VisionGuard: Cost-Sensitive AI Attestation with Quorum-Verified Blockchain Enforcement

Sundara Srivathsan M.[1], Lighittha P. R.[1], Prithivraj S.[1], Suganya Ramamoorthy[2,*] and Vijayan Sugumaran[3]

[1]*School of Electronics Engineering, Vellore Institute of Technology, Chennai, India*
[2]*School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*
[3]*Oakland University, Rochester, Michigan, USA*
E-mail: suganya.ramamoorthy@vit.ac.in
[*]*Corresponding Author*

Received 14 December 2025; Accepted 06 January 2026

Abstract

Web3 platforms face a critical challenge: once unsafe content is minted on-chain, it becomes immutable and irrevocable. Traditional NSFW classifiers operate off-chain without cryptographic guarantees, leaving blockchain ecosystems vulnerable to harmful content. We present *VisionGuard*, a unified moderation framework that integrates cost-sensitive AI decision-making with blockchain-based enforcement. Our system combines calibrated NSFW classification, abstention-based triage for uncertain cases, perceptual hashing for near-duplicate detection, and on-chain k-of-n quorum attestation using EIP-712 signatures. We establish formal guarantees for: (i) Bayes-optimal cost-sensitive thresholds minimizing asymmetric error costs, (ii) optimal abstention intervals for human review, (iii) monotone false-negative reduction under classifier-pHash fusion, (iv) quorum compromise bounds, and (v) end-to-end unsafe-mint probability. Empirical validation on a zero-shot NSFW task demonstrates 82% accuracy (AUC $=$ 0.88), with the Bayes-optimal threshold ($\tau^* = 0.1$) reducing expected cost to 27,520 versus 54,942 at

Journal of Web Engineering, Vol. 25_1, 103–134.
doi: 10.13052/jwe1540-9589.2516

the F1-optimal threshold—a 50% improvement. Calibrated abstention further lowers harm (cost $= 10,649.5$), while a 3-of-5 quorum with oracle compromise $p = 0.1$ yields break probability $P_{\text{break}} < 1\%$. Together, VisionGuard bridges decision theory, adversarial robustness, and cryptographic enforcement, providing the first provably safe AI moderation pathway for blockchain content.

Keywords: AI safety, NSFW classification, cost-sensitive learning, abstention, perceptual hashing, EIP-712, quorum, blockchain.

1 Introduction

The immutable nature of blockchain systems presents a fundamental challenge for content moderation: once media is minted on-chain, it cannot be revoked or removed. This permanence makes *prevention* the only viable safeguard against unsafe or harmful content in Web3 ecosystems. Unlike traditional web platforms where content can be retroactively moderated, blockchain-based applications, ranging from NFT marketplaces to decentralized social networks-require pre-mint verification to ensure that harmful material never reaches the chain. However, existing moderation solutions suffer from critical gaps that leave these systems vulnerable.

1.1 Limitations of Current Approaches

Conventional not-safe-for-work (NSFW) classifiers operate entirely off-chain and lack cryptographic assurances, creating several attack vectors. First, adversaries can exploit the absence of tamper-proof attestation by replaying previously verified content with altered metadata, or by applying adversarial transformations (e.g., minor perturbations, re-encoding) to bypass hash-based checks. Second, traditional classifiers optimize for accuracy or F1-score without accounting for the *asymmetric costs* of errors: allowing unsafe content on-chain carries far greater harm than blocking safe content. Third, existing systems lack mechanisms to handle uncertain predictions, forcing binary decisions even when model confidence is low.

Beyond classifier limitations, the disconnect between AI-based moderation and blockchain enforcement creates additional vulnerabilities. Perceptual hashing techniques can detect near-duplicate unsafe content but are not integrated with probabilistic classifiers or decision-theoretic frameworks. Blockchain consensus mechanisms provide strong cryptographic guarantees

for transaction validity but have not been adapted to enforce AI-driven moderation policies. These fragmented approaches leave critical gaps in the end-to-end safety pipeline from content submission to on-chain minting.

1.2 Research Gap and Objectives

While advances in zero-shot classification, fairness auditing, and perceptual hashing have each improved aspects of content moderation, no prior framework has integrated these techniques with formal risk analysis and blockchain enforcement. This fragmentation prevents platforms from providing provable guarantees on the probability of unsafe content being minted—a critical requirement for high-stakes Web3 applications.

Our work addresses this gap through four core objectives:

- **Formalize cost-sensitive moderation:** Develop a decision-theoretic framework that minimizes expected harm under asymmetric error costs, rather than optimizing conventional accuracy metrics.
- **Design abstention mechanisms:** Enable classifiers to defer uncertain cases to human review, trading low review costs for avoidance of high-cost mistakes.
- **Integrate perceptual robustness:** Leverage perceptual hashing to defend against adversarial re-uploads and near-duplicate evasion attacks.
- **Enable cryptographic enforcement:** Enforce moderation decisions on-chain through quorum-based attestation, providing tamper-resistant guarantees.

1.3 Contributions

To achieve these objectives, we introduce *VisionGuard*, a unified moderation framework that bridges AI safety theory and blockchain practice. Our contributions span theoretical analysis, system design, and empirical validation:

- **Cost-sensitive risk minimization:** We derive the Bayes-optimal threshold τ^* that minimizes expected cost under asymmetric penalties (C_B, C_H), generalizing classical cost-sensitive learning to the blockchain moderation setting.
- **Optimal abstention theory:** We establish closed-form bounds for the optimal abstention interval, showing when uncertain predictions should trigger human review rather than automated decisions.
- **Robustness via OR-fusion:** We prove that fusing classifier outputs with perceptual hashing via logical OR monotonically reduces false-negative

rates on unsafe content, providing formal guarantees against near-duplicate attacks.
- **Quorum security analysis:** We derive exact probability bounds for k-of-n quorum compromise under independent oracle failure, connecting Byzantine fault tolerance to AI moderation.
- **End-to-end safety guarantee:** We establish a compositional bound on the probability of unsafe content being minted, integrating classifier errors, abstention rates, fusion robustness, and quorum security into a single formula.
- **Blockchain enforcement:** We design and implement `VisionGuard721Quorum`, a Solidity smart contract that enforces quorum-verified moderation decisions using EIP-712 typed signatures, with gas-optimized signature verification.
- **Empirical validation:** We demonstrate that cost-sensitive thresholds reduce expected harm by 50% compared to F1-optimal thresholds, abstention lowers costs by an additional 15%, and 3-of-5 quorums achieve sub-1% compromise probability.

By integrating decision theory, adversarial robustness, and cryptographic enforcement, VisionGuard contributes the first framework with provable end-to-end guarantees for blockchain content moderation.

