T0 Casual	Hosted app, default model	Gateway moderation, provenance
T1 Local	Open weights, local inference	Synthesis detection + attribution
T2 Tuner	Fine-tune / LoRA on a target	Adapter attribution, takedown
T3 Adaptive	Knows detectors, evades	Adversarial training, ensembles
T4 Builder	Trains from contaminated data	Dataset governance, memorization audit

4. Related Work

Explicitness and nudity classification. The sub-problem of detecting sexual content predates deepfakes: statistical skin-colour models [3] and bag-of-visual-words classifiers such as Skin Sheriff [4] established the classical baseline, and modern comparative studies benchmark deep nudity classifiers across architectures and robustness regimes [5]. Our prior work targeted the deepnude sub-problem specifically with classical ML pipelines [1]; the present paper generalizes it from a nudity classifier toward a theory that composes explicitness with synthesis detection, attribution, and prevention.

Operational and open-source forensic tooling. Nudity detection reached the field as deployable forensic software well before deepnudes. NuDetective, built by forensic experts for crime-scene triage, combines skin-pixel detection and computational geometry with file-hash and file-name matching—and later video frame sampling—to flag child-abuse material on seized media [63]. Open-source NSFW detectors generalize the explicitness classifier: `adult-image-detector` wraps an `open_nsfw`-style network over OpenCV for image and PDF scoring [64], and the widely used NudeNet supplies a reusable nudity classifier/detector. These systems share a blind spot—they score *content*, not *synthesis*, so a photorealistic deepnude that reads as "nude" gives no signal that it was *generated*. Recent tooling begins to close that gap by fusing nudity detection with deepfake/deepnude detection and preemptive protection: DeepNudeDetect combines NudeNet, CLIP, a custom deepnude CNN, and a deepfake model with Fawkes-style cloaking and adversarial perturbation [65], and Mary Poppins integrates perceptual hashing (pHash/PDQ/PhotoDNA), a staged CSAM/age AI classifier, grooming NLP, and a tamper-evident chain-of-custody audit into a zero-visual-exposure investigative platform [66]. This

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

paper provides the theory and methodology that such tools instantiate, and pushes them past the *deepnude barrier* defined in §7.6.

Deepfake / synthetic-image detection. Media-forensics surveys frame the field and its arms race [15]. Early detectors exploited GAN spectral artifacts [18] and the surprising transferability of a simple CNN trained on a single generator [17]. Face-manipulation benchmarks (FaceForensics++ [20]) drove video deepfake detection, and camera-forensic residuals (noiseprint [52]) supplied the conceptual template for model fingerprints. Diffusion outputs defeated many GAN-era detectors; DIRE reframed detection around *reconstruction error* under a diffusion model [16], autoencoder reconstruction gave a training-free variant [54], and foundation-model features (CLIP) yielded strong universal detectors [19]. DE-FAKE studied detection and attribution of text-to-image outputs jointly [21]. Physiological-signal detectors (rPPG) add an orthogonal video cue [55].

Attribution and provenance. GAN fingerprints [22] and weight-modulation user attribution (WOUAF) [23] attribute an image to a model or user; CLIP/DINOv2 features enable open-set origin attribution [24]. Asnani et al. (Meta AI / MSU) push this furthest with *model parsing*—inferring a generative model's network-architecture hyperparameters and training-loss type from a single generated image, explicitly beyond closed-set attribution [62]. Provenance standards (C2PA [25]) and watermarking (SynthID [26], Stable Signature [27], Tree-Ring [28]) aim to mark content at generation time.

Preemptive protection. PhotoGuard raises the cost of malicious edits by adversarially perturbing the source image [30]; Glaze and Nightshade protect against style mimicry and poison training [31,32]. A growing body documents that these protections are *brittle* under purification [33] and adaptive attack [34], and that watermarks can be removed or forged [29].

Generation advances as a moving target. The detection literature races a generation literature: subject-agnostic face swapping [43], motion-transfer reenactment [44], audio-driven talking heads [45], zero-shot identity adapters [47,48], few-step distillation [49], rectified-flow backbones [50], and video diffusion [46] each erode a detection assumption (§3.2). Any durable forensic claim must therefore be stated against unseen generators, not a fixed panel.

Dataset and safety failures. Thiel et al. established CSAM contamination of LAION-5B [2]; Rando et al. red-teamed the Stable Diffusion safety filter [35]; Carlini et al. showed diffusion models memorize and can regurgitate training images [36]—directly relevant to re-victimization.

Law, policy, and privacy infrastructure. A parallel literature governs response: statutory instruments criminalizing virtual and AI-generated abuse material and NCII [38–41], forensic-admissibility doctrine [37], and privacy-preserving matching proposals with their documented perceptual-hash attack surface [60,61].

Gap. No prior work unifies a theory of deep nude detection, open-set attribution, no-possession training, preemptive immunization, provenance, cybersecurity hardening, and operational/legal integration—together with a reusable research methodology—specifically for the synthetic-CSAM forensic setting. This paper targets that gap.

5. The Uncensored Generation Ecosystem: A Risk Taxonomy

Open model hubs host hundreds of thousands of image checkpoints. The overwhelming majority are legitimate. A long tail, however, is explicitly oriented toward unrestricted sexual output. We characterize this ecosystem *structurally*, without naming or linking abuse-enabling artifacts, because the structure—not any one checkpoint—is what defenses and policy must address.

5.1. Historical Evolution of the Abuse Toolchain

Understanding the defense requires understanding how the threat arrived at its present form. The abuse toolchain has passed through four overlapping generations (Table 5), each lowering cost and each shifting the detectable artifact. The 2019 "DeepNude" application was a single conditional GAN performing clothed→nude translation: narrow, low-fidelity, and leaving strong GAN spectral fingerprints. Latent-diffusion inpainting (2022 onward) generalized this to arbitrary edits at photographic fidelity,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

collapsing fingerprints toward the VAE's softer signature and introducing the composite structure of §7.5. Personalization (LoRA/DreamBooth, 2022–2023) made the attack *victim-specific* from a handful of images; zero-shot identity adapters (2023–2024, §3.2) removed even the fine-tuning step. The trajectory is monotone in three quantities—rising fidelity, falling cost, and migrating signature—and every defense in this paper is positioned against one of those axes.

Table 5. Four generations of the synthetic-abuse toolchain.

Era	Technology	Forensic shift
2019	Conditional GAN (DeepNude)	Strong GAN spectral print; narrow
2022	LDM inpainting	Photographic composites; softer print
2022–23	LoRA/DreamBooth	Victim-specific; adapter signature
2023–24	Zero-shot ID adapters	No persistent adapter; signature in conditioning

5.2. The Economics of Abuse: An Asymmetry Model

Defense is ultimately an economic problem: the goal is to invert a cost asymmetry that currently favours the attacker. Let c_A be the adversary's marginal cost to produce one laundered artifact and c_D the defender's marginal cost to detect, attribute, and action it. The retrospective hashing paradigm had $c_A \gg 0$ (abuse material had to be captured and distributed) and c_D small (a hash lookup); synthesis inverts this to $c_A \to 0$ and c_D dominated by human review. The defensive levers of this paper map directly onto the terms:

- **Raise c_A:** immunization and poisoning (§12) add per-target compute and uncertainty; provenance and KYC (§13, 16) add attribution risk.
- **Lower c_D:** automated synthesis detection and attribution (§8) cut per-item human cost; clustering by model family amortizes one decision over many artifacts.

Writing the defender's expected cost as $c_D = p_{\text{review}}\, c_{\text{human}} + (1 - p_{\text{review}})\, c_{\text{auto}}$, where p_{review} is the fraction escalated to humans, shows why triage calibration is the highest-leverage variable: a detector that safely reduces p_{review} at fixed error attacks the dominant term. This framing also bounds optimism—because c_A cannot be driven arbitrarily high against an open-weights adversary (§7.8), the durable strategy is to minimize c_D and concentrate the raised-c_A levers on the cooperating and hosted parts of the ecosystem.

5.3. Nudify-as-a-Service: The Commercial Layer

The open-weights ecosystem has a commercialized front end that makes the attack accessible to non-technical users: a layer of "nudify" / "undress" web and mobile applications and chatbots that wrap uncensored generators behind a one-click, often freemium, interface. Public-interest investigations document the scale and business model of this layer—dozens of such services, millions of visits, freemium and referral monetization, and payment, hosting, and app-distribution dependencies [8,67]. We treat it, like the rest of the ecosystem (§5), *structurally and without profiling or linking specific services*: naming or routing to them would aid the abuse this paper exists to counter, and our analysis needs only the structure.

Structurally, these services are the tier-T0 (casual) and tier-T2 (victim-targeting) attack made turnkey: the user supplies a photograph—frequently of a real woman or a minor, non-consensually—and the service returns a deepnude, folding the inpainting/identity-transfer pipeline of §3.2 into a hosted product. Their harm profile is accordingly severe: mass NCII, sextortion, and, when the input depicts a minor, synthetic CSAM.

Crucially, *because this layer is commercial and hosted, it is the part of the ecosystem most exposed to intervention*—it depends on chokepoints an open-weights offender does not. The levers follow directly:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Financial:** payment processors and app stores denying service to undressing apps removes the monetization that sustains them.
- **Infrastructure:** hosting providers, CDNs, and domain registrars acting on abuse, and search/ad platforms de-listing and de-monetizing, raise operating cost and discoverability.
- **Legal:** NCII statutes and platform-duty regimes (§15, e.g. the TAKE IT DOWN Act [39]) create takedown and liability pressure specifically on hosted services.
- **Technical:** source-image immunization and poisoning (§12.5) degrade the service's output quality, while provenance and attribution (§6.8) erode the anonymity the business model assumes.

In the cost model above, the nudify-service layer is where raising c_A is most tractable: it has a balance sheet, a domain name, and a payment rail, none of which a local script has. Concentrating enforcement here shrinks the *casual* and *at-scale* attack surface even though it cannot reach the determined local offender—who is instead the target of detection and attribution. Table 6 summarizes the chokepoints and the actor who holds each.

Table 6. Chokepoints against the commercial deepnude-service layer.

Chokepoint	Action	Who holds it
Financial	Deny payment / in-app billing	Processors, app stores
Infrastructure	Suspend hosting/CDN/domain	Providers, registrars
Discovery	De-list, de-monetize ads	Search, ad networks
Legal	Takedown, liability	Legislators, regulators
Technical	Immunize, poison, attribute	Platforms, researchers

5.4. How Alignment Is Removed (Category Level)

Safety behaviour in a released model is enforced by *separable* components, which is precisely why it is fragile:

- **Post-hoc output classifiers** (NSFW/safety checkers) sit *outside* the generator and can simply be disabled when weights run locally. Rando et al. showed the Stable Diffusion filter was a thin, reverse-engineerable classifier [35].
- **Alignment fine-tuning** can be *undone* by further fine-tuning, overwriting suppression without retraining from scratch—the same mechanism that strips watermarks [27].
- **Latent capability** persists because unsafe concepts are present in pretraining data [2]; "removal" often suppresses rather than erases, and personalization re-surfaces it.

We present these as *detector/policy design drivers*: any defense assuming the safety layer is intact is defeated by local inference (adversary tier $\geq$ T1).

5.5. Model Lifecycle and Lineage

Fig. 2 shows the lifecycle through which risk propagates: a base model derives community fine-tunes and adapters, which recombine and redistribute. Each edge is an *attribution opportunity* (lineage can be fingerprinted) and an *enforcement point* (scanning, licensing, takedown).

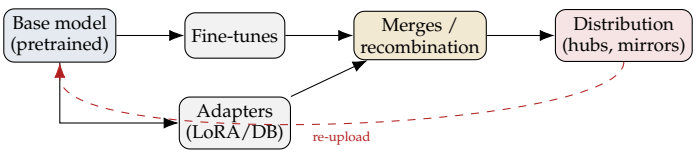


Figure 2. Model lifecycle and lineage. Each edge is simultaneously an attribution handle and a policy enforcement point.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Table 7. Layered risk taxonomy of the open generation ecosystem.

Layer	Abuse-enabling property	Defensive lever
Data	Contaminated web-scrape; memorization [2,36]	Dataset hashing/scrub; training-data provenance
Weights	Separable safety layer; fine-tune reversibility	Model fingerprinting; hub scanning; licensing
Adapters	LoRA/DreamBooth target identities	Adapter attribution; hashing; takedown
Inference	Local, unmoderated; inpainting	Gateway controls; client-side provenance
Distribution	Hubs, mirrors, torrents	KYC, scanning, legal duties
Output	Novel pixels defeat hashing	Pipeline-forensic detection

5.6. Discovery and Monitoring (Defensive OSINT)

Child-protection and platform-integrity teams need to *measure* this ecosystem without participating in it. Defensible monitoring uses metadata only: model cards, tags, licenses, download-count trends, upload cadence, and the lineage graph (base→derivative). This supports risk scoring, hub-level enforcement, and early warning—and it is the *only* reconnaissance this paper endorses. A risk score $r(m)$ for a hosted model can be assembled from observable, non-content features:

$$r(m) = \sigma\Big(\alpha_1 \text{tags} + \alpha_2 \text{lineage-risk} + \alpha_3 \text{velocity} + \alpha_4 \text{license}\Big), \tag{5}$$

enabling triage of enforcement attention without any model being executed or any output being generated by the defender.

6. Reverse Engineering Generative Pipelines for Attribution

"Reverse engineering" here means the forensic inverse problem: given an artifact, recover evidence about the *process* that produced it. This is the pivot from retrospective matching to pipeline forensics, and it supplies the courtroom-relevant question "where did this come from?" rather than only "is this fake?"