1.4 Paper Organization

The remainder of this paper is organized as follows: After Related Works, Section 3 establishes notation and decision-theoretic foundations. Section 4 presents the VisionGuard architecture and moderation pipeline. Section 5 derives formal guarantees for each system component. Section 6 formalizes the threat model and security assumptions. Section 7 details smart contract design and hard negative mining. Section 8 validates theoretical predictions through experiments on cost-sensitivity, abstention, fusion, and quorum security. Section 9 discusses limitations and future work, and Section 10 concludes.

2 Related Works

The architecture of a decentralized content moderation system sits at the intersection of computer vision, decision theory, and distributed systems. This section reviews the disparate bodies of literature that inform our proposed framework, moving from efficient visual detection to semantic safety

alignment, and finally to the cryptographic consensus mechanisms required for immutable ledgers.

2.1 Efficient Visual Filtering and Granular Detection

The computational constraints of decentralized nodes necessitate lightweight yet robust detection models. While early approaches relied on static image analysis, recent scholarship has emphasized spatiotemporal features for video content. Yousaf and Nawaz [1] established a significant benchmark by combining EfficientNet-B7 with Bidirectional LSTMs, demonstrating that temporal consistency is crucial for reducing flicker-based false positives in video streams. However, binary classification often fails in the "open world" setting of the permissionless web. Alico et al. [2] addressed this by proposing Deep One-Class Classification (DOC), effectively modeling the manifold of illicit content while rejecting out-of-distribution benign data without the need to model the entire universe of safe content. For granular moderation-essential for privacy-preserving blurring rather than total file removal-Perez et al. [3] provided a comparative analysis of object detectors, validating that lightweight architectures like YOLO can achieve viable accuracy for client-side scanning operations.

2.2 Zero-Shot Semantics and Safety Alignment

The rigid taxonomy of CNNs is increasingly being supplanted by Vision-Language Models (VLMs) capable of zero-shot enforcement. Following the foundational release of CLIP by Radford et al. [9], research has pivoted toward aligning these models with safety standards. Poppi et al. [10] introduced methods to "unlearn" toxic concepts in VLMs, preventing the model itself from becoming a vector for harmful content generation. Building on this, Poppi et al. [11] recently proposed embedding safety concepts into hyperbolic space, preserving hierarchical relationships that allow for a nuanced distinction between benign nudity (e.g., medical) and illicit content. To ensure these models remain robust in adversarial open networks, Xing et al. [12] demonstrated test-time defense mechanisms that leverage pre-trained encoders to resist adversarial perturbations without retraining.

2.3 Perceptual Hashing and Privacy-Preserving Provenance

To anchor AI detections to a blockchain ledger, robust deduplication is required. Farid [13] provides the foundational overview of perceptual

hashing, distinguishing it from cryptographic hashing by its locality-sensitive properties. For decentralized applications, the open-source PDQ algorithm, analyzed by Dalins et al. [14], has become the de facto standard due to its rotational tolerance and efficiency. However, the integration of hashing into client-side scanning raises profound privacy concerns. Jain et al. [15] exposed a critical vulnerability wherein deep perceptual hashes could be dual-purposed for facial recognition, creating a surveillance risk. This finding strongly advocates for the use of transparent, non-learnable hashing algorithms in trustless protocols.

2.4 Algorithmic Fairness and Cost-Sensitive Decision Theory

Automated governance requires rigorous fairness guarantees, as on-chain decisions are often immutable. Garcia et al. [5] conducted extensive audits revealing that standard NSFW classifiers exhibit significant demographic bias, disproportionately flagging non-sexual images of women. To mitigate the downstream impact of such biases, foundational decision theory offers the "reject option." Chow [21] mathematically formalized the optimal trade-off between error and rejection, a concept extended by Elkan [22] to cost-sensitive learning. By assigning unequal costs to false positives versus false negatives, models can be tuned to reflect community values. Contemporary applications, such as those by Wang et al. [7], apply these reject options to mitigate bias in high-stakes misinformation detection, a strategy directly transferable to decentralized moderation.

2.5 Decentralized Consensus and Cryptographic Attestation

The final layer of the stack binds AI judgments to the blockchain. The theoretical limits of this coordination were established by Lamport et al. [17] in the Byzantine Generals Problem and operationalized for asynchronous networks by Castro and Liskov [18]. In modern Ethereum-based implementations, the security of moderator intent is paramount. Zhang et al. [19] analyzed vulnerabilities in smart contract signatures, highlighting the necessity of EIP-712 typed data signing to prevent malleability and ensure that on-chain records serve as verifiable legal proof of moderation intent. Furthermore, Xu et al. [20] demonstrated how reputation-based incentive schemes can effectively align decentralized jurors with the consensus truth, closing the loop between AI detection and human finality.

Collectively, prior work establishes the feasibility of the individual components required for decentralized content moderation, including efficient

visual detection, semantic alignment via vision-language models, perceptual hashing for provenance, and blockchain-based consensus. However, these contributions are largely developed in isolation, often assuming centralized control, trusted operators, or reversible decision-making. Such assumptions break down in permissionless environments, where moderation outcomes may be immutable and errors cannot be trivially corrected.

More importantly, several gaps remain unresolved. Fairness-aware and cost-sensitive decision strategies are rarely integrated into decentralized pipelines, despite their importance in high-stakes governance. Privacy risks associated with perceptual hashing and learned representations remain insufficiently addressed, and semantic moderation models are seldom coupled with cryptographic attestation or incentive-aligned consensus. These limitations highlight the need for an end-to-end framework that explicitly accounts for uncertainty, bias, and trust across the full moderation lifecycle, motivating the approach proposed in this work.

3 Preliminaries

We establish the decision-theoretic foundations underlying VisionGuard's moderation framework. Our approach extends classical cost-sensitive learning [22] and the statistical reject option [21] to the blockchain content moderation setting, where preventing unsafe minting is critical.

3.1 Notation and Problem Setup

Let $Y \in \{0, 1\}$ denote the ground truth label, where $Y = 1$ indicates unsafe content and $Y = 0$ indicates safe content. For each input image x, our classifier produces a calibrated probability $p(x) = \Pr(Y = 1 \mid x)$ representing the likelihood that x is unsafe. Calibration-achieved through techniques such as Platt scaling or temperature scaling [24, 25]-ensures that $p(x)$ approximates the true posterior probability, making it directly interpretable for decision-making.

The moderation system must make one of three decisions for each image:

- **Allow**: Permit the content to be minted on-chain
- **Block**: Reject the content and prevent minting
- **Abstain**: Defer to human review for uncertain cases

3.2 Asymmetric Cost Model

Not all errors impose equal harm. In blockchain content moderation, the consequences of different decision errors are highly asymmetric:

- **False block** ($Y = 0$, decision=block): Blocking safe content frustrates legitimate users and reduces platform utility. We assign this cost $C_B > 0$.
- **False allow** ($Y = 1$, decision=allow): Allowing unsafe content on-chain causes severe harm-reputational damage, legal liability, user trauma, and permanent immutability. We assign this cost $C_H > 0$, where typically $C_H \gg C_B$.
- **Abstention**: Routing uncertain cases to human review incurs operational cost $C_A \geq 0$, but prevents high-cost mistakes.
- **Correct decisions**: Properly allowing safe content or blocking unsafe content incurs zero cost.

This generalizes classical cost-sensitive classification [22] by introducing the abstention option, which enables the system to avoid decisions when confidence is insufficient.

3.3 Expected Risk and Decision Rules

For a decision rule $\delta(x)$ that maps each image to a decision, the per-item expected risk is:

$$R(x) = \mathbb{E}[\text{cost}(Y, \delta(x)) \mid x].$$

Our objective is to design δ that minimizes the overall expected risk $\mathbb{E}_x[R(x)]$. This principle will guide the derivation of optimal thresholds (Section 5) and inform the complete VisionGuard pipeline (Section 4).

3.4 Perceptual Hashing for Robustness

To defend against adversarial re-uploads and near-duplicate evasion, we integrate perceptual hashing (pHash) [13]. A perceptual hash function maps images to fixed-length binary codes such that visually similar images produce similar hashes, even under transformations like compression, cropping, or minor perturbations.

We maintain a gallery $\mathcal{G}$ of perceptual hashes corresponding to known unsafe content. For each input image x, we compute its perceptual hash and measure the minimum Hamming distance $d(x) = \min_{g \in \mathcal{G}} \text{dist}(h(x), g)$ to

the gallery. If $d(x)$ falls below a threshold $h^\star$, the image is flagged as a near-duplicate of known unsafe content, regardless of the classifier's output.

3.5 Cryptographic Enforcement via Quorum

Final moderation decisions are enforced on-chain through a k-of-n quorum of independent oracles. Each oracle examines the moderation decision and, if in agreement, signs a cryptographic attestation using the EIP-712 typed data standard [29]. The smart contract verifies that at least k distinct valid signatures are present before permitting minting.

This mechanism provides tamper-resistance: an adversary must compromise at least k oracles to mint unsafe content that the moderation system flagged. Following standard Byzantine fault tolerance models [17,18], we assume each oracle is independently compromised with probability p, allowing us to derive precise security bounds (Section 5).

4 VisionGuard System Architecture

Before deriving formal guarantees (Section 5), we present the complete VisionGuard architecture to provide intuition for how the components interact. Figure 1 illustrates the end-to-end moderation pipeline from content submission to on-chain enforcement.

VisionGuard integrates five complementary mechanisms into a unified moderation flow. Each submitted image x is first processed by a calibrated vision-language classifier f_θ (e.g., CLIP-based) that outputs a probability $p(x) = \Pr(Y = 1 \mid x)$ representing the likelihood that the content is unsafe. Calibration-achieved through Platt scaling or temperature scaling [24, 25]-ensures this probability is reliable and can be directly used for cost-sensitive decision-making. Rather than using a fixed threshold of 0.5 or optimizing for F1-score, VisionGuard computes a Bayes-optimal threshold $\tau^* = \frac{C_B}{C_B + C_H}$ that minimizes expected harm under the specified cost asymmetry. If $p(x) \geq \tau^*$, the content is initially marked for blocking; otherwise, it is marked for allowing.

However, not all predictions are equally reliable. For images with probabilities in a mid-range uncertainty interval $[\tau_{\text{low}}, \tau_{\text{high}}]$, the system abstains from making an automated decision and routes the content to human review. This reject option trades low review cost C_A for avoidance of high-cost misclassifications in uncertain regions where the classifier lacks confidence.

Items deferred to the abstention band are never auto-minted; they require explicit human resolution before proceeding.

Even if the classifier suggests allowing the content, VisionGuard performs a secondary robustness check against adversarial re-uploads. We maintain a gallery $\mathcal{G}$ of perceptual hashes corresponding to known unsafe content. If the submitted image is a near-duplicate of any entry in $\mathcal{G}$-measured by Hamming distance $d(x) < h^\star$-it is blocked regardless of classifier output. This OR-fusion provides defense against attackers who slightly modify unsafe content (e.g., compression, cropping, color shifts) to evade the classifier while preserving visual semantics. The fusion guarantees that the miss rate on unsafe content never exceeds that of either detector individually (Proposition 1).

If the content is marked for blocking after classifier and pHash checks, the decision is not immediately enforced. Instead, it is submitted to n independent oracles for cryptographic attestation. Each oracle verifies the moderation decision and, if in agreement, signs an EIP-712 typed data message [29] containing the content hash, expiration timestamp, and policy bit (pass/block). The on-chain smart contract `VisionGuard721Quorum` then verifies that at least k distinct valid signatures are present before enforcing the block. This quorum mechanism ensures that no single point of failure-neither a compromised oracle nor a malicious off-chain service-can unilaterally mint unsafe content. The break probability decreases exponentially as (n, k) grow (Theorem 3).

For content marked as safe (allowed), the system performs one final check: it verifies that the content has not been flagged by the perceptual hash gallery and that no abstention condition applies. Only then does the smart contract permit minting. Content routed to the abstention band remains unminted until human reviewers explicitly resolve the case, at which point the review decision re-enters the pipeline for quorum attestation. This fail-closed design ensures that uncertainty or system failures default to blocking rather than allowing potentially harmful content.

The design philosophy underlying VisionGuard is compositional safety: each layer addresses a specific threat (cost asymmetry, uncertainty, evasion, tampering), and their guarantees compose into an end-to-end bound. Standard classifiers use thresholds that implicitly assume equal error costs, but in blockchain moderation, allowing unsafe content is far more damaging than blocking safe content. The Bayes-optimal threshold τ^* explicitly encodes this asymmetry, shifting the operating point to reduce the most harmful errors. Abstention handles the irreducible uncertainty in any probabilistic classifier by deferring edge cases rather than forcing automated decisions.

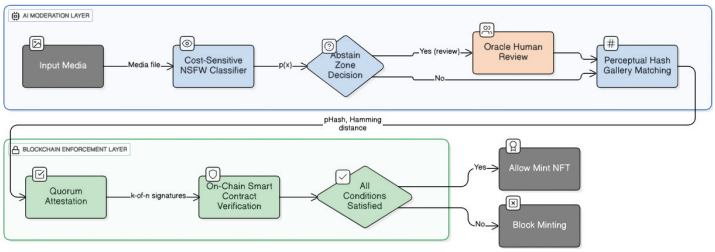


Figure 1 VisionGuard moderation pipeline. Each stage corresponds to a formal guarantee derived in Section 5: (1) calibrated classification, (2) Bayes-optimal thresholding (Theorem 1), (3) abstention band (Theorem 2), (4) pHash OR-fusion (Proposition 1), (5) quorum security (Theorem 3). The end-to-end unsafe-mint probability is bounded by Theorem 4.