An attribution ladder, stated conservatively. It is essential not to overclaim what attribution yields: each rung below licenses only a bounded conclusion, and reaching a person requires evidence *outside* the image (Table 8). In particular, image-only methods can support "consistent with family m"; they do *not*, on their own, identify a checkpoint, a user, or an offender, nor does a shared family fingerprint establish that two artifacts belong to one coordinated campaign—that is a hypothesis to corroborate with external records, not a conclusion. We use the strong words ("identify the operator," "dismantle the service") only for the *combined* pipeline of image evidence plus lawful external evidence, and flag where each applies.

Table 8. Attribution ladder: each rung's warranted conclusion and what it additionally requires.

Result	Warrants	Needs more
Synthesis detection	"signs of generation/edit"	—
Family attribution	"consistent with family m"	open-set calib.
Checkpoint/adapter ID	specific model instance	extra evidence, validation
User/operator ID	a person/account	external: logs, infra, payments
Campaign linkage	artifacts co-originate	corroborating metadata

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Algorithm 1 DIRE feature extraction (defensive)

Require: image x; pretrained diffusion ϕ; inversion steps S
$z \leftarrow \mathcal{I}_\phi(x, S)$ ▷ DDIM inversion to noise
$\hat{x} \leftarrow \mathcal{R}_\phi(z, S)$ ▷ deterministic re-denoising
$d \leftarrow |x - \hat{x}|$ ▷ per-pixel reconstruction error
return d, $\|d\|_1$ ▷ feature map and scalar score

6.1. Diffusion Reconstruction Error (DIRE)

A diffusion model can near-reconstruct its own outputs through deterministic DDIM inversion and re-denoising, whereas real photographs reconstruct poorly [16]. Let ϕ be a pretrained diffusion model, $\mathcal{I}_\phi(\cdot)$ its deterministic inversion to latent noise, and $\mathcal{R}_\phi(\cdot)$ the reconstruction. The DIRE representation of image x is

$$\mathrm{DIRE}(x) = \big|\, x - \mathcal{R}_\phi(\mathcal{I}_\phi(x)) \,\big|, \tag{6}$$

and a classifier $f_\theta(\mathrm{DIRE}(x))$ separates synthetic from real. The intuition is geometric: a diffusion output already lies on (or very near) the manifold the model can generate, so the forward–backward map $\mathcal{R}_\phi \circ \mathcal{I}_\phi$ is close to the identity on it, whereas a real photograph carries sensor noise, optical blur, and off-manifold detail that inversion cannot reproduce, inflating $\|\mathrm{DIRE}(x)\|$. The error *generalizes across unseen diffusion models* because the manifold proximity, not a model-specific artifact, is what it measures—a key property against an adversary who uses tomorrow's checkpoint. Two caveats shape deployment: (i) the inverting model ϕ need not be the generator, but a large architectural gap (e.g. inverting a flow-matched image with a DDPM, §3.2) weakens the signal, so a small panel of inverters is used; (ii) aggressive purification [33] itself projects a real image toward the manifold, raising false positives, which is why DIRE is fused rather than used alone (§8.3). Algorithm 1 gives the procedure.

6.2. Model Fingerprints: Spectral and Weight-Space

Generators imprint systematic traces along two complementary axes—a passive one recoverable from the pixels alone, and an active one that a licensed deployer can plant.

- **Spectral (passive).** Transposed-convolution and pixel-shuffle up-samplers inject periodic high-frequency structure absent from camera optics and sensor pipelines. The azimuthally averaged power spectrum

$$P(f) = \frac{1}{|\{\|k\| = f\}|} \sum_{\|k\| = f} |\mathcal{F}\{x\}(k)|^2 \tag{7}$$

  concentrates this into a compact, model-discriminative descriptor [18,22]: GAN up-samplers produce sharp spectral spikes at stride-dependent radial frequencies, latent-diffusion VAE decoders produce a softer but still periodic grid, and the *peak location and sharpness* discriminate architecture and even up-sampler stride. The same statistic also distinguishes a passive *model fingerprint* (an unintended, systematic trace shared by a family) from a per-instance artifact, which matters when arguing in court that a signature is characteristic of a family rather than coincidental.
- **Weight-modulation attribution (WOUAF, active).** Per-user weight modulation makes each licensed deployment's outputs carry a recoverable identifier at minimal quality cost [23]. Because the signature lives in the weights rather than a strippable metadata field, it survives re-encoding and screenshotting, giving hosted/licensed generators a proactive attribution primitive—the cooperative counterpart to the passive spectral print that is all one can recover from an uncooperative open-weights offender.

A camera-forensics analogy. The conceptual template for generative attribution is classical camera forensics, where a sensor's photo-response non-uniformity (PRNU) acts as a device fingerprint and a learned noise residual (noiseprint [52]) isolates the camera-model trace from scene content. Generative fingerprints play the same role one level up the pipeline: the "device" is the model family, and the trace is the systematic residual its sampler and decoder leave in every output. This analogy transfers a

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Algorithm 2 Open-set attribution with abstention

Require: artifact x; families $\mathcal{M}$; cues $\{g_k\}$; weights $\{w_k\}$; threshold τ
for $m \in \mathcal{M}$ **do**
  $s_m \leftarrow \sigma(\sum_k w_k\, g_k(x, m))$
end for
$m^\star \leftarrow \arg\max_m s_m$
if $s_{m^\star} < \tau$ **or** max-gap$(s) < \delta$ **then**
  return $\varnothing$ ▷ open-set: abstain ("unknown")
else
  return $m^\star$, calibrated $s_{m^\star}$, per-cue breakdown
end if

mature forensic vocabulary—reference fingerprint, correlation detector, and the distinction between *source-device* and *source-class* identification—directly onto the synthesis setting, and it is why the same residual-domain reasoning used for inpainting localization (§6.5) also serves attribution.

For forensics, the key metric is *closed-set vs. open-set* attribution accuracy under laundering T. Open-set attribution—assigning "unknown" rather than forcing a known label—is essential and is best served today by foundation-model features (CLIP/DINOv2) [24]. We note that spectral fingerprints degrade under JPEG and resize, so they must be *fused* with reconstruction- and feature-based cues rather than relied on alone.

6.3. Membership and Memorization as Linkage Evidence

Carlini et al. showed diffusion models memorize training images and can emit near-duplicates [36]. Forensically this is double-edged. It enables *re-victimization* (a known survivor's abuse image reproduced), but it also enables *linkage*: membership inference and near-duplicate detection can connect a synthetic artifact back to a specific contaminated dataset or model, strengthening attribution and takedown. A near-duplicate test between a candidate x and an enclave-held reference set $\mathcal{R}$ computes $\min_{r\in\mathcal{R}} \mathrm{dist}(\psi(x), \psi(r))$ in a perceptual embedding ψ, run *only* inside the accredited enclave, surfacing re-victimization without exporting references.

6.4. An Attribution Fusion Score

We combine evidence streams into a calibrated attribution posterior. For model family m and artifact x:

$$s(x, m) = \sigma\Big(\sum_k w_k\, g_k(x, m)\Big), \tag{8}$$

where g_k are per-cue log-likelihood ratios (DIRE score, spectral match, latent-fingerprint correlation, CLIP-feature distance, watermark decode), w_k learned weights, and σ the logistic link. Algorithm 2 assembles the decision, including the open-set "unknown" rule. Calibration (temperature scaling; reliability curves) is *mandatory*: forensic outputs must express defensible uncertainty, not raw confidence (Section 14).

6.5. Conditioning Recovery: Prompt, Seed, and Inpainting Forensics

Beyond "which model," the pipeline leaks *how it was driven*. Three inverse problems are forensically valuable.

Prompt inversion. Hard-prompt recovery optimizes a discrete token sequence whose embedding reproduces a target image under the model (PEZ-style gradient projection [51]). A recovered prompt is intelligence—it exposes intent, targeting language, and the conditioning vocabulary of an offender toolchain—and can corroborate authorship across a seized corpus.

Seed and sampler inference. For deterministic samplers, the initial noise and schedule constrain the output; recovering (or bounding) the seed links multiple artifacts to a single generation session and supports Tree-Ring-style noise-space detection [28]. Sampler and step-count signatures (distilled vs. 50-step, flow vs. DDPM, §3.2) further narrow the pipeline.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Inpainting-boundary forensics. The dominant NCII vector is masked regeneration, which fuses synthetic pixels into a real photograph along a mask seam. The seam is recoverable: noise-residual (noiseprint-style) inconsistency [52] across the boundary, a step change in the local power spectrum $P(f)$, and double-compression traces outside vs. inside the mask localize the edit. We estimate a per-pixel manipulation heatmap

$$h(x) = \lambda_1 \left|\nabla \text{noiseprint}(x)\right| + \lambda_2 \, \Delta_{\text{spec}}(x) + \lambda_3 \, \text{DIRE}(x), \tag{9}$$

whose connected high-response regions estimate the edited mask $\hat{M}$. Localizing the edit both raises detection accuracy on swap/reenactment content and yields courtroom-legible evidence ("this region was synthesized").

6.6. Weight-Space Lineage and Adapter Recovery

When weights are seized (tier $\geq$ T2), attribution becomes a white-box problem. Merged checkpoints are approximately linear combinations of a base and task vectors $\theta = \theta_{\text{base}} + \sum_i \alpha_i \tau_i$ in the task-arithmetic sense [53]; correlating a candidate's weight delta $\theta - \theta_{\text{base}}$ against a library of known task vectors recovers lineage and can *un-mix* contributing fine-tunes. For LoRA, the low-rank update $\Delta W = BA$ is a compact, matchable signature: the singular-subspace of ΔW is stable across export formats, so subspace correlation attributes a shipped adapter to a known family even after renaming or re-quantization. This complements the black-box latent/weight-modulation fingerprints (WOUAF [23]) usable when only outputs are available.

6.7. Zero-Shot Reconstruction Detection (No Generator Library)

DIRE needs a diffusion model to invert through. A lighter primitive uses only the *autoencoder*: a latent-diffusion output reconstructs through its own (or a surrogate) VAE with lower error than a real photograph, because the output already lies on the VAE manifold. The autoencoder reconstruction error

$$\text{ARE}(x) = \|x - \text{Dec}(\text{Enc}(x))\| \tag{10}$$

is a training-free detection score (AEROBLADE-style [54]) that also discriminates *which* LDM family by trying a panel of VAEs and taking the argmin. It is cheap enough for intake triage and degrades gracefully to the open-set case.

6.8. Model Parsing: Recovering Architecture from One Image

The deepest form of reverse engineering recovers not just *which known model* but *the design of an unknown one*. Asnani et al. (Meta AI and Michigan State) frame this as *model parsing*: estimating a generative model's network-architecture hyperparameters (number of layers and blocks, operation types) and training-loss family from a single generated image [62]. A fingerprint-estimation network is trained unsupervised under general fingerprint priors—magnitude, repetitiveness, frequency range, and symmetry of the frequency response—so the estimate captures the model's trace rather than image content; a parsing head then regresses the hyperparameters. The forensic significance is open-setness by construction: because it predicts *properties* of the generator rather than selecting from a fixed list, it extends attribution to models never seen in training—"recognizing the components of a car from how it sounds," in the authors' analogy. For CSAM investigation this converts a novel, unseen generator from a dead end into a set of recoverable design attributes that can cluster cases and narrow the toolchain behind a campaign.

6.9. Attacking the Generator: Inversion and Algorithmic Reversal

Reversing a pipeline for attribution shades into *neutralizing* it: the same differentiability that makes a generator powerful is the lever that lets a defender invert, fingerprint, and disrupt it. We propose a defensive program built on three mathematical primitives, all usable without redistributing any generative capability.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

(1) Gradient-domain inversion. For a differentiable generator or editor $\mathcal{G}$, recovering the latent/conditioning that produced an artifact is the optimization $\min_{z,c} \|\mathcal{G}(z,c) - x\|^2 + \mathcal{R}(z,c)$, solved by back-propagation with a prior $\mathcal{R}$. A solution yields the seed/prompt evidence of §6.5; a *failure* to invert within a model's reachable set is itself evidence the artifact came from elsewhere, an open-set signal.

(2) Score / Jacobian probing. A diffusion model's learned score $\nabla_x \log p_t(x)$ and the local Jacobian of its denoiser encode a family-specific operator. Estimating these around a candidate—by finite differences or autodiff on a surrogate—gives a behavioural fingerprint that complements the passive spectral one (§6.6) and is harder to launder because it probes *dynamics*, not static pixels.

(3) Boundary and cross-attention extraction. Where query or white-box access to an abuse service exists under lawful authority, the decision/generation boundary can be mapped by controlled queries, and cross-attention maps expose which conditioning tokens drive which regions—localizing and attributing an edit. The same machinery underlies model-extraction *defenses*: a defender who knows how cheaply a boundary leaks can rate-limit and calibrate to deny it (§16).

The unifying idea—"kill the algorithm to detect it"—is adversarial symmetry: every property that makes generation controllable (differentiability, a smooth manifold, stable fingerprints, guidance) is a handle for detection, attribution, or disruption. Table 9 maps each primitive to its forensic product and its limit.

Table 9. Reverse-engineering primitives and their forensic yield.

Primitive	Yield	Limit
DIRE/ARE recon [16,54]	Generated? which family	Purification
Model parsing [62]	Architecture/loss of unseen model	Needs fingerprint est.
Gradient inversion	Seed/prompt; open-set flag	Non-convex; cost
Score/Jacobian probe	Dynamic fingerprint	Needs surrogate
Weight un-mix [53]	Lineage, LoRA signature	Needs weights

6.10. Dismantling a Hosted Generation Service (Defensive, Black-Box)

The commercial "nudify" layer (§5.3) is a hosted black box, but a lawful investigator with authority to collect a service's *own outputs* can reverse-engineer and dismantle its pipeline without any operational detail about the service and without redistributing capability. We describe the methodology at the category level; active probing of a live service is undertaken only under appropriate legal authority, and we deliberately give no service-specific procedure.

The method chains the primitives above on a sample of outputs. (i) *Base-model identification*: run DIRE/ARE and spectral analysis (§§6.7, 6.6) and model parsing [62] over sampled outputs to recover the architecture family and, often, the specific open base model the service wraps. (ii) *Pipeline fingerprinting*: sampler/step signatures, guidance-scale imprint (§3.2), and fixed post-processing (upscalers, watermarks) narrow the exact toolchain. (iii) *Provenance/watermark probing*: decode any embedded marks to attribute to a hosted upstream generator that should have marked its output. The forensic payoff is leverage, not exploitation: identifying the *base model* converts a faceless service into a takedown/notice target against the model host and hub (§16), feeds family fingerprints into the detector so the service's future output is caught downstream, and supplies evidence for the financial/infrastructure/legal chokepoints of §5.3. "Dismantling the algorithm" therefore means attributing and cutting off the pipeline a service depends on—base model, hosting, payment, and anonymity—rather than attacking the site directly. Algorithm 3 states the workflow.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Algorithm 3 Attribution-driven dismantling (defensive, lawful)

Require: lawfully collected outputs $X = \{x_i\}$ of a hosted service; VAE/diffusion panel Φ; base-model library $\mathcal{M}$

for $x_i \in X$ **do**
  $d_i \leftarrow$ DIRE/ARE over Φ ▷ §6.7
  $p_i \leftarrow$ model-parsing hyperparameters ▷ §6.8
  $w_i \leftarrow$ watermark/provenance decode ▷ §13
end for
$m^\star \leftarrow$ consensus base family over $\{d_i, p_i\}$ against $\mathcal{M}$
register fingerprint of $m^\star$ in detector ▷ catch future output
if w_i attributes a hosted upstream **then**
  notice to that provider
end if
return $m^\star$, evidence bundle $\rightarrow$ chokepoints (§5.3)

7. Foundations of Deep Nude Detection: Theory and Techniques

This section develops the conceptual and mathematical basis for deep nude detection, generalizing the classifier-centric treatment of our prior study [1] into a theory that the architecture of §8 then instantiates. The goal is to give researchers a principled decomposition, the lineage of techniques that address each part, and the decision-theoretic and information-theoretic framing that explains why a single classifier is insufficient and a calibrated multi-cue system is required.

7.1. Problem Decomposition

"Is this a harmful deepnude?" is not atomic. It factors into three largely orthogonal latent questions about an artifact x:

$$E(x) \in [0,1] \quad \textit{explicitness}\text{: sexual/nude content present;} \tag{11}$$

$$M(x) \in [0,1] \quad \textit{manipulation}\text{: synthesized or edited;} \tag{12}$$

$$S(x) \in \mathcal{A} \quad \textit{subject attributes}\text{: age band, identity.} \tag{13}$$

The operational verdicts of §2 are functions of these factors *together with information the pixels do not contain*. Two refinements are essential and were missing from a naive reading. First, the factors must be split finer than a single "explicitness" score: *nudity* is not *sexual content*, and *apparent age* (an image-estimable quantity, with wide uncertainty and demographic bias) is not *verified age* (an external fact). We therefore treat $E = (E_{\text{nude}}, E_{\text{sex}})$ and $S = (\hat{a}, \text{id})$ with $\hat{a}$ an *apparent*-age estimate carrying a calibrated confidence. Second, and decisively, *consent and context are not image-intrinsic*: the same factorization cannot, from pixels alone, distinguish a consensual edit from non-consensual intimate imagery. Consent, authorization, and context enter as *external/unknown* variables U (provenance, reports, case metadata), never inferred from x.

Thus NCII corresponds to $E_{\text{sex}} \uparrow \wedge M \uparrow \wedge$ identity-linked-to-a-real-person $\wedge \neg\text{consent}(U)$; synthetic CSAM to $E_{\text{sex}} \uparrow \wedge \hat{a}$=minor (with M either value, since photographic and synthetic CSAM are both contraband); and the image-estimable *conjunction* the classifier targets is only $C(x) = \Pr[E_{\text{sex}} \uparrow \wedge \hat{a}\text{=minor}]$—a *triage signal with an abstention option*, not a legal determination, which requires U and human adjudication. Factoring the problem this way is also what makes the no-possession design of §8 possible: E and M are learnable from legal data, and only the conjunction with $\hat{a}$=minor is contraband and enclave-restricted. Each factor has its own technique lineage, which we trace below.

7.2. The Common Substrate: Neural Networks, ML, and Deep Learning

Detection and generation are two faces of one substrate, and stating it plainly clarifies why the arms race is intrinsic. *Machine learning* fits a parameterized function f_θ to data by minimizing an

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

empirical risk $\hat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_i \ell(f_\theta(x_i), y_i)$ via stochastic gradient descent, $\theta \leftarrow \theta - \eta\,\nabla_\theta \ell$. A *neural network* realizes f_θ as a composition of linear maps and nonlinearities, $f_\theta = \sigma_L \circ W_L \circ \cdots \circ \sigma_1 \circ W_1$, trained end-to-end by back-propagation; *deep learning* is the regime where many such layers learn a hierarchy of representations—edges, textures, parts, objects, scene semantics—directly from pixels, removing the hand-crafted feature stage of the classical pipeline (§7.3).

Two distinctions organize everything that follows. First, *discriminative* models estimate $p(y \mid x)$—the detectors, nudity classifiers, and attributors of this paper—while *generative* models estimate or sample $p(x)$ or $p(x \mid c)$—the GANs and diffusion models of §3 that produce the abuse. The adversary trains the second; the defender trains the first; and because both minimize related objectives over the same image manifold, an improvement in generation is, by construction, a harder discrimination problem (§7.8). Second, the *convolutional* inductive bias (weight sharing, locality) that makes CNNs strong image classifiers also makes their transposed-convolution inverses leave the periodic spectral fingerprints that attribution exploits (§6.6)—the same architecture choice aids both sides.

Three deep-learning mechanisms recur throughout the paper and are worth naming once. *Transfer learning* adapts a model pretrained on a large corpus (an ImageNet CNN, a CLIP or DINOv2 encoder) to a target task with a small labelled head—the recipe behind both modern nudity classifiers and the strongest universal detectors [1,19,24]. *Representation learning* means the penultimate-layer features, not just the label, carry transferable signal—why the same backbone serves detection, attribution, and near-duplicate linkage. *Adversarial examples*—imperceptible input perturbations that cross a decision boundary—are the shared mechanism behind both detector evasion (§12) and the immunization/-cloaking defenses (§12.2), a duality that recurs because both exploit the differentiability of f_θ. This substrate is what lets the factors E, M, and S be learned, fused, and attacked with one mathematical toolkit.

7.3. Classical Computer-Vision Pipeline

Early explicitness detection (E) was dominated by hand-crafted colour, region, and texture features, and the lineage still informs feature design and fast pre-filters.

Skin modelling. The oldest reliable cue is skin chrominance. Mapping x into a luminance-decoupled space (YCbCr or HSV) concentrates skin pixels in a compact chrominance region. A pixel p is scored by a likelihood ratio between skin and non-skin colour models,

$$\Lambda(p) = \frac{P(\mathbf{c}_p \mid \text{skin})}{P(\mathbf{c}_p \mid \overline{\text{skin}})} \gtrless \theta, \tag{14}$$

where the class-conditional densities are estimated either parametrically (a Gaussian $\mathcal{N}(\mu, \Sigma)$ over (Cb, Cr)) or non-parametrically (Bayesian colour histograms, the classic large-scale estimate of Jones and Rehg [3]). Connected skin regions then yield shape, extent, and placement features.

Mid-level representations. Aggregating local descriptors into a bag-of-visual-words histogram and classifying with an SVM (the Skin Sheriff design [4]) captures texture and composition beyond raw skin ratio. These systems are fast and interpretable, which keeps them useful as edge pre-filters.

Intrinsic limits. Colour-based explicitness detection confounds skin with skin-coloured backgrounds, fails across illumination and ethnicity without careful modelling, and is blind to *context*—it cannot tell a medical image from abuse, nor a real photo from a synthetic one. These limits are exactly what motivated the shift to learned representations and, orthogonally, to the manipulation factor M.

7.4. Deep Learning Pipeline

Convolutional and transformer classifiers learn a hierarchy—edges, parts, bodies, scene semantics—that subsumes hand-crafted skin and texture features and adds the contextual discrimination the classical pipeline lacks [1]. Transfer learning from large pretrained backbones (ImageNet CNNs; CLIP/ViT encoders) is the dominant recipe: a general visual prior is adapted to explicitness

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

with a modest labelled head, and modern comparative studies map the accuracy/robustness trade space of such nudity classifiers [5]. Three theoretical points matter for this paper:

1. **Representation transfer.** Features learned for E (explicitness) and features learned for M (synthesis, §8.3) are complementary, not redundant: the former keys on semantic content, the latter on generation artefacts. Fusing them is a design consequence, not an engineering convenience.
2. **Data asymmetry.** Deep models are data-hungry, but the positive class for the *conjunction* is contraband. This is the central tension the no-possession methodology (§8) resolves by learning E and M on legal proxies and reserving only the conjunction for accredited enclaves.
3. **Shortcut risk.** Classifiers readily latch onto dataset shortcuts (watermarks, compression, backgrounds) rather than the intended signal, which inflates in-distribution accuracy and collapses under laundering—formalized in the robustness regime of §19.

7.5. The Deepnude-Specific Layer: Composite Detection

A deepnude is rarely a wholly synthetic image; it is a *composite*—a real photograph whose intimate region has been regenerated, or a real body with a grafted face (§3.2). Theoretically this means E and M are spatially *localized* and often disjoint: explicitness concentrates in the edited region, while the manipulation signal concentrates on its boundary. A global "90% real" score is therefore the wrong statistic; the right one is a per-region joint estimate, which is precisely the localization pipeline of §6.5 and §10. This composite structure is the formal reason deepnude detection cannot be reduced to either a nudity classifier or a synthesis classifier alone.

7.6. The Deepnude Barrier and the Tooling Lineage

The operational tools of the field were built for the pre-synthesis world, and the deepnude breaks each of their assumptions in turn—we call this the *deepnude barrier*. A crime-scene triage tool such as NuDetective [63] rests on three cues: file-hash matches against known-bad lists, file-name heuristics, and skin/geometry nudity detection. Synthesis defeats all three: a freshly generated artifact has a novel hash (Property 1), arrives with an innocuous name, and—this is the crux—a nudity detector that fires on a deepnude confirms only that it *looks* nude, never that it was *generated*. Open-source NSFW detectors [64] and reusable nudity networks (NudeNet) inherit the same limitation: they estimate E (explicitness) with no estimate of M (manipulation). The barrier is therefore not a weakness of any one tool but the structural consequence of the factorization in §7.1: content-only tooling measures one factor of a three-factor problem.

Crossing the barrier requires fusing the content factor with synthesis detection, attribution, and localization, which is what the second generation of tooling begins to do. DeepNudeDetect composes a nudity network (NudeNet), CLIP content features, a custom deepnude CNN, and a deepfake model, and—going beyond detection—adds Fawkes-style cloaking and adversarial perturbation to *hamper* generation in the first place [65]. Mary Poppins embeds the detection factors inside a forensic operating model: perceptual hashing (pHash/PDQ/PhotoDNA), a staged CSAM/age classifier, grooming-language NLP, and, decisively for admissibility, a zero-visual-exposure design (raw pixels never touch disk) with a hash-chained, tamper-evident chain of custody [66]. These two systems illustrate the paper's thesis concretely: the first adds the missing M factor and a preemptive layer; the second supplies the no-possession (§8) and admissibility (§14) scaffolding that turns a classifier into forensic evidence. The theory of §§7.1–7.7 is what both need to generalize past today's generators. Table 10 places the tooling on the factor axes of §7.1.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Table 10. Tooling lineage against the detection factors (§7.1): E=explicitness, M=manipulation, S=subject, prevention, forensic scaffolding. For existing tools, ✓=implemented capability; for *this paper*, "prop."=proposed/analyzed, not an evaluated system.

Tool	E	M	S	Prev.	Forens.
NuDetective [63]	✓				hash/CoC
adult-detector [64]	✓				
DeepNudeDetect [65]	✓	✓		✓	
Mary Poppins [66]	✓	partial	✓		✓
This paper (proposed)	prop.	prop.	prop.	prop.	prop.