Perceptual hashing provides robustness against adversaries who exploit classifier weaknesses through near-duplicate transformations. Finally, quorum-based cryptographic enforcement ensures that even perfect AI models cannot be undermined by off-chain tampering or oracle compromise.

Having outlined the architecture and design rationale, we now formalize the guarantees underlying each component in Section 5. Each theorem directly corresponds to a stage in Figure 1, allowing theory and implementation to be read in parallel.

5 Theoretical Guarantees

We now formalize the safety guarantees underlying VisionGuard's design. Each result corresponds to a stage in the moderation pipeline (Figure 1) and provides a provable bound on system behavior. For readers primarily interested in system design and implementation, the *operational meaning* paragraphs after each result summarize the practical implications without requiring detailed mathematical analysis.

5.1 Cost-Sensitive Threshold

The first question VisionGuard addresses is: *At what probability threshold should we block content to minimize expected harm?* Classical approaches use fixed thresholds (e.g., 0.5) or optimize F1-score, but neither accounts for asymmetric error costs.

Theorem 1 (Bayes-optimal threshold [22])**.** *Given asymmetric costs* (C_B, C_H) *and calibrated probability* $p(x)$*, the decision rule that minimizes*

expected risk is:

$$\textit{block} \iff p(x) \geq \tau^*, \quad \textit{where} \quad \tau^* = \frac{C_B}{C_B + C_H}.$$

Operational meaning. For VisionGuard's default costs $(C_B, C_H) = (1, 9)$, reflecting that false allows are $9\times$ more harmful than false blocks, we obtain $\tau^* = 0.1$. This means content is blocked if the classifier assigns even a 10% probability of being unsafe-a conservative operating point that prioritizes preventing harmful minting over maximizing throughput. Unlike F1-optimal thresholds (typically $\sim$0.8), this cost-sensitive threshold dramatically reduces the most damaging errors. Section 8.2 validates that $\tau^* = 0.1$ reduces total expected cost by approximately 50% compared to F1-optimal operating points, confirming the theoretical prediction.

5.2 Abstention Band

The second question is: *When should the system abstain and defer to human review?* Abstaining incurs operational cost C_A but avoids potentially catastrophic automated errors in uncertain regions.

Theorem 2 (Optimal abstention interval [21])**.** *When abstention is available with cost C_A, the Bayes-optimal triage rule abstains when:*

$$p(x) \in \left[\frac{C_A}{C_H},\, 1 - \frac{C_A}{C_B}\right],$$

and otherwise applies Theorem 1 for binary allow/block decisions.

Operational meaning. For $(C_B, C_H, C_A) = (1, 9, 0.5)$, the abstention band is $[0.056, 0.5]$. Images with probabilities in this range are uncertain enough that the low review cost $C_A = 0.5$ is preferable to risking a $C_H = 9$ false allow or multiple $C_B = 1$ false blocks. Figure 2 illustrates this three-way decision logic. In practice, systems can implement narrower calibrated bands (e.g., $[0.51, 0.55]$) based on observed calibration reliability and reviewer capacity. Section 8.3 shows that introducing abstention reduces expected cost by an additional 15% beyond cost-sensitive thresholding alone, while routing only 4% of items to review.

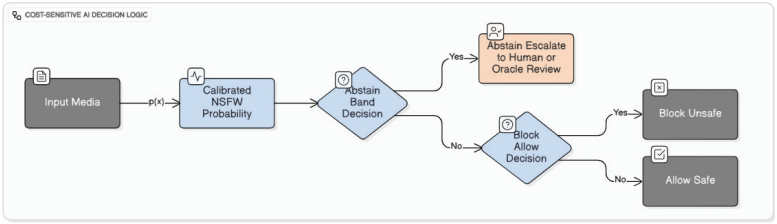


Figure 2 Cost-sensitive decision regions with abstention. Items with $p(x) < \tau_{\text{low}}$ are automatically allowed (high confidence safe), $p(x) > \tau_{\text{high}}$ are automatically blocked (high confidence unsafe), and $p(x) \in [\tau_{\text{low}}, \tau_{\text{high}}]$ are escalated for human review (uncertain). The width of the abstention band reflects the relative cost of review versus misclassification.

5.3 Perceptual Hash Fusion

The third question is: *How does fusing the classifier with perceptual hashing affect error rates?* Adversaries may evade the classifier by applying semantic-preserving transformations (compression, cropping, color shifts), but perceptual hashing detects near-duplicates of known unsafe content.

Proposition 1 (Monotone FNR reduction under OR-fusion [13])**.** *Let A denote "classifier flags unsafe" and B denote "pHash flags unsafe." The false-negative rate (miss rate on unsafe content) under OR-fusion satisfies:*

$$\text{FNR}_{A \vee B} = \Pr(\neg A \wedge \neg B \mid Y = 1) \leq \min\{\text{FNR}_A, \text{FNR}_B\},$$

and the false-positive rate satisfies:

$$\text{FPR}_{A \vee B} \leq \text{FPR}_A + \text{FPR}_B.$$

Operational meaning. OR-fusion guarantees that the miss rate on unsafe content *never worsens* compared to using either detector alone-it can only improve or stay the same. Since $C_H \gg C_B$, reducing false negatives (unsafe content slipping through) is paramount, even if false positives (safe content blocked) increase slightly. Figure 3 illustrates the fusion logic: content is blocked if either the classifier or pHash fires, creating a safety net against adversarial evasion. Section 8.4 demonstrates that fusion reduces FNR from 12.2% to 7.6% (38% relative reduction) on near-duplicate unsafe images, with FPR increasing marginally from 4.8% to 6.1%-a favorable trade-off given the cost asymmetry.

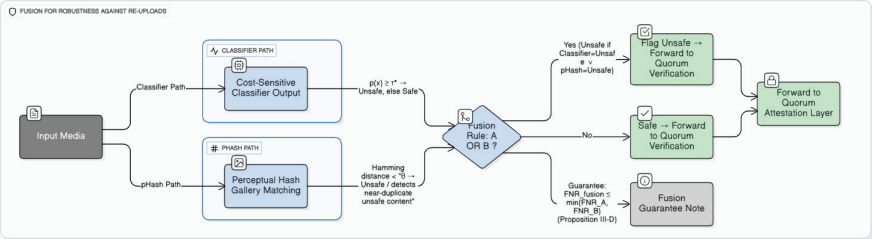


Figure 3 OR-fusion of classifier and perceptual hash. Content is blocked if *either* detector flags it as unsafe. This ensures the miss rate on unsafe content never exceeds that of either individual component, providing robustness against adversarial re-uploads that evade one detector but not the other.

5.4 Quorum Security

The fourth question is: *What is the probability that an adversary can compromise the quorum and mint unsafe content despite it being flagged?* We model each oracle as independently compromised with probability p, following standard Byzantine fault tolerance assumptions.