7.7. Decision-Theoretic Fusion

Given per-cue evidence $g_1, \ldots, g_K$ (skin/explicitness score, CNN posterior, DIRE/ARE reconstruction, spectral match, provenance decode), the Bayes-optimal combiner under conditional independence is the sum of log-likelihood ratios,

$$\mathrm{LLR}(x) = \sum_{k=1}^{K} \log \frac{P(g_k(x) \mid \mathcal{G})}{P(g_k(x) \mid \mathcal{R})}, \tag{15}$$

recovering the logistic attribution fusion of §6 and Algorithm 2 as its calibrated, weight-learned relaxation when independence is only approximate. Two decision-theoretic constraints are specific to this domain:

- **Asymmetric costs.** The loss of a false negative on CSAM is not commensurate with a false positive; the operating threshold must be set by a cost-sensitive / Neyman–Pearson criterion (fix platform-scale FPR, maximize TPR), not by accuracy. This is why §19 reports TPR at FPR$= 10^{-3}$ rather than headline accuracy.
- **Calibration as a first-class output.** A forensic verdict must be a calibrated probability with a defensible error rate (§14); an uncalibrated "confidence" is inadmissible. Temperature scaling and reliability analysis are therefore part of the method, not an afterthought.

7.8. An Information-Theoretic View of the Arms Race

Detection can be cast as communication over a noisy channel: the label $Y \in \{\mathcal{R}, \mathcal{G}\}$ is transmitted through the artifact X and the laundering channel T, and the detector recovers $\hat{Y}$. The achievable accuracy is bounded by the mutual information $I(Y; T(X))$, and one direction is rigorous: every laundering transform T is a data-processing step that, by the data-processing inequality, cannot increase it, $I(Y; T(X)) \leq I(Y; X)$. The other direction we state as a *conditioned hypothesis*, not a theorem, because the reviewer's caution is correct—better perceptual quality does not *automatically* mean smaller statistical distance between the real and generated distributions. **Hypothesis (shrinking margin).** *If* a sequence of generators reduces a chosen divergence $d(p_{\mathcal{R}}, p_{\mathcal{G}})$ between the real and generated distributions over the relevant support, *then* $I(Y; X)$ and hence the achievable detection margin decrease. Under that hypothesis—which must be checked empirically per generator family and divergence, since fidelity gains in one region need not reduce distance everywhere—the realistic objective is *cost inflation and triage*, not a permanent separating boundary. The hypothesis also clarifies why *provenance* (§13) is strategically different: a watermark *injects* side information, raising $I(Y; X)$ by construction rather than racing a possibly shrinking natural signal.

7.9. Cybersecurity Aspects of the Detector

Deep nude detection is itself a security system and therefore an attack surface—a theme this paper treats as co-equal with accuracy.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Evasion (test-time).** An adaptive adversary perturbs x to cross the decision boundary (§12); robustness must be stated under the ℓ_∞/purification threat model of §12, not on clean data.
- **Poisoning (train-time).** A contributor can corrupt training data or federated gradients; defences are curated/hashed data and the DP-clipped aggregation of Algorithm 4.
- **Model extraction and inversion.** Query access can leak the decision boundary or training data; rate limiting, output calibration, and DP bound the exposure.
- **Privacy of deployment.** On-device matching must not expose the known-bad corpus nor transmit user content, linking detection theory to the PSI/perceptual-hash security analysis of §16.1.

Thus the detector inherits the full confidentiality–integrity–availability frame: it must stay accurate (integrity) under evasion and poisoning, available at platform scale, and confidential about both the corpus it matches and the content it scans.

8. A No-Possession Detection Architecture

8.1. Design Constraints

A lawful detection program faces a hard constraint absent from ordinary ML: *the positive class is contraband*. The architecture must let researchers build and evaluate detectors *without possessing* illegal material. We adopt four principles: (P1) train on *structure*, not identity; (P2) keep real material inside accredited enclaves; (P3) use benign proxies and the synthetic-vs-real signal as the learnable target; (P4) deploy via on-device/edge hashing so platforms match without transmitting content.

8.2. Multi-stream detector

Fig. 3 shows the pipeline. Four parallel streams feed a calibrated fusion head.

1. **Synthesis stream:** DIRE + spectral features ⇒ "is this generated, and by which family?"
2. **Content stream:** a nudity/age-risk estimator trained on legal adult and synthetic data to flag the *sexualized-minor* risk region—never trained by researchers on real CSAM (enclave-only fine-tune, distributed as frozen weights).
3. **Provenance stream:** C2PA manifest validation + watermark decode (SynthID/Stable Signature) [25–27].
4. **Hash stream:** PhotoDNA / perceptual + cryptographic hashes against accredited corpora for the retrospective case.

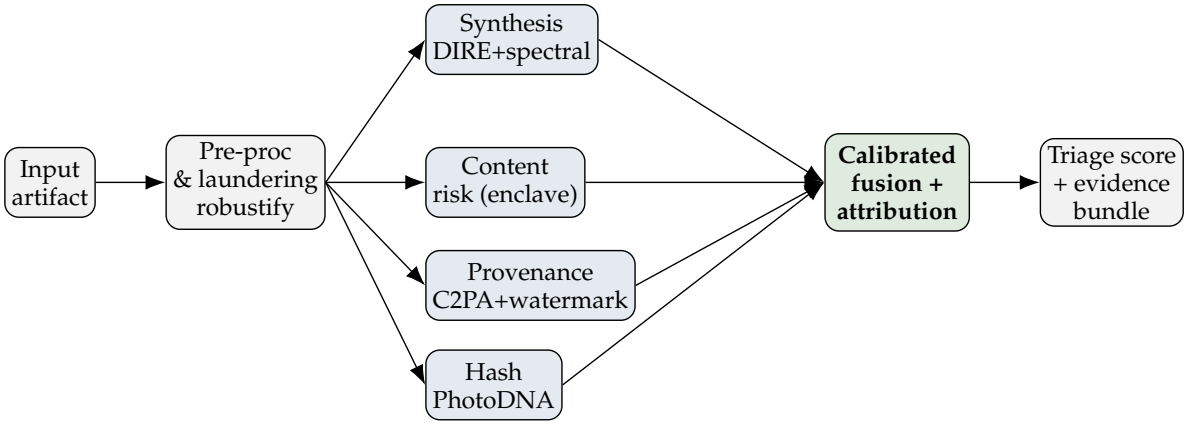


Figure 3. No-possession, multi-stream detection and attribution pipeline. The content-risk stream is fine-tuned only inside accredited enclaves; researchers work with the other three streams and benign proxies.

8.3. Detection Cue Taxonomy

The synthesis stream is not a single model but a fusion of complementary cue families, each with a distinct failure mode under laundering T and under the generation advances of §3.2. Diversity is the point: an adaptive adversary who defeats one cue rarely defeats all.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Spatial / learned.** A CNN or ViT trained on real-vs-generated transfers surprisingly well across generators [17]; CLIP/DINOv2 features give the strongest open-set generalization [19,24]. Weakness: distribution shift to unseen samplers.
- **Frequency.** Up-sampling leaves periodic spectral peaks [18]; azimuthal $P(f)$ and co-occurrence/texture statistics are compact and fast. Weakness: JPEG and resize attenuate high-frequency evidence.
- **Reconstruction.** DIRE (diffusion inversion) [16] and ARE (autoencoder, §6.7) [54] key on the output lying on the generative manifold; strong cross-model generalization. Weakness: purification [33] partially closes the gap.
- **Physiological / physical (video).** Remote photoplethysmography detects the absent blood-flow signal in synthetic faces [55]; blink rate, gaze, and lighting/geometry consistency add orthogonal video cues. Weakness: higher-fidelity video closes some gaps.
- **Provenance.** Watermark/C2PA decode (§13) is near-certain *when present*; absent for open local inference.

Table 11 summarizes the trade space.

Table 11. Detection cue families, strength, and primary weakness.

Cue family	Signal	Primary weakness
Learned feat. [19,24]	CLIP/DINOv2 real-vs-gen	Novel-sampler shift
Frequency [18]	Up-sample spectral peaks	JPEG/resize
Reconstruction [16,54]	On-manifold low error	Purification
Physiological [55]	rPPG, blink, gaze (video)	Fidelity gains
Provenance [25,26]	Watermark/manifest	Absent locally

8.4. Training Without Contraband

- **Synthetic-vs-real objective.** The synthesis stream needs only "real photo" vs. "model output" labels—both legally obtainable in bulk—yet transfers to the contraband case because it keys on *generation artifacts*, not sexual content [16,17].
- **Benign proxies for the content stream.** Age estimation and nudity detection are trained on legal datasets; the *conjunction* (minor $\wedge$ sexual) classifier is fine-tuned and validated *only* by accredited bodies on lawfully held data, then distributed as frozen weights to field tools.
- **Federated / split learning.** Where multiple agencies hold data, federated learning with secure aggregation lets a shared model improve without any party exporting images; gradients are clipped and DP-noised (Algorithm 4). To be a *guarantee* rather than a heuristic, the DP claim must be fully specified, which we state as a requirement, not an achieved bound: the *adjacency* relation and *protected unit* (we target example-level/per-image DP, and note that per-person or per-agency protection needs group-DP with the group size folded into the accounting), the per-round mechanism, the *composition across rounds* via a formal accountant (e.g. Rényi/moments), and the reported overall budget (ε, δ) with the sampling assumptions it depends on. Absent this accounting, clipping-plus-noise provides empirical resistance to reconstruction, not a certified privacy level; and "frozen weights" limit but do *not* prove the absence of memorization, so distributed models additionally require a memorization audit (§6) before field release.
- **Curated-benign public track, suspect-stays-enclave.** The public research track ingests *only* pre-curated benign corpora (legal adult, scene, and fully synthetic imagery); it is never fed field-collected material. Because a hash non-match does not certify an item as safe (Property 1), any field-collected or unverified candidate remains inside the accredited enclave *regardless* of its hash result—hashing routes known-bad items but is never used to release the remainder into the public track.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Algorithm 4 DP federated update across accredited enclaves

Require: enclaves $1..N$ with local data $\mathcal{D}_i$; global model θ; clip C; noise σ

for round $= 1, 2, \ldots$ **do**
  for each enclave i **in parallel do**
    $g_i \leftarrow \nabla_\theta \mathcal{L}(\theta; \mathcal{D}_i)$ ▷ data never leaves enclave
    $\tilde{g}_i \leftarrow g_i / \max(1, \|g_i\|_2 / C)$ ▷ clip
  end for
  $\bar{g} \leftarrow \frac{1}{N} \sum_i \tilde{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ ▷ secure-agg + DP noise
  $\theta \leftarrow \theta - \eta\, \bar{g}$
end for
return θ ▷ distributed as frozen weights

8.5. Robustness to Laundering and Adaptive Evasion

Detectors must survive the full channel T: JPEG, resize, blur, screenshot, and diffusion *purification* [33]. We adopt: augmentation with the full laundering chain at train time; test-time multi-crop/multi-scale fusion; and adversarial training against gradient-based evasion by solving the inner maximization

$$\max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}\big(f_\theta(T(x+\delta)), y\big) \tag{16}$$

within the training loop. Our evaluation *protocol* requires reporting under this *adaptive* threat model, not just clean accuracy; clean-only numbers systematically overstate field performance (Section 19). We specify the protocol here; we do not report measured numbers, and the ROC curves in Fig. 6 are illustrative placeholders.

9. Illustrative End-to-End Walkthrough (Benign)

To make the pipeline concrete without touching illegal material, consider a benign triage scenario: an analyst receives a batch of images suspected to be AI-generated portraits distributed in a harassment campaign (adult NCII proxy). The walkthrough exercises every stream using only legal data.

1. **Ingest & hash-gate.** Each image is cryptographically and perceptually hashed; any match to an accredited corpus is routed to the enclave, the rest proceed in the research pipeline.
2. **Synthesis stream.** DIRE (Alg. 1) yields a low reconstruction error and a radial spectrum peak consistent with a latent-diffusion up-sampler $\Rightarrow D(x) = \mathcal{G}$ with high confidence.
3. **Attribution.** Fusion (Alg. 2) assigns the batch to a single base-model family with a recoverable LoRA-style latent fingerprint; two outliers abstain ($\varnothing$, open-set), flagged for manual review.
4. **Provenance.** No C2PA manifest and no decodable watermark is *consistent with* local, unmoderated inference (tier T1–T2), but it is weak and non-specific evidence: legitimate images routinely lack these marks, and manifests are easily stripped. Absence lowers priors only slightly and is used, if at all, as a soft tie-breaker—never as a positive indicator of abuse.
5. **Fusion & bundle.** The system emits a calibrated triage score, the per-cue breakdown, saliency overlays, and an immutable evidence bundle (input hash, model-weight hash, scores, timestamps) for downstream admissibility.

The same control flow applies to the contraband case, with the content stream executed inside the enclave; the researcher-facing path is unchanged, demonstrating the no-possession property end to end.

10. Deepnude and Deepfake Identification Pipeline

Generic "is it synthetic?" detection (§8) is necessary but insufficient for the two headline abuse classes. A deepnude is often *mostly real pixels* with a synthesized intimate region; a face-swap deepfake is a real body/scene with a grafted face. Identification therefore adds three questions beyond synthesis: *where* is the manipulation, *whose* identity is implicated, and *what* is depicted. We specify a dedicated

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Algorithm 5 Deepnude / swap identification (no-possession)

Require: artifact x; optional source x_{src}; enclave flag
$\hat{M} \leftarrow \textsc{LocalizeEdit}(x)$ ▷ §6.5 heatmap
$s_{\text{syn}} \leftarrow$ synthesis score on $\hat{M}$ and on x
$s_{\text{exp}} \leftarrow$ exposure estimator on region $\hat{M}$ ▷ benign-trained
if x_{src} given **then**
    $s_{\text{id}} \leftarrow \text{sim}(\psi(x_{\text{face}}), \psi(x_{\text{src}}))$ ▷ NCII linkage
end if
if enclave **and** s_{exp} high **then**
    $s_{\text{minor}} \leftarrow$ conjunction classifier ▷ accredited-only
end if
return $\hat{M}$, calibrated $\{s_{\text{syn}}, s_{\text{exp}}, s_{\text{id}}, s_{\text{minor}}\}$, evidence bundle

pipeline (Fig. **??** would mirror Fig. 3; omitted for space) that is co-designed with the no-possession constraints of §8.

10.1. Localization: finding the synthesized region

Deepnudes and swaps are *composites*. We run the inpainting-boundary forensics of §6.5—noiseprint discontinuity [52], local spectral step, double-JPEG and DIRE/ARE heatmaps—to estimate the edited mask $\hat{M}$. A whole-image detector reports "90% real"; the localizer reports "the torso region is synthetic," which is both more accurate on composites and directly admissible. For face swaps, blending-boundary artifacts around the face oval and resolution/colour mismatch between grafted face and native skin are the dominant cues.

10.2. Exposure and Content Estimation

A nudity/exposure estimator scores the suspected region. Trained on *legal* adult and synthetic data, it localizes sexual content without any CSAM in the training set. Its output gates—but never alone decides—the minor-risk conjunction classifier, which is enclave-only (§8). Separating "sexual content present" (benign-trainable) from "subject is a minor" (age estimation, benign-trainable) from their *conjunction* (accredited-only) is what keeps the pipeline lawful.

10.3. Identity Linkage (NCII Case)

NCII implicates a *specific* real victim. Given a claimed or candidate source image, a face-recognition embedding ψ tests whether the deepnude's preserved facial region matches the source, $\text{sim}(\psi(x_{\text{face}}), \psi(x_{\text{src}})) > \tau_{\text{id}}$, while the body/intimate region is flagged synthetic by the localizer—the signature pattern of "real face, synthetic body." This linkage is what distinguishes victim-targeted NCII (high urgency, a real person harmed) from victimless-appearing synthetic content, and it drives takedown and victim-notification duties.

10.4. Video Deepfakes

For video, the image cues extend with temporal ones: inter-frame identity drift, boundary flicker around the swapped region, lip-sync phoneme/viseme mismatch for talking-head fakes [45], and the physiological rPPG signal [55]. Temporal aggregation (per-frame scores pooled with a sequence model) markedly outperforms per-frame voting against reenactment [44] and video-diffusion [46] content, because the manipulation artifact is intermittent per frame but temporally correlated.

11. Multimodal and Audio Threats

Abuse is increasingly multimodal, and a forensic program scoped to still images alone will miss a growing share of harm. We outline the adjacent modalities and how the paper's principles extend to them.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Voice cloning. Few-shot text-to-speech and voice conversion can clone a target voice from seconds of reference audio. In the child-safety context this enables voice-based grooming and sextortion (impersonating a peer or a parent) and lends false authenticity to synthetic video. The forensic response mirrors the image case: synthesis detection on the audio (vocoder and prosody artifacts), speaker-identity linkage, and provenance/watermarking of generated audio. The no-possession and calibration principles transfer unchanged.

Audio–visual consistency. When audio accompanies video, cross-modal consistency is a powerful, hard-to-forge cue: lip-sync phoneme/viseme alignment [45], and the correlation between speech prosody and facial motion. A talking-head fake that is individually convincing in each modality frequently fails their *joint* statistics, so fusing audio and visual streams raises the adversary's cost super-additively.

Coordinated and context abuse. Beyond single artifacts, abuse is organized: a cloned face, a cloned voice, and a fabricated backstory combine in sextortion and disinformation campaigns. Attribution clustering (§14) can therefore *generate hypotheses* that extend from "which model" toward "possibly one campaign," grouping artifacts by shared model fingerprints, reused identities, and temporal co-occurrence—a lead to corroborate with external evidence (Table 8), not a conclusion of common origin on its own. Used this way it elevates forensics from per-item verdicts to an intelligence *lead*.

These extensions are deliberately sketched rather than solved; they define the multimodal frontier of the agenda in §22 and show that the decomposition, no-possession, and calibration principles are modality-general.

12. Preemptive Defense: Immunization and Poisoning

Detection is reactive. Preemptive defense aims to make a target image *hard to misuse* before it is ever edited—shifting cost onto the adversary.

12.1. Adversarial immunization (PhotoGuard class)

PhotoGuard adds an imperceptible perturbation δ so a diffusion editor either misreads the image (encoder attack) or steers edits toward a decoy (diffusion attack) [30]. The encoder attack solves

$$\min_{\|\delta\|_\infty \le \epsilon} \left\| E(x+\delta) - z_{\text{target}} \right\|_2^2, \tag{17}$$

driving the VAE encoding $E(\cdot)$ toward a useless target latent z_{target} so inpainting degrades. The stronger diffusion attack backpropagates through the full denoising chain so that any edit of $x+\delta$ converges to a fixed decoy image; formally it solves $\min_{\|\delta\|_\infty \le \epsilon} \mathbb{E}_{M,c} \left\| \text{Inpaint}(x+\delta, M, c) - y_{\text{decoy}} \right\|_2^2$, taking the expectation over masks and prompts so protection generalizes across edits rather than overfitting one. The cost is the backward pass through many denoising steps, which is the practical reason strong immunization does not scale to "immunize everything" and why transferability across models—optimizing δ against a surrogate and hoping it holds on the attacker's actual model—is the crux of whether these defenses survive contact (§12.2). Glaze and Nightshade apply the same adversarial-perturbation idea to *style cloaking* and to *training-time poisoning* respectively [31,32]: Nightshade perturbs images so that models trained on them learn corrupted concept associations.

12.2. Image Cloaking: A Taxonomy

"Image cloaking" spans several mechanisms with different protection targets; conflating them causes the over-selling the next subsection warns against. We organize the family by *what* it protects against.

- **Recognition cloaking.** Fawkes [56] perturbs a face so that a recognition model learns a corrupted template, frustrating unauthorized enrollment—protection against *identification*, not editing. Relevant to NCII because it raises the cost of building a victim-specific model from scraped photos.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Edit immunization.** PhotoGuard [30] (encoder and diffusion-attack variants) targets *inpainting/img2img* of an existing photo—the direct deepnude vector.
- **Personalization poisoning.** Anti-DreamBooth [57] and AdvDM [58] perturb a person's images so that DreamBooth/LoRA fine-tuning on them fails or produces degraded identities—protection against *victim-specific model building* rather than one-off editing, which matters because the identity-adapter advances of §3.2 make model building the cheaper attack.
- **Robust / transferable cloaks.** MetaCloak [59] uses meta-learning over a model pool plus transformation-robust objectives to survive the laundering and purification that defeat single-model perturbations.

The shared math is a constrained perturbation $\min_{\|\delta\|_\infty \le \epsilon} \mathcal{J}(x+\delta)$ where $\mathcal{J}$ encodes the protection goal (corrupt encoding, steer edit to a decoy, poison fine-tuning). The differences are the objective $\mathcal{J}$ and the *threat the cloak assumes*; a cloak optimized against one model or without transformation robustness transfers poorly.

Table 12. Image-cloaking families by protection target.

Family	Protects against	Example / note
Recognition	Face enrollment	Fawkes [56]
Edit immun.	Inpaint / img2img	PhotoGuard [30]
Personaliz. poison	DreamBooth/LoRA build	Anti-DB [57], AdvDM [58]
Robust cloak	Purification/transform	MetaCloak [59]
Training poison	Concept learning	Nightshade [32]

Evaluation protocol. A cloak claim is only meaningful under an adaptive, purification-aware test: report protection success after DiffPure [33], JPEG, upscale, and cross-model transfer (cloak optimized on model A, attacked with model B), never on the single clean model it was tuned against.

12.3. Documented brittleness

These defenses are valuable but must not be oversold.

- **Purification removes perturbations.** DiffPure and simple JPEG/blur/upscale pipelines strip much of the protection [33,34].
- **Cost.** The strong PhotoGuard variant is compute- and memory-heavy per image, limiting any "immunize everything" deployment.
- **False sense of security.** Independent evaluation finds protected images are often still editable to match the prompt; some newer attacks exploit the perturbation itself. Immunization is *friction*, not a guarantee.

Implication: immunization belongs in a *portfolio*—platform-side opt-in protection for high-risk uploads (e.g., minors' images on social platforms)—paired with detection and provenance, never as a standalone solution. Newer directions (optimization-free universal immunization, protective backdoors [34]) trade per-image cost for robustness and deserve forensic evaluation.

12.4. Noise as a double-edged tool

The same perturbation machinery that *protects* victims can *evade* detectors in adversarial hands. A forensic program must therefore (a) treat an "immunized-looking" artifact as a detection signal in itself (perturbation residue is detectable), and (b) adversarially train detectors against the public perturbation methods. This duality is why Sections 8 and 12 are co-designed rather than independent.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Table 13. Preemptive defenses: mechanism, target, and known weakness.

Method	Mechanism / target	Known weakness
PhotoGuard [30]	Perturb source vs. editing	Purification; heavy per-image cost
Glaze [31]	Cloak artistic style	Robustness contested under transforms
Nightshade [32]	Poison training concepts	Needs adoption at scale; mitigable
Universal imm. [34]	Optimization-free / backdoor	New; limited independent eval

12.5. A Program to End or Hamper Deepnude Generation

No single mechanism stops synthesis, but a coordinated program can make victim-specific deepnude generation materially harder and riskier. We propose five layers, ordered from the victim outward, each raising the adversary cost c_A (§5.2) at a different point in the kill chain.

1. **Immunize images at the source.** Platforms hosting minors' photos opt-in high-risk uploads into edit-immunization (PhotoGuard-class [30]) and recognition cloaking (Fawkes-class [56]), so a scraped photo resists both model-building and direct editing—the protection path already prototyped in open tooling [65]. Deployed with transformation-robust, purification-aware objectives [59], not the brittle single-model variants.
2. **Poison the personalization step.** Images released publicly can carry Anti-DreamBooth/AdvDM perturbations [57,58] so that fine-tuning a victim-specific adapter on them fails—attacking the cheapest targeted-NCII route.
3. **Provenance by default.** Hosted generators sign and watermark output (§13) so lawful synthetic content is fast-cleared and unmarked content is itself a triage flag.
4. **Harden the supply chain.** Uncensored checkpoints proliferate through open hubs (§5); hub-level scanning, lineage tracking, KYC for high-risk capabilities, and takedown raise distribution cost and attribution risk.
5. **Deter via attribution.** Model parsing and fingerprinting (§6.8) make "anonymous" generation traceable to a toolchain and, with provenance, to an account—shifting the adversary's expected cost from compute to exposure.

The honest bound (§7.8) remains: against a determined open-weights offender running locally, layers (1)–(2) are friction, not a wall, and (3)–(5) bite hardest on the *cooperating and hosted* ecosystem. The realistic goal is to shrink the casual and at-scale attack surface, raise cost and traceability everywhere else, and detect what still gets through.

13. Provenance and Watermarking

Provenance shifts proof from "spot the fake" to "verify the real"—often the more tractable problem at scale. Conceptually it *injects* side information at generation time, which (§7.8) is categorically different from racing the shrinking natural detection signal: a watermark raises $I(Y;X)$ by construction. Two sub-mechanisms carry distinct threat models. A *signed manifest* binds metadata to content with a public-key signature, so tampering is detectable but *stripping* the manifest is trivial and leaves no trace—provenance-by-signature answers "is this attested?" not "is this generated?". A *pixel watermark* embeds a decodable payload in the content itself, surviving re-encoding but facing removal and forgery attacks [29]. Robustness must therefore be stated against a specified adversary (which transforms, which query access), exactly as for detection.

- **C2PA** cryptographically signed manifests record capture/edit/generation as tamper-evident "ingredients" [25]. It verifies the *integrity and origin of the provenance assertions*—who signed what, and whether the manifest was altered—*not* the truth of the depicted content: a validly signed manifest can still describe a misleading image, and an unsigned image is not thereby suspect. Strength: rich,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

verifiable lineage when present. Weakness: metadata is strippable and adoption is partial, so it must be paired with pixel-level marks and never read as a content-truth oracle.

- **SynthID** embeds an imperceptible, decodable watermark reported to survive compression and cropping [26].
- **Stable Signature** binds a message by fine-tuning the LDM decoder [27]; later work shows further fine-tuning can strip it—an availability-vs-robustness tension.
- **Tree-Ring** plants a detectable pattern in the initial noise [28], robust to many transforms but model-specific.

Watermark robustness is actively contested: studies show many schemes can be removed or *forged* under realistic attack [29]. For CSAM specifically, watermarking the *generators* is effective only for cooperating (licensed, hosted) models; it does nothing for open local inference. Its real forensic value is therefore (1) rapid exculpation/triage of *marked legitimate* synthetic content, and (2) attribution leverage over hosted services that *should* have marked their output—i.e., accountability, not prevention.

Table 14. Provenance/watermark approaches and forensic role.

Approach	Layer	Forensic role / limit
C2PA [25]	Signed metadata	Rich lineage; strippable
SynthID [26]	Pixel watermark	Survives compress/crop; hosted-only
Stable Sig. [27]	Decoder-bound	Fine-tune can strip
Tree-Ring [28]	Noise-space	Robust; model-specific; forgeable [29]

14. Operational Integration with Law Enforcement

Technology only matters if it plugs into existing interdiction.

Table 15. Operational toolset mapping (defensive, as publicly documented).