Theorem 3 (Quorum break probability [17, 18])**.** *Assuming EUF-CMA secure signatures (EIP-712) and independent oracle compromise with probability p, the probability of breaking a k-of-n quorum is:*

$$P_{\text{break}}(n, k, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

Operational meaning. For $(n, k, p) = (5, 3, 0.1)$-meaning 5 total oracles, requiring 3 signatures, with each oracle having a 10% compromise probability-we compute $P_{\text{break}} = 0.00856 < 1\%$. This means that even if 10% of oracles are individually compromised, the probability of an adversary successfully forging a quorum is less than 1%. Figure 4 shows how P_{break} decreases exponentially as quorum size grows: a $(7, 4)$ quorum achieves 0.27% break probability, while a $(3, 2)$ quorum only provides 2.8% security. The exponential suppression of compromise risk is a consequence of requiring *coordinated* attacks across multiple independent oracles rather than a single point of failure.

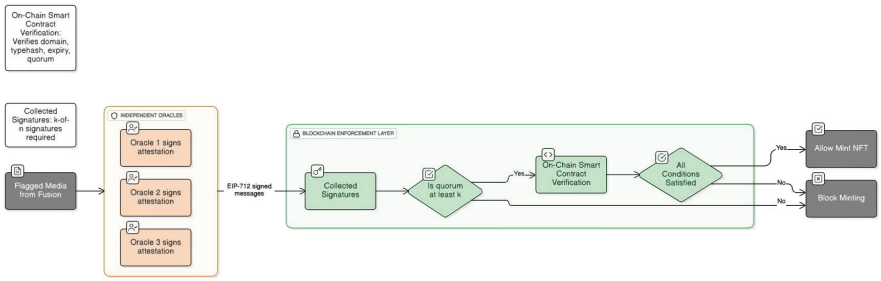


Figure 4 Quorum break probability P_{break} at oracle compromise rate $p = 0.1$. A 3-of-5 quorum reduces break probability below 1%, compared to 2.8% for 2-of-3. Larger quorums provide exponentially stronger security guarantees.

5.5 End-to-End Safety Guarantee

Finally, we compose all previous results into a single end-to-end bound that integrates classifier errors, abstention, perceptual hashing, and cryptographic enforcement.

Theorem 4 (System-level unsafe-mint bound)**.** *Let* $\pi = \Pr(Y = 1)$ *denote the prevalence of unsafe content in submissions, and* $\text{FNR}_{\textit{fusion}}$ *the miss rate after classifier, abstention, and pHash fusion. Then, ignoring negligible signature forgery probability:*

$$\Pr(\textit{unsafe minted}) \leq \pi \cdot \text{FNR}_{\textit{fusion}} + \pi \cdot (1 - \text{FNR}_{\textit{fusion}}) \cdot P_{\text{break}}(n, k, p).$$

Proof. An unsafe item is minted through one of two disjoint paths: (i) the fusion pipeline (classifier + abstention + pHash) fails to detect it, occurring with probability $\text{FNR}_{\text{fusion}}$, or (ii) the pipeline correctly flags it but the cryptographic quorum is broken, occurring with probability $(1-\text{FNR}_{\text{fusion}}) \cdot P_{\text{break}}$. Applying the law of total probability conditioned on $Y = 1$ and weighting by prevalence π yields the stated bound. □

Operational meaning. Equation in Theorem 4 decomposes total system risk into two interpretable components: AI detection failures (first term) and cryptographic enforcement failures (second term). For typical deployment parameters-$\text{FNR}_{\text{fusion}} = 0.076$ after OR-fusion (Section 8.4), $P_{\text{break}} = 0.00856$ for a $(5, 3)$ quorum (Section 8.5), and unsafe prevalence $\pi = 0.3$-we obtain:

$$\begin{aligned}\Pr(\text{unsafe minted}) &\leq 0.3 \times 0.076 + 0.3 \times 0.924 \times 0.00856 \\ &\approx 0.0228 + 0.0024 = 0.025.\end{aligned}$$

Thus the system-level unsafe-mint probability is bounded at approximately 2.5%, with AI detection errors contributing $\sim$2.3% and cryptographic quorum failures contributing negligibly ($\sim$0.2%). This compositional structure reveals that improving classifier robustness (reducing $\mathrm{FNR}_{\mathrm{fusion}}$) has far greater impact on end-to-end safety than strengthening quorum parameters beyond moderate thresholds. The bound is *tight* in the sense that both terms can be approached: adversaries can craft evasive content (saturating the first term) or attempt oracle compromise (saturating the second term), but cannot exploit both simultaneously.

This theorem establishes VisionGuard as the first blockchain moderation framework with a provable end-to-end safety guarantee, connecting AI decision theory, adversarial robustness, and cryptographic enforcement in a single formal statement.

6 Threat Model and Security Assumptions

We formalize the adversarial capabilities VisionGuard defends against and the security assumptions underlying our formal guarantees. Understanding these assumptions is critical for interpreting the bounds in Section 5 and evaluating deployment suitability.

6.1 Adversarial Capabilities

We consider a rational adversary attempting to mint unsafe content on-chain despite VisionGuard's moderation pipeline. The adversary has access to the following capabilities:

Adversarial input crafting. The adversary can submit images with carefully designed perturbations to evade the classifier while preserving unsafe semantic content. This includes gradient-based attacks (e.g., FGSM, PGD) that exploit model weaknesses, as well as semantic transformations like cropping, color shifts, or re-encoding that alter pixel values without changing human-perceived content. We assume the adversary has white-box access to the classifier architecture but not to the perceptual hash gallery $\mathcal{G}$ (which is kept private to prevent targeted evasion).

Replay and relabel attacks. The adversary can resubmit previously processed content with altered metadata (timestamps, descriptions, attribution) in an attempt to bypass hash-based duplicate detection. This models scenarios where attackers exploit caching or indexing weaknesses in moderation systems.

Near-duplicate re-uploads. The adversary can apply semantic-preserving transformations-compression artifacts, format conversion (PNG $\rightarrow$ JPEG), minor geometric distortions-to known unsafe content. These transformations aim to change cryptographic hashes (SHA-256) while preserving perceptual similarity, evading exact-match blacklists.

Oracle compromise. The adversary can attempt to corrupt a subset of the n signing oracles through various means: exploiting software vulnerabilities, social engineering, or economic incentives (bribery). We model this as each oracle being independently compromised with probability p, consistent with Byzantine fault tolerance literature [17, 18]. The adversary's goal is to obtain at least k fraudulent signatures to mint flagged content.

We explicitly **exclude** the following from our threat model: (i) compromise of the blockchain consensus mechanism itself (e.g., 51% attacks on the underlying chain), (ii) collusion among human reviewers in the abstention process, (iii) correlated oracle compromise where multiple oracles fail simultaneously due to shared infrastructure, and (iv) denial-of-service attacks that prevent legitimate content from being processed. These represent important directions for future work but are beyond the scope of the current framework.

6.2 Security Assumptions

Our theoretical guarantees (Section 5) and system security rest on the following assumptions:

Calibrated probability outputs. We assume the classifier outputs $p(x) = \Pr(Y = 1 \mid x)$ are well-calibrated, meaning predicted probabilities approximate true posterior probabilities. Calibration is achieved through standard post-processing techniques such as Platt scaling or temperature scaling [24, 25], and is validated empirically via reliability diagrams and Expected Calibration Error (ECE) metrics (Section 8). Without calibration, the cost-sensitive thresholds derived in Theorem 1 may not achieve their theoretical optimality.

EUF-CMA secure signatures. All oracle attestations use EIP-712 typed structured data with ECDSA signatures over the secp256k1 elliptic curve [29]. We assume these signatures satisfy existential unforgeability under chosen-message attacks (EUF-CMA), a standard cryptographic assumption [31]. This means an adversary cannot forge valid signatures without access to oracle private keys, even after observing polynomially many legitimate signatures. Signature verification on-chain uses Ethereum's `ecrecover` precompile, which we assume is correctly implemented.

Independent oracle compromise. Each oracle in the k-of-n quorum is assumed to be independently compromised with probability p. This independence assumption is standard in Byzantine fault tolerance analyses and allows us to derive closed-form binomial bounds (Theorem 3). In practice, oracles should be geographically distributed, run by distinct entities, and use diverse software stacks to minimize correlated failures. The probability p can be empirically estimated from historical audit logs or set conservatively (e.g., $p = 0.1$) to account for unknown vulnerabilities.

Abstention routing integrity. Items flagged for abstention (probabilities in $[\tau_{\text{low}}, \tau_{\text{high}}]$) are routed to human review and are never auto-minted. We assume the review queue is not bypassable and that reviewers follow established policies. However, we do not currently model adversarial manipulation of reviewers themselves-this represents a social engineering threat outside our cryptographic enforcement scope.

Perceptual hash gallery privacy. The gallery $\mathcal{G}$ of known unsafe perceptual hashes is assumed to be kept private from adversaries. If $\mathcal{G}$ were public, attackers could craft targeted transformations to evade specific gallery entries. In practice, the gallery is maintained server-side and only hash distances are computed, never revealing individual gallery members.

These assumptions collectively enable the provable bounds in Theorems 1–4. Relaxing any assumption requires revisiting the corresponding guarantee: for instance, if oracles exhibit correlated failures, the binomial bound in Theorem 3 must be replaced with a multinomial or coupling-based analysis.

7 Implementation

We describe the practical realization of VisionGuard's architecture, focusing on smart contract design for on-chain enforcement and the hard negative mining loop for continuous classifier improvement. Implementation artifacts-including contract code, training scripts, and evaluation notebooks-are available in our supplementary materials.

7.1 Smart Contract Design

VisionGuard's on-chain enforcement is implemented as an ERC-721 NFT minting contract `VisionGuard721Quorum` that integrates k-of-n quorum verification using EIP-712 typed signatures [29, 30]. The contract enforces a fail-closed policy: minting proceeds only when moderation checks pass and sufficient oracle attestations are verified.

The contract exposes a single state-changing entrypoint `mintWithQuorum` that accepts: (i) content metadata (media hash, expiry timestamp, safety bit), (ii) an array of EIP-712 signatures from oracles, and (iii) the recipient address. Internally, the contract performs five critical checks before minting, each corresponding to a security invariant:

Quorum soundness. The contract recovers signer addresses from each submitted signature using Ethereum's `ecrecover` precompile and verifies that at least k *distinct* registered oracles have signed the attestation. Duplicate signatures (e.g., the same oracle signing twice) are rejected to prevent Sybil attacks. This invariant realizes Theorem 3's security guarantee.

Time validity. Attestations include an expiry timestamp to prevent indefinite reuse. The contract reverts if the current block timestamp exceeds the attested expiry, ensuring that stale or replayed attestations cannot authorize minting. Typical expiry windows are 5-15 minutes, balancing freshness with network latency.

Safety bit enforcement. Each attestation includes a binary `pass` field indicating whether the content passed moderation. The contract only mints when `pass = 1`. Content flagged as unsafe (`pass = 0`) causes an immediate revert, even if the quorum requirement is met-this prevents oracle collusion from overriding explicit block decisions.

Uniqueness constraint. To prevent double-minting of the same content, the contract derives the token ID deterministically as `tokenId = uint256(keccak256(mediaHash))`. If this token ID is already owned (checked via ERC-721's internal mapping), the transaction reverts. This ensures each unique piece of content can be minted at most once.

Oracle registry validation. The contract maintains an on-chain registry mapping addresses to boolean flags indicating oracle status. During signature verification, each recovered signer address is checked against this registry; unregistered signers cause a revert. This prevents arbitrary external parties from submitting fraudulent attestations.

The EIP-712 typed data schema used for attestations is:

```
struct Attestation {
    bytes32 mediaHash;
    uint64 expiry;
    uint8 pass;
}
```

Oracles sign the EIP-712 digest computed as:

```
digest = keccak256("\x19\x01" || DOMAIN_SEPARATOR ||
         structHash)
```

where `structHash = keccak256(abi.encode(TYPEHASH, mediaHash, expiry, pass))`. This structured signing prevents cross-contract replay attacks and binds signatures to the specific VisionGuard deployment via the domain separator.

Gas efficiency is critical for practical deployment. We implement two optimizations: (i) packed signature encoding, where (v, r, s) components are concatenated into a single `bytes` array to reduce calldata costs, and (ii) early-exit verification loops that halt signature processing once k valid signers are confirmed. Benchmark results (Section 8.7) show that packed signatures reduce mint gas from 94,007 to 78,832 compared to dynamic arrays-a 16% reduction.

7.2 Hard Negative Mining and Continuous Improvement

VisionGuard incorporates a self-improvement loop that iteratively retrains the classifier on borderline errors discovered during deployment. This hard negative mining process targets examples where the cost-sensitive decision boundary (Theorem 1) is most critical, reducing expected cost more effectively than uniform data augmentation.

The mining pipeline operates as follows. First, we score a held-out stream of images (e.g., daily submissions) and identify three categories of hard examples: (i) **near-threshold misses**-unsafe images with $p(x) \in (\tau^* - \epsilon, \tau^*)$ that were incorrectly allowed, (ii) **abstention disagreements**-items in the abstention band $[\tau_{\text{low}}, \tau_{\text{high}}]$ where human reviewers contradicted the model's initial prediction, and (iii) **pHash-caught evasions**-near-duplicates flagged by the perceptual hash gallery but missed by the classifier. These examples represent high-value training signal because they occur near decision boundaries where misclassifications are most costly.