Capability	Representative system	Role / gap vs. synthetic
Perceptual hash	PhotoDNA; ICSE; Arachnid [6,7]	Retrospective; blind to novel synthesis
Near-dup / crawl	Project Arachnid [7]	Finds re-shares; misses first-seen
Classifier triage	CSAI Match; Krunam; AviaTor	Prioritizes unknowns; needs synthesis-aware upgrade
Victim ID	INTERPOL ICSE	Weakened when no real victim
Synthesis forensics	*proposed here*	DIRE + attribution + provenance

14.1. Workflow Insertion Points

1. **Intake triage.** Run the synthesis + content streams first to *prioritize* human review and reduce examiner exposure—a welfare as well as throughput gain.
2. **Attribution for de-confliction.** Cluster artifacts by model family/fine-tune fingerprint to link cases, identify prolific sources, and distinguish synthetic-only from mixed photographic cases (the latter implying a real victim and higher urgency).
3. **Provenance pre-filter.** Validate C2PA/watermarks to fast-track clearly marked legitimate content.
4. **Hash feedback.** Confirmed synthetic abuse is itself hashed and shared so the retrospective layer catches re-shares.

Fig. 4 places these steps in sequence.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

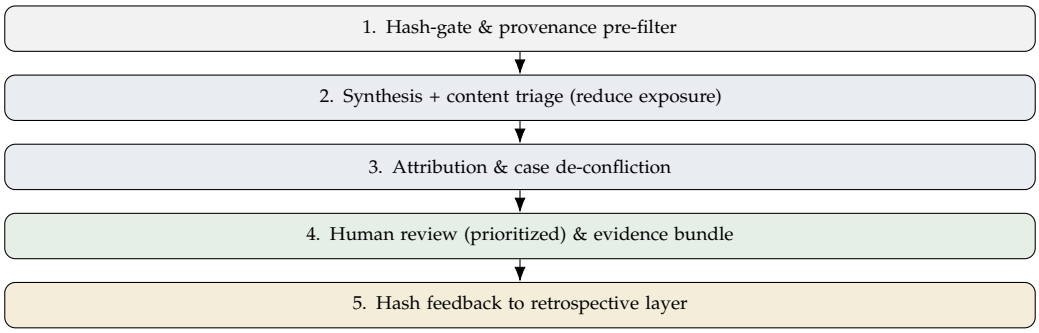


Figure 4. Investigator workflow with pipeline-forensics insertion points.

14.2. Forensic Admissibility and Chain of Custody

For a classifier's output to support *evidence*, several properties help, though admissibility is ultimately decided by courts under jurisdiction-specific rules and is not something a system design can guarantee in the abstract. The properties we target: (1) *reproducibility*—fixed model versions, seeds, and hashes of weights recorded; (2) *calibrated, documented error rates* (ROC/AUC, FPR at operating point), the kind of information Daubert [37]/Frye-style scrutiny tends to weigh in jurisdictions that apply them; (3) *explainability*—saliency and the per-cue attribution breakdown exhibited; (4) *immutable logging* of the evidence bundle (input hash, model hash, scores, timestamps) under standard chain-of-custody. As a design heuristic rather than a legal rule, an opaque "high-accuracy" score is far weaker than a versioned, calibrated, explained, logged pipeline. This shapes the architecture: Algorithm 2 and the evidence bundle in Section 9 exist to support these properties, not only to maximize accuracy; whether they suffice in a given case is for the court to decide.

14.3. Triage Scoring and Case Linkage

The operational output is not a binary verdict but a *ranked queue*. Each artifact carries a calibrated triage score combining synthesis confidence, content-risk, attribution certainty, and provenance status; the queue is ordered to put the highest-confidence, highest-harm, real-victim-likely items in front of a human first. Two signals re-order aggressively: a *real-face/synthetic-body* composite (§10) implies a real victim and is escalated; a confident *marked-legitimate* provenance decode is de-prioritized. Case linkage then clusters the queue by model-family and fine-tune fingerprint, so a single adjudication can resolve a batch and prolific sources surface as clusters rather than isolated items. The metrics that matter here—precision@k, time-to-triage, and exposure reduction (§19)—are operational, not just statistical, and should be reported alongside AUC.

14.4. Examiner Welfare

Human examiners who review abuse material suffer documented psychological harm. Pipeline forensics has a direct welfare dividend: automated triage reduces the volume and salience of material a human must view, blurred-preview and region-masked review modes can be driven by the content stream, and attribution clustering lets one decision cover many artifacts. We argue welfare should be an explicit design objective and evaluation metric (Section 19), not a side effect.

15. Legal Landscape (Multi-Jurisdiction)

Synthetic CSAM is illegal in many jurisdictions even absent a real victim, and NCII law is expanding rapidly. Table 16 summarizes representative instruments; it is a technical orientation for system designers, not legal advice, and statutes evolve.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Table 16. Representative legal instruments (orientation only; verify current text).

Jurisd.	Instrument	Relevance
US	PROTECT Act (2003) [38]; TAKE IT DOWN Act (2025) [39]	Obscene virtual CSAM; NCII incl. AI deepfakes, takedown duties
EU	Proposed CSA Regulation [40]	Detection/reporting duties for services
UK	Online Safety Act (2023) [41]	Platform duties; illegal-content codes
BR	ECA arts. 240–241 [42]; LGPD; Lei 12.737/12.015	Production/possession of child porn; data protection; cyber offenses
Intl	INTERPOL / national referrals	Cross-border victim ID and takedown

15.1. Jurisdictional Analysis

The instruments differ in ways that directly constrain system design. In the **United States**, the PROTECT Act criminalizes obscene virtual depictions [38] after earlier overbreadth jurisprudence narrowed a purely "virtual" prohibition, so U.S. deployments must reason about the obscenity and indistinguishability tests rather than assume all synthetic depictions are per se illegal; the 2025 TAKE IT DOWN Act adds NCII—expressly including AI-generated imagery—and platform takedown timelines [39]. The **European Union**'s proposed CSA Regulation [40] would place detection and reporting duties on services, raising the privacy-preserving-matching questions of §16.1 to a legal requirement and making on-device, minimal-disclosure designs a compliance strategy rather than an option. The **United Kingdom**'s Online Safety Act [41] imposes duties of care and illegal-content codes enforced by a regulator, favoring auditable, documented pipelines. **Cross-border** cases route through INTERPOL and national central bureaus, where attribution evidence must be packaged for mutual legal assistance.

In **Brazil**—the operational context of NuDetective and the tooling of §7.6—the Estatuto da Criança e do Adolescente (ECA), as amended in 2008, criminalizes producing, possessing, and storing child-pornographic material (arts. 240–241), which is precisely what motivated crime-scene triage tools [42,63]; synthetic depictions are reached through the "simulated" and morphing provisions, and the LGPD (data protection) plus the cyber-offense statutes (Leis 12.737/2012 and 12.015/2009) shape how seized media may be processed. Brazil thus illustrates the general lesson: possession law makes the no-possession, enclave design (§17) a legal necessity, not a convenience.

A recurring tension cuts across all jurisdictions: *detection duties* press toward scanning more content, while *possession and privacy* law presses toward holding and transmitting less. The no-possession, tiered architecture (§17) is, in part, an answer to that tension.

15.2. Design Consequences

Concretely: (i) *possession* rules force the no-possession architecture; (ii) *reporting* duties favor on-device detection with minimal data transfer; (iii) *takedown* timelines favor fast triage and provenance pre-filtering; (iv) cross-border cases require attribution evidence packaged for mutual legal assistance; and (v) divergent definitions of what counts as unlawful synthetic content mean a global platform must make the *jurisdiction* a parameter of its policy engine, not a hard-coded constant.

16. Cybersecurity Hardening of the Model Supply Chain

Abuse models are also a *supply-chain* problem, and the defensive program is itself an attack surface.

- **Deserialization risk.** Legacy `pickle`-based checkpoints execute arbitrary code on load; mandate `safetensors`, sandboxed loading, and scanning of hub artifacts.
- **Hub hygiene.** Continuous metadata scanning, lineage graphing, and automated takedown workflows; sign and pin the weights the defense itself depends on.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Inference-gateway controls.** For hosted services: input/output moderation, rate limits, KYC for high-risk capabilities, provenance marking by default, and abuse telemetry.
- **Detector security.** Protect the detector from poisoning (curated, hashed training data), model extraction (query limits, output calibration), and adversarial evasion (adversarial training, ensembling).
- **Data governance.** Enclave isolation, DP in federated training (Alg. 4), strict access logging, and secure deletion—handling contraband is a confidentiality and legal-exposure risk, not just an ML task.
- **Red-teaming.** Periodic authorized evaluation of the whole stack (including the hash and provenance layers) under adaptive adversaries [35].

16.1. Privacy-Preserving Matching and Its Attack Surface

On-device scanning must reconcile two hard requirements: match candidate media against a known-bad corpus *without* the client learning the corpus and *without* the server seeing non-matching user content. The canonical design is a perceptual hash (e.g., a learned NeuralHash) composed with private set intersection and threshold secret sharing, so the server decrypts only once enough matches accrue [60]. Two failure modes are forensically decisive and must be designed around:

- **Adversarial collisions.** Learned perceptual hashes are differentiable, so an adversary can craft a benign image that collides with a known-bad hash (a false-positive flooding / framing attack) or, conversely, perturb contraband to *evade* the hash [61]. Robust deployment needs a non-differentiable or server-side second-stage confirmation and rate-limited human adjudication.
- **Scope creep and corpus integrity.** The known-bad corpus is a trusted root; its signing, multi-party governance, and auditability are security controls, not policy footnotes. Attribution evidence inherits the corpus's chain of custody.

The defensive lesson is that privacy-preserving matching is a *cryptographic* and *adversarial-ML* problem simultaneously: the PSI protects privacy, but the hash function is the attack surface, and both must be hardened together.

Table 17. Supply-chain hardening checklist.

Surface	Control
Checkpoint load	safetensors only; sandbox; AV/secret scan
Model registry	Signed, pinned, lineage-tracked weights
Training data	Hashed, scrubbed, provenance-tracked
Federated exchange	Secure aggregation + DP; gradient clipping
Inference API	Moderation, rate-limit, KYC, default provenance
Detector itself	Anti-poisoning, anti-extraction, adv. training
Evidence store	Append-only, access-logged, chain-of-custody

17. Deployment Reference Architecture

The theory and algorithms of this paper become operational only inside a deployment that respects the legal, privacy, and scale constraints already established. We give a conceptual reference architecture in three tiers, distinguished by *what data each tier may hold* and *where each model runs* (Fig. 5).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

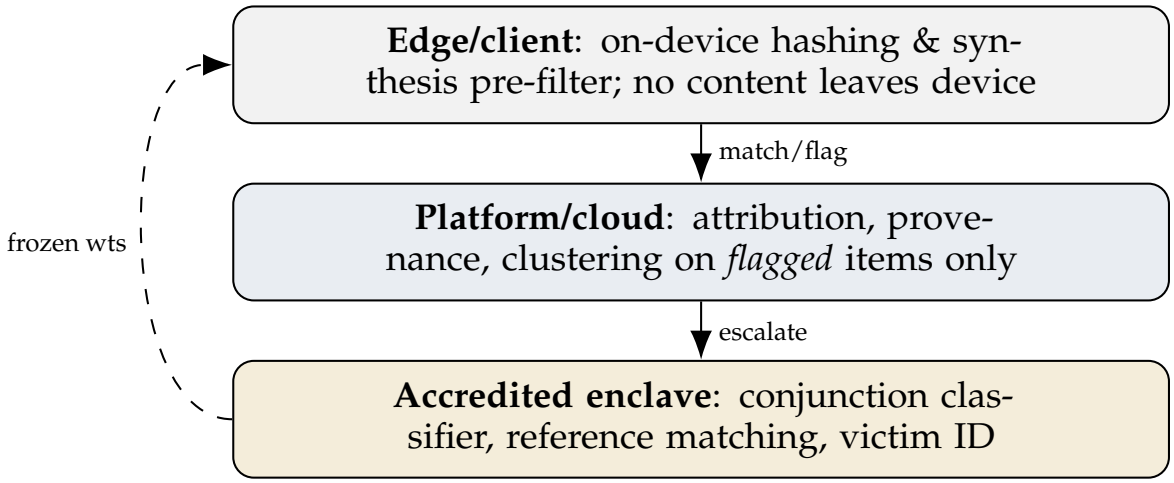


Figure 5. Three-tier deployment: data sensitivity rises downward while data volume falls; models trained in the enclave are distributed outward as frozen weights.

Edge/client tier. Perceptual hashing and the synthesis pre-filter run on-device so that non-matching content never leaves the user's device (the data-minimization principle behind privacy-preserving matching, §16.1). Only a match or a high synthesis score produces a signal.

Platform/cloud tier. Flagged items undergo attribution, provenance validation, and family clustering. This tier never needs the *conjunction* (minor ∧ sexual) decision; it operates on the legally obtainable synthesis/attribution signals, and it amortizes one attribution decision over many artifacts for case de-confliction (§14).

Accredited-enclave tier. The conjunction classifier, reference near-duplicate matching, and victim identification run only here, on lawfully held data, under the DP federated protocol of Algorithm 4. Models trained in the enclave flow outward as *frozen weights*, closing the loop without proliferating contraband. This tiering is the architectural expression of the no-possession principle (P1–P4, §8): capability is distributed, data is contained.

18. Ethics, Governance, and Dual-Use Control

Because this work sits at a dual-use boundary, its governance is part of the contribution, not a disclaimer. We make the controls explicit so they can be adopted and audited.

Dual-use minimization. Every mechanism here is presented at the level of *category and consequence*. We withhold abuse-enabling specifics—no named checkpoints, prompts, fine-tuning recipes, or filter-bypass procedures—following the norm established for safety-filter red-teaming [35]. The asymmetry is deliberate: detection and attribution detail aids defenders far more than attackers, whereas generation detail aids attackers more, so the two are disclosed asymmetrically.

No-possession as an ethical control. The decomposition of §7.1 is what lets researchers contribute without handling contraband: explicitness and synthesis are learnable on legal proxies, and only their conjunction with minority is enclave-restricted. This is an ethics mechanism encoded in the architecture, not merely a policy.

Research governance. Work in this area should carry institutional/ethics-board approval, a documented data-handling and incident-response plan, access logging, and secure deletion; handling real material is reserved to accredited bodies under their legal mandates. Negative results—what evasion defeats—are reported rather than suppressed, because hidden failure modes become field surprises.

Examiner welfare and bias. Governance includes the humans in the loop: automated triage to reduce exposure (§14) and explicit auditing for demographic bias in age and explicitness estimation, since a biased classifier both misses harm and inflicts it. Fairness evaluation across skin tone and apparent age is a required construct-validity check (§20).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

19. Evaluation Methodology and Benchmark Design

We specify a protocol executable *without* illegal data by researchers, extensible to enclave validation by accredited bodies.

19.1. Datasets (benign / proxy)

- **Real:** large legal photo corpora (faces, bodies, scenes).
- **Synthetic:** outputs from a panel of *legitimate* open and hosted models spanning GAN, LDM, and newer architectures—labeled by source for attribution evaluation.
- **Laundering suite (T):** JPEG (multiple quality factors), resize, crop, blur, screenshot re-capture, DiffPure.
- **Adversarial:** PhotoGuard/Glaze-perturbed benign images to test evasion and robustness.

The *conjunction* (minor $\wedge$ sexual) classifier is validated only on accredited, lawfully held data; researchers evaluate the surrounding streams on proxies. For the synthesis-detection and attribution streams, public synthetic-image benchmarks (families of real-vs-generated corpora spanning GAN, latent-diffusion, and newer samplers, with laundering variants) supply a lawful training and evaluation base entirely disjoint from the contraband region, so the D and A tasks are developed openly while only C is enclave-gated. The crucial split discipline is *by generator family*: train on families $\mathcal{M}_{\text{tr}}$, test on disjoint $\mathcal{M}_{\text{te}}$, because an adversary always uses a model the detector has not seen (§3.2); a split that leaks a family across train and test reports an optimistic number with no field validity.

19.2. Metrics

We fix the metric formalism so results are comparable across groups. From the confusion counts (TP, FP, TN, FN),

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}, \quad \text{Prec} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{18}$$

with $F_1 = 2\,\text{Prec}\cdot\text{TPR}/(\text{Prec}+\text{TPR})$ and AUC the area under the ROC TPR(FPR). Because the domain is deployed at platform scale with asymmetric costs (§7.7), the headline detection metric is **TPR at a fixed low FPR** (10^{-3}), not accuracy, which is misleading under heavy class imbalance. Calibration is quantified by the Expected Calibration Error over B confidence bins,

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\, \text{acc}(S_b) - \text{conf}(S_b) \,\right|, \tag{19}$$

reported with a reliability diagram. **Attribution** uses closed-set accuracy, an open-set AUROC that scores the "unknown" decision, and a per-family confusion matrix. **Localization** (composites, §7.5) uses mask IoU and boundary F_1. **Operational** metrics are examiner-exposure reduction and triage precision@k. We require that *clean*, *post-laundering*, and *adaptive-evasion* curves all be reported (Fig. 6); reporting clean-only, or accuracy in place of TPR@FPR, is treated as a methodological defect.

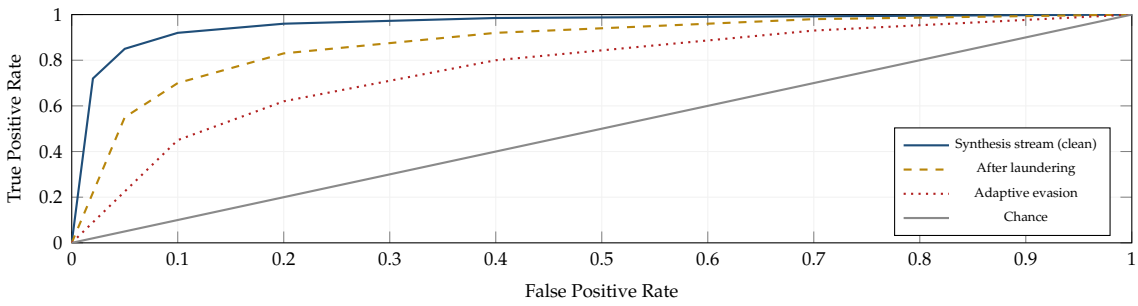


Figure 6. Mandatory reporting regime: clean, post-laundering, and adaptive-evasion ROC must all be published. Curves are schematic placeholders for the protocol, not measured results.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

19.3. Ablations and statistical rigor

We prescribe ablations over each stream (synthesis-only, +provenance, +hash, full fusion) and over the laundering suite, with confidence intervals from bootstrap resampling and significance tests between configurations. Attribution is reported as a confusion matrix over families plus open-set rejection rate. Cross-generator generalization (train on families $\mathcal{M}_{\text{tr}}$, test on disjoint $\mathcal{M}_{\text{te}}$) is the headline number, because real adversaries use unseen checkpoints.

19.4. Reproducibility and Ethics Controls

Record model/weight hashes, seeds, and library versions; publish code and benign data; gate contraband-touching components behind accredited access; obtain institutional/ethics approval; and document a data-handling and incident-response plan. Negative results—what evasion defeats—are reported, not hidden.

20. A Methodological Framework for Research

Beyond a specific system, this paper intends to give the field a reusable research scaffold—a way to pose, test, and report work on synthetic-abuse forensics that is comparable across groups and defensible in court. We organize it as research questions, a hypothesis-and-design template, threats to validity, benchmark governance, and a capability-maturity model.

20.1. Core Research Questions

We frame the agenda as five questions, each mapped to a measurable construct:

RQ1 Detection.
How accurately, and at what platform-scale FPR, can synthesis be detected under laundering T and adaptive evasion? *(Constructs: TPR@FPR, AUC, robustness curves.)*

RQ2 Attribution.
Can an artifact be traced to a model family, including *open-set* "unknown"? *(Open-set AUROC, confusion matrix, rejection rate.)*

RQ3 Localization.
Can the manipulated region be delineated on composites (§7.5)? *(Mask IoU, boundary F1.)*

RQ4 Prevention.
Do immunization/provenance measurably raise adversary cost and survive purification? *(Protection-success-after-attack, compute cost.)*

RQ5 Operations.
Does the pipeline reduce examiner exposure and time-to-triage without unacceptable error? *(Exposure reduction, precision@k, calibration.)*

20.2. Hypothesis and Experimental-Design Template

A well-formed study here states: (i) a falsifiable hypothesis tied to one RQ (e.g. "reconstruction fusion improves cross-generator TPR@10^{-3} over a CLIP baseline"); (ii) the generator panel and the *disjoint* train/test split by model family (cross-generator generalization is the headline, §19); (iii) the full laundering and adaptive-evasion regime as a fixed factor, not an optional robustness appendix; (iv) a pre-registered primary metric and operating point; and (v) ablations isolating each cue. Reporting clean-only accuracy, or testing on the same family trained on, is treated as a design defect rather than a positive result.

20.3. Threats to Validity

Explicit threat-to-validity analysis should accompany every result (Table 18): *internal* (shortcut learning, data leakage across splits), *external* (generalization to unseen generators, demographics, and real field distributions), *construct* (does the benchmark's "synthetic" label capture the operational

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

threat?), and *conclusion* (statistical power, multiple-comparison control). The contraband-data constraint adds a domain-specific validity gap: proxy-trained components must be validated in enclaves before operational claims.

Table 18. Threats to validity and standard mitigations.

Type	Threat	Mitigation
Internal	Shortcut/leakage	Family-disjoint splits; shortcut audits
External	Unseen generators/demographics	Cross-generator & cross-demographic tests
Construct	Proxy $\neq$ operational threat	Enclave validation; proxy justification
Conclusion	Underpowered, *p*-hacking	Pre-registration; CIs; corrections

20.4. Benchmark Governance

A durable benchmark for this domain needs more than data: a versioned generator panel with recorded weights/seeds; a fixed laundering suite; an open-set protocol with held-out families; a leaderboard that reports the robustness and calibration curves (not a single accuracy); and an ethics gate separating the public, benign-proxy track from the accredited-enclave track. Governance—who curates, signs, and audits the corpus—is a methodological requirement because attribution evidence inherits the benchmark's chain of custody (§16.1).

20.5. A Capability-Maturity Model

To let agencies and platforms self-assess and to give researchers targets, we propose a five-level maturity model (Table 19), from retrospective hashing only (L0) to a fully integrated, calibrated, enclave-backed, provenance-aware forensic program (L4). The levels are cumulative: each adds a defensive layer from Fig. 1 without discarding the previous one.

Table 19. Capability-maturity model for synthetic-abuse forensics.

Lvl	Capability	Gap
L0	Perceptual hashing only	Blind to novel synthesis
L1	+ synthesis detection (triage)	No attribution; clean-only
L2	+ attribution & localization	No prevention; weak calibration
L3	+ provenance & immunization	Not enclave-integrated
L4	+ enclave, calibration, admissibility, welfare	Research frontier (open-set)

20.6. Reporting Standard

A paper or deployment is *complete*, in this framework, when it reports: RQ and hypothesis; generator panel and splits; clean, laundering, and adaptive curves; open-set behaviour; calibration (ECE, reliability); threats-to-validity; compute cost; and the ethics/enclave protocol. This standard is the bridge from a research result to the admissibility criteria of §14, and it is what we invite the community to adopt so that results compose into cumulative progress rather than incomparable point claims.

21. Discussion and Limitations

Arms race. Every detector invites an adaptive counter; immunization and watermarks are removable; attribution degrades under aggressive laundering. Under the conditioned shrinking-margin hypothesis of §7.8—whenever improving generators actually reduce the real–generated divergence—

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

the recoverable signal falls, so the honest claim is *cost inflation and triage*, not a permanent solution; where that hypothesis fails to hold, so much the better for detection, but we do not assume it does. This is precisely why the strategy also weights provenance (which injects signal) and economics (which raises c_A on cooperating actors) alongside detection.

Generalization. Open-set attribution to genuinely novel architectures remains the hardest open problem [24]; the fast-sampler and flow-matching shifts of §3.2 show how quickly a fixed detector's priors expire.

Dependence on cooperation. Provenance helps only where generators cooperate; open local inference is out of its reach, which bounds how far the raise-c_A levers extend.

Data asymmetry. The no-possession design trades some accuracy for legality; enclave validation is essential and bottlenecked by accreditation, and the proxy-to-operational validity gap (§20) is real and must be measured, not assumed away.

Bias and fairness. Age and explicitness estimation can carry demographic bias; an unaudited classifier both misses harm and inflicts it, so fairness is a limitation to measure, not a footnote.

Measurement. Field prevalence of synthetic CSAM is hard to measure precisely due to legal constraints on data, so operational metrics must be interpreted carefully. These limitations motivate the agenda below rather than undercut the approach: a layered, cost-raising defense is the realistic objective.

22. Future Directions and Recommendations

22.1. Research Agenda

1. **Open-set, laundering-robust attribution** as the central forensic primitive—generalizing to *unseen* model families is the hard, high-value problem [24].
2. **Enclave-native learning:** federated/split training (Alg. 4) and synthetic-data distillation so capability improves without proliferating contraband.
3. **Robust provenance:** watermarks resilient to fine-tune stripping and forgery [27,29]; standardized C2PA adoption across hosted generators.
4. **Immunization at scale:** optimization-free, purification-resistant perturbations [34]; honest robustness benchmarks.
5. **Memorization auditing:** dataset- and model-level tests to pre-empt re-victimization [2,36].
6. **Video and multimodal** extension of DIRE/attribution, following the trajectory of face-video benchmarks [20], including audio–visual consistency and voice-clone forensics (§11).
7. **Fairness-audited detection:** demographic bias evaluation of age and explicitness estimation as a standing requirement, not a one-off study.
8. **A community benchmark and reporting standard** per the methodological framework (§20), so results compose into cumulative progress rather than incomparable point claims.

22.2. Recommendations for Authorities and Platforms

- **Upgrade the stack from matching to forensics:** fund synthesis-aware triage and attribution inside NCMEC/INTERPOL/C3P workflows while retaining hashing for the retrospective case.
- **Mandate provenance by default** for hosted generators and define legal duties for model hubs (scanning, lineage, takedown, KYC for high-risk capabilities).
- **Create accredited enclaves and shared federated models** so smaller agencies gain capability without holding data.
- **Require admissibility-grade pipelines:** versioned, calibrated, explainable, logged—codify this in forensic standards and expert-evidence guidance.
- **Fund examiner welfare** as a first-class objective: automated triage specifically to cut human exposure.
- **Invest in attribution intelligence:** cluster and track prolific model families/fine-tunes as an intelligence product, not just case-by-case evidence.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

- **Support dataset governance:** mandatory CSAM scrubbing and provenance for training corpora, closing the LAION-class failure [2].