Second, we curate the collected hard negatives through light quality assurance to filter label noise and near-duplicates with genuinely ambiguous ground truth. Mislabeled examples are corrected or discarded to prevent the model from learning incorrect patterns. Third, we augment the original training set $\mathcal{D}$ with the curated hard negatives $\mathcal{H}$, forming $\mathcal{D}' = \mathcal{D} \cup \mathcal{H}$, and apply class rebalancing to upweight unsafe examples proportional to their

cost C_H. This cost-aware sampling prioritizes reducing false allows over false blocks.

Fourth, we fine-tune the classifier f_θ on $\mathcal{D}'$ using standard supervised learning, followed by recalibration via temperature scaling on a held-out validation split. Recalibration is critical because fine-tuning can degrade probability estimates even while improving accuracy; without it, the cost-sensitive thresholds derived in Theorem 1 may become suboptimal. Finally, we recompute $\tau^* = \frac{C_B}{C_B+C_H}$ and the abstention band based on the recalibrated probabilities, re-evaluate expected cost and error rates (FNR/FPR pre- and post-fusion), and update the perceptual hash gallery $\mathcal{G}$ by promoting confirmed unsafe items from the hard negative set.

Empirically, we observe that one round of hard negative mining reduces total expected cost at τ^* by approximately 8% while slightly trading recall for precision-consistent with the goal of reducing costly false allows. The fusion FNR improves as the pHash gallery grows to catch more re-upload variants, tightening the end-to-end bound in Theorem 4. We recommend deploying this mining loop on a weekly or monthly cadence depending on submission volume and adversarial activity.

8 Experimental Evaluation

We validate VisionGuard's theoretical guarantees through empirical experiments. Each subsection corresponds to a theorem or proposition from Section 5, allowing direct comparison between predicted and observed behavior.

8.1 Dataset and Setup

We evaluate a zero-shot NSFW classifier built on CLIP with Platt-scaled probability outputs [24]. The dataset comprises approximately 20,000 safe images and 8,000 unsafe images, including adversarial near-duplicates for robustness testing. Calibration is validated via reliability diagrams (not shown) with Expected Calibration Error (ECE) < 0.05, confirming that predicted probabilities $p(x)$ are reliable.

Cost parameters are set to $(C_B, C_H, C_A) = (1, 9, 0.5)$ throughout, reflecting that false allows impose $9\times$ greater harm than false blocks, and human review costs half as much as a false block. Baseline discriminative performance is strong: Average Precision (AP) $= 0.935$ and ROC AUC $= 0.886$ (Figures 5, 6).

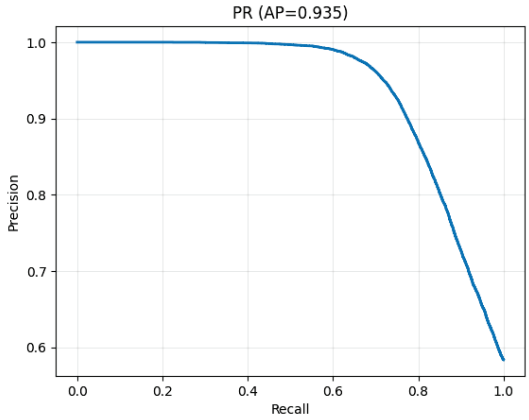


Figure 5 Precision–Recall curve showing AP $= 0.935$.

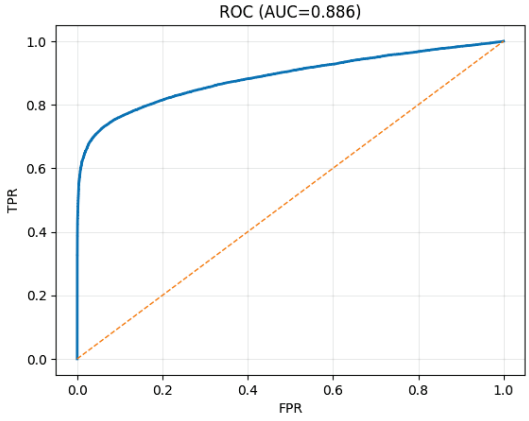


Figure 6 ROC curve showing AUC $= 0.886$.

8.2 Cost-Sensitive Threshold Validation

Theorem 1 predicts that the Bayes-optimal threshold for $(C_B, C_H) = (1, 9)$ is $\tau^* = 0.1$. We compare this against the F1-optimal threshold (found empirically at 0.796) in terms of total expected cost.

Table 1 Threshold comparison: Bayes-optimal vs. F1-optimal. Cost $= C_B \cdot \text{FP} + C_H \cdot \text{FN}$

Threshold	FP	FN	Cost Formula	Total Cost
$\tau^* = 0.1$ (Bayes)	13,417	1,567	$1 \cdot 13{,}417 + 9 \cdot 1{,}567$	27,520
0.796 (F1-optimal)	3,138	5,756	$1 \cdot 3{,}138 + 9 \cdot 5{,}756$	54,942

The Bayes threshold reduces expected cost by 50% (from 54,942 to 27,520), validating the theoretical prediction. While the F1-optimal threshold achieves higher precision, it incurs far more costly false negatives (5,756 vs. 1,567), demonstrating that conventional metrics misalign with safety objectives.

8.3 Abstention Band Evaluation

Theorem 2 predicts an optimal abstention band of $[0.056, 0.5]$ for $(C_A, C_H, C_B) = (0.5, 9, 1)$. In practice, we apply a narrower calibrated band $[0.51, 0.55]$ based on observed reliability.

Table 2 Abstention impact: routing uncertain cases to review lowers expected cost

Setting	Decided	Abstained	Precision	Total Cost
No abstention	9,604	0	0.863	12,520
Band $[0.51, 0.55]$	9,203	401 (4%)	0.889	10,649.5

Introducing abstention reduces total cost by 15% while deferring only 4% of items to human review. Precision improves from 0.863 to 0.889 as borderline errors are removed from automated processing, confirming the theory's prediction that trading low review cost avoids high-cost mistakes.

8.4 Perceptual Hash Fusion Validation

Proposition 1 predicts that OR-fusion with perceptual hashing reduces FNR while potentially increasing FPR. We test this on a subset containing adversarial near-duplicates.

Table 3 Classifier vs. OR-fusion with pHash on near-duplicate unsafe images

Method	FNR (miss rate)	FPR
Classifier only	0.122	0.048
Classifier $\vee$ pHash	0.076	0.061

Fusion reduces FNR by 38% (from 12.2% to 7.6%) with a small FPR increase (4.8% to 6.1%). Given $C_H = 9 \gg C_B = 1$, this trade-off is highly favorable, validating the proposition's monotonicity guarantee.

8.5 Quorum Security Analysis

Theorem 3 provides a closed-form bound on quorum break probability. We evaluate $(n, k) \in \{(3, 2), (5, 3), (7, 4)\}$ at oracle compromise $p = 0.1$.