22.3. Synthesis: Layers, Tiers, and Cost

Table 20 draws the whole program together, mapping each defensive layer to the adversary tier (§3) it reaches and the economic term (§5.2) it moves. The pattern is the paper's central strategic claim in one view: prevention and chokepoint layers bite on the casual and commercial tiers and raise c_A; detection, reverse engineering, and attribution reach every tier and lower c_D; and no row covers all tiers alone, which is why the defense must be layered.

Table 20. Defensive layers mapped to adversary tiers (T0–T4) and the cost term moved.

Layer	Reaches	Cost effect
Hashing (retrospective)	re-shares	$c_D \downarrow$ (known)
Synthesis detection	T0–T4	$c_D \downarrow$
Reverse eng. / parsing	T1–T4	$c_D \downarrow$, deter
Attribution	T1–T4	deter ($c_A \uparrow$)
Immunization / poison	T0–T2	$c_A \uparrow$
Provenance	T0 (hosted)	$c_D \downarrow$, deter
Chokepoints (NaaS)	T0, T2 (commercial)	$c_A \uparrow\uparrow$
Supply-chain / hubs	T2–T4	$c_A \uparrow$

23. Conclusion

The abuse frontier has moved from *distributing* illegal images to *synthesizing* them on demand—increasingly through turnkey commercial "nudify" services (§5.3)—breaking the perceptual-hash paradigm that underpins global child protection. Extending our prior deepnude-detection work, we gave the theoretical foundations of deep nude detection (decomposing the task into explicitness, manipulation, and subject factors over a shared neural-network substrate) and argued for a shift to *generative-pipeline forensics*: detect synthesis, reverse-engineer and attribute it to its source—down to the architecture of an unseen generator (§6.8)—verify provenance, and raise the adversary's cost, all under strict no-possession, admissibility, and examiner-welfare constraints.

Two themes bind the program. First, *the deepnude barrier* (§7.6): content-only tools (NuDetective, NSFW classifiers) measure one factor of a three-factor problem and are blind to synthesis; crossing the barrier requires fusing explicitness with manipulation, attribution, and localization, which modern tooling (DeepNudeDetect, Mary Poppins) begins to do and this paper gives the theory for. Second, *adversarial symmetry*: every property that makes generation controllable—differentiability, a smooth manifold, stable fingerprints, guidance—is a handle for detection, attribution, or disruption, so the same mathematics that builds the threat can be turned to dismantle it. No single mechanism suffices: hashing, synthesis detection, reverse engineering, attribution, immunization, provenance, supply-chain hardening, and commercial-chokepoint enforcement are complementary layers of one defense-in-depth, bounded honestly by the information-theoretic limit that the recoverable signal shrinks as generators improve. The decisive move for authorities is to fund and standardize this layered, forensics-grade program—and to act on the hosted, commercial layer where intervention is most tractable—before the asymmetry widens further.

Acknowledgments: The author thanks collaborators in the digital-forensics and child-protection communities. This work is defensive; all handling of illegal material is reserved to accredited authorities under their legal mandates.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Appendix A Notation

Symbol	Meaning
x, $H \times W \times 3$	image and dimensions
$\mathcal{G}_m$, $m \in \mathcal{M}$	generator and model family
D, A, C	detection, attribution, risk-classification tasks
T	laundering channel (JPEG, resize, DiffPure, ...)
$\mathcal{I}_\phi, \mathcal{R}_\phi$	diffusion inversion and reconstruction
$E(\cdot)$	VAE encoder (latent diffusion)
g_k, w_k	attribution cue and fused weight
ϵ, δ	perturbation budget and perturbation
C, σ (Alg. 4)	gradient clip and DP noise scale
$\varnothing$	open-set "unknown" label

Appendix B Glossary of Key Terms

Attribution
Tracing an artifact to the model family (or user) that produced it; *open-set* attribution permits an "unknown" verdict.

C2PA
Coalition for Content Provenance and Authenticity: a standard for cryptographically signed content-provenance manifests.

Composite
An image mixing real and synthesized regions (e.g. a real photo with an inpainted region or a swapped face); the defining structure of deepnudes.

DIRE
Diffusion Reconstruction Error: the residual between an image and its diffusion inversion–reconstruction, low for generated images.

ARE
Autoencoder Reconstruction Error: a training-free latent-diffusion detection score using VAE reconstruction.

Enclave
An accredited, legally authorized environment in which contraband may lawfully be held and processed.

Immunization
A preemptive adversarial perturbation added to an image to frustrate later malicious editing (e.g. PhotoGuard).

Laundering channel (T)
Transformations (JPEG, resize, blur, screenshot, purification) applied to evade detection/attribution.

NCII
Non-Consensual Intimate Imagery, including AI-generated "deepnudes."

No-possession
A design in which detectors are built and evaluated without researchers holding illegal material.

PRNU
Photo-Response Non-Uniformity: a sensor fingerprint in classical camera forensics; the conceptual template for model fingerprints.

Provenance
Evidence of an artifact's origin and edit history (signed manifests and/or embedded watermarks).

Purification
Diffusion-based removal of adversarial perturbations (DiffPure); defeats many immunization schemes.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

Synthetic CSAM
Model-generated sexual depictions of minors, with no corresponding photographic victim or by re-victimizing a known survivor.

Watermark
An imperceptible, decodable mark embedded at generation time to enable later identification (e.g. SynthID, Stable Signature).

Appendix C Reproducibility Checklist

(i) Weight and dataset hashes recorded; (ii) seeds fixed and logged; (iii) library/CUDA versions pinned; (iv) benign data and code released; (v) contraband-touching components gated behind accredited access; (vi) clean/laundering/adaptive results all reported; (vii) calibration (ECE, reliability) reported; (viii) ethics/IRB approval and data-handling plan documented; (ix) negative/evasion results included; (x) chain-of-custody evidence bundle schema published.

References

1. J. A. dos Santos and D. P. Franco, "Deep nude detection using machine learning techniques," Preprint, Authorea/TechRxiv, 2024.
2. D. Thiel, "Identifying and eliminating CSAM in generative ML training data and models," Stanford Internet Observatory, Stanford Digital Repository, 2023.
3. M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.
4. C. Platzer, M. Stuetz, and M. Lindorfer, "Skin sheriff: A machine learning solution for detecting explicit images," in *Proc. 2nd Int. Workshop on Security and Forensics in Communication Systems*, ACM, 2014.
5. F. Cermelli *et al.*, "State-of-the-art in nudity classification: A comparative analysis," arXiv:2312.16338, 2023.
6. INTERPOL, "Global efforts to identify child abuse victims via INTERPOL boosted with Microsoft technology (PhotoDNA integration into the ICSE database)," 2015.
7. Canadian Centre for Child Protection, "Project Arachnid: online availability of child sexual abuse material," Technical report and platform documentation, 2017–2024.
8. Internet Watch Foundation, "How AI is being abused to create child sexual abuse imagery," IWF Report, 2023.
9. National Center for Missing & Exploited Children, "CyberTipline data and generative-AI reporting trends," 2024.
10. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.
11. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020, arXiv:2006.11239.
12. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, arXiv:2112.10752.
13. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," arXiv:2106.09685, 2021.
14. N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation," in *Proc. CVPR*, 2023, arXiv:2208.12242.
15. L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020.
16. Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for diffusion-generated image detection," in *Proc. ICCV*, 2023, arXiv:2303.09295.
17. S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot...for now," in *Proc. CVPR*, 2020.
18. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. ICML*, 2020.
19. U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. CVPR*, 2023, arXiv:2302.10174.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

20. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. ICCV*, 2019.
21. Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models," in *Proc. ACM CCS*, 2023, arXiv:2210.06998.
22. N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. ICCV*, 2019.
23. C. Kim, J. Min, M. Jiang, R. Krishnan, X. Fu, and H. Zhao, "WOUAF: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models," arXiv:2306.04744, 2023.
24. D. Cozzolino, G. Poggi, M. Nießner, and L. Verdoliva, "Are CLIP features all you need for universal synthetic image origin attribution?," arXiv:2408.09153, 2024.
25. Coalition for Content Provenance and Authenticity (C2PA), "C2PA technical specification," 2021–2024.
26. S. Dathathri *et al.*, "Scalable watermarking for identifying large language model outputs (SynthID)," *Nature*, 2024; and Google DeepMind, "SynthID" technical documentation, 2023–2024.
27. P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The Stable Signature: Rooting watermarks in latent diffusion models," in *Proc. ICCV*, 2023, arXiv:2303.15435.
28. Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-Ring watermarks: Fingerprints for diffusion images that are invisible and robust," in *Proc. NeurIPS*, 2023, arXiv:2305.20030.
29. M. Saberi, V. S. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi, "Robustness of AI-image detectors: Fundamental limits and practical attacks," arXiv:2310.00076, 2023.
30. H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious AI-powered image editing (PhotoGuard)," in *Proc. ICML*, 2023, arXiv:2302.06588.
31. S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," in *Proc. USENIX Security*, 2023.
32. S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *Proc. IEEE S&P*, 2024, arXiv:2310.13828.
33. W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification (DiffPure)," in *Proc. ICML*, 2022, arXiv:2205.07460.
34. "GuardDoor: Safeguarding against malicious diffusion editing via protective backdoors," arXiv:2503.03944, 2025; see also "Optimization-free image immunization against diffusion-based editing," arXiv:2411.17957, 2024. (Preprints; author lists to be verified against the published versions.)
35. J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the Stable Diffusion safety filter," arXiv:2210.04610, 2022.
36. N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *Proc. USENIX Security*, 2023, arXiv:2301.13188.
37. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).
38. U.S. Congress, "PROTECT Act of 2003" (Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today), Pub. L. 108–21, 2003.
39. U.S. Congress, "TAKE IT DOWN Act," 2025 (non-consensual intimate imagery, including AI-generated, and platform takedown obligations).
40. European Commission, "Proposal for a Regulation laying down rules to prevent and combat child sexual abuse," COM(2022) 209, 2022.
41. United Kingdom, "Online Safety Act 2023," 2023.
42. Brazil, "Estatuto da Criança e do Adolescente (Lei n. 8.069/1990), arts. 240–241 as amended by Lei n. 11.829/2008"; see also Lei n. 12.737/2012 and Lei Geral de Proteção de Dados (Lei n. 13.709/2018).
43. R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. ACM Multimedia*, 2020, arXiv:2106.06340.
44. A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. NeurIPS*, 2019.
45. K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild (Wav2Lip)," in *Proc. ACM Multimedia*, 2020.
46. A. Blattmann *et al.*, "Stable Video Diffusion: Scaling latent video diffusion models to large datasets," arXiv:2311.15127, 2023.
47. H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," arXiv:2308.06721, 2023.
48. Q. Wang *et al.*, "InstantID: Zero-shot identity-preserving generation in seconds," arXiv:2401.07519, 2024.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 11 November 2026 https://doi.org/10.20944/preprints202611.1122.v1

49. A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation (SDXL-Turbo)," arXiv:2311.17042, 2023.
50. P. Esser *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis (Stable Diffusion 3)," in *Proc. ICML*, 2024, arXiv:2403.03206.
51. Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery (PEZ)," in *Proc. NeurIPS*, 2023, arXiv:2302.03668.
52. D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 144–159, 2020.
53. G. Ilharco *et al.*, "Editing models with task arithmetic," in *Proc. ICLR*, 2023, arXiv:2212.04089.
54. J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *Proc. CVPR*, 2024, arXiv:2401.17879.
55. U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
56. S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proc. USENIX Security*, 2020.
57. T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. Tran, and A. Tran, "Anti-DreamBooth: Protecting users from personalized text-to-image synthesis," in *Proc. ICCV*, 2023, arXiv:2303.15433.
58. C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: Preventing painting imitation from diffusion models (AdvDM)," in *Proc. ICML*, 2023, arXiv:2302.04578.
59. Y. Liu *et al.*, "MetaCloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning," in *Proc. CVPR*, 2024, arXiv:2311.13127.
60. Apple Inc., "CSAM detection—technical summary" (private set intersection and threshold secret sharing design), 2021.
61. L. Struppek, D. Hintersdorf, D. Neider, and K. Kersting, "Learning to break deep perceptual hashing: The use case NeuralHash," in *Proc. ACM FAccT*, 2022, arXiv:2111.06628.
62. V. Asnani, X. Yin, T. Hassner, and X. Liu, "Reverse engineering of generative models: Inferring model hyperparameters from generated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023, arXiv:2106.07873; see also Meta AI, "Reverse engineering generative models from a single deepfake image," 2021.
63. P. M. S. Eleutério and M. C. Polastro, "NuDetective: A forensic tool to help combat child pornography through automatic nudity detection," in *Proc. 21st Int. Workshop on Database and Expert Systems Applications (DEXA)*, 2010, pp. 349–353, doi:10.1109/DEXA.2010.74.
64. Open-Dating contributors, "adult-image-detector: Use deep neural networks and other algorithms to detect nude images," open-source software, https://github.com/open-dating/adult-image-detector, 2017–2024.
65. J. A. dos Santos, "DeepNudeDetect: Comprehensive NSFW/deepfake detection and protection system," open-source software, https://github.com/JoasASantos/DeepNudeDetect, 2024.
66. J. A. dos Santos, "Mary Poppins: A zero-visual-exposure digital-intelligence and forensic platform for child-protection investigation," technical documentation, 2025.
67. Graphika, "A revealing picture: The rise of synthetic NCII (non-consensual intimate imagery) 'undressing' services," Graphika Report, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.