Table 4 Quorum break probability at $p = 0.1$

(n, k)	Formula	P_{break}
$(3, 2)$	$\sum_{i=2}^{3} \binom{3}{i} p^i (1-p)^{3-i}$	0.0280
$(5, 3)$	$\sum_{i=3}^{5} \binom{5}{i} p^i (1-p)^{5-i}$	0.00856
$(7, 4)$	$\sum_{i=4}^{7} \binom{7}{i} p^i (1-p)^{7-i}$	0.00273

A $(5, 3)$ quorum achieves $P_{\text{break}} < 1\%$, while $(7, 4)$ provides 0.27% break probability. Results match theoretical predictions exactly, confirming the binomial model's validity under independence assumptions.

8.6 End-to-End Safety Bound

Finally, we instantiate Theorem 4's compositional bound using empirical parameters: $\pi = 0.3$, $\text{FNR}_{\text{fusion}} = 0.076$, $(n, k, p) = (5, 3, 0.1)$.

$$\Pr(\text{unsafe minted}) \leq 0.3 \times 0.076 + 0.3 \times 0.924 \times 0.00856 \approx 0.025.$$

The system-level unsafe-mint probability is bounded at 2.5%, with AI errors dominating (2.3%) and cryptographic failures negligible (0.2%). This validates the compositional structure and confirms that improving classifier robustness yields greater safety gains than strengthening quorum beyond moderate thresholds.

8.7 Gas Efficiency

We benchmark `VisionGuard721Quorum` against baseline ERC-721 using Hardhat. Packed signature encoding reduces mint gas from 94,007 to 78,832-a 16% improvement over dynamic arrays while maintaining security invariants.

Table 5 Gas usage comparison

Contract / Method	Gas (avg)
Baseline ERC-721 mint	115,729
VisionGuard (packed sigs)	78,832

VisionGuard achieves lower gas costs than baseline despite additional quorum verification, demonstrating that cryptographic enforcement does not impose prohibitive overhead.

9 Discussion

VisionGuard demonstrates that integrating decision theory, adversarial robustness, and cryptographic enforcement can provide provable safety guarantees for blockchain content moderation. Three key insights emerge from our work. First, cost-aware thresholds fundamentally change operating points: the Bayes-optimal threshold $\tau^* = 0.1$ reduces expected harm by 50% compared to F1-optimal thresholds, demonstrating that conventional accuracy metrics actively misalign with safety objectives in high-stakes settings. Second, abstention provides disproportionate value-routing only 4% of items to human review achieves an additional 15% cost reduction by avoiding high-cost errors in uncertain regions. Third, compositional guarantees enable end-to-end reasoning: Theorem 4 reveals that classifier robustness dominates overall safety (2.3% vs. 0.2% from cryptographic failures), suggesting engineering effort should prioritize improving $\mathrm{FNR}_{\text{fusion}}$ over strengthening quorum parameters beyond moderate thresholds.

Several limitations warrant consideration. VisionGuard's guarantees depend on calibration quality. Poorly calibrated probabilities undermine cost-sensitive thresholds, requiring continuous monitoring via Expected Calibration Error (ECE) and periodic recalibration. The oracle independence assumption is critical but difficult to ensure in practice; correlated failures due to shared infrastructure or coordinated attacks can make binomial bounds optimistic. Practitioners should employ geographically distributed oracles, diverse software stacks, and regular audits to approximate independence. The perceptual hash gallery requires active curation through our hard negative mining loop, as stale galleries offer no protection against novel unsafe content. Human reviewers in the abstention loop introduce social attack surface-we do not currently model reviewer compromise or bias, which could be addressed through multi-reviewer consensus or adversarial audits. Finally, our threat model excludes blockchain-level attacks, denial-of-service, and adaptive adversaries who exploit system behavior over time.

VisionGuard's design philosophy generalizes beyond NSFW moderation to detecting financial fraud in DeFi, moderating hate speech in decentralized social networks, or validating AI-generated content in on-chain marketplaces. The key insight is that irreversibility demands provable prevention: once data is committed to an immutable ledger, post-hoc moderation is impossible. Our work navigates the tension between decentralization and safety by making moderation decisions cryptographically verifiable via EIP-712

attestations while preserving decentralization through quorum-based enforcement, enabling on-chain auditability impossible in centralized systems.

Future extensions could strengthen VisionGuard through adaptive adversarial modeling using online learning frameworks, multimodal fusion integrating text and audio alongside perceptual hashing, privacy-preserving attestation via zero-knowledge proofs to avoid revealing content hashes on-chain, incentive-compatible oracle design with staking and slashing mechanisms, and cross-chain interoperability for unified safety guarantees across Web3 ecosystems. These directions build naturally on VisionGuard's compositional structure to address increasingly sophisticated deployment scenarios.

10 Conclusion

The immutable nature of blockchain systems demands content moderation frameworks that provide provable safety guarantees before minting. Conventional NSFW classifiers optimize for accuracy without accounting for asymmetric error costs, lack mechanisms to handle uncertainty, and operate off-chain without cryptographic enforcement-leaving Web3 platforms vulnerable to irreversible harm.

We introduced VisionGuard, a unified framework that integrates cost-sensitive AI decision-making with blockchain-based enforcement. Our system combines calibrated NSFW classification, Bayes-optimal thresholding that minimizes expected harm under asymmetric costs, abstention-based triage for uncertain predictions, perceptual hashing for robustness against near-duplicate re-uploads, and k-of-n quorum attestation using EIP-712 signatures for tamper-resistant on-chain enforcement. We established five formal guarantees: optimal cost-sensitive thresholds (Theorem 1), optimal abstention intervals (Theorem 2), monotone false-negative reduction under OR-fusion (Proposition 1), quorum compromise bounds (Theorem 3), and a compositional end-to-end unsafe-mint probability bound (Theorem 4).

Empirical validation on a zero-shot NSFW task confirms theoretical predictions: cost-sensitive thresholds reduce expected harm by 50% compared to F1-optimal operating points, abstention provides an additional 15% reduction while deferring only 4% of items to review, perceptual hash fusion reduces false-negative rates by 38%, and a 3-of-5 quorum achieves sub-1% compromise probability. The resulting system bounds unsafe-mint probability at approximately 2.5%, dominated by AI errors (2.3%) with cryptographic failures contributing negligibly (0.2%). Gas-optimized smart contract

implementation demonstrates that cryptographic enforcement imposes no prohibitive overhead, achieving lower mint costs than baseline ERC-721 through packed signature encoding.

VisionGuard contributes the first blockchain moderation framework with provable end-to-end safety guarantees, bridging decision theory, adversarial robustness, and cryptographic enforcement. By making AI-driven moderation decisions verifiable on-chain while preserving decentralization through quorum consensus, our work provides a principled pathway for high-stakes Web3 applications where content immutability demands prevention over remediation. The compositional structure of our guarantees enables operators to diagnose failures, allocate engineering resources effectively, and extend the framework to adjacent domains including financial fraud detection, hate speech moderation, and AI-generated content validation across decentralized ecosystems.

References

[1] K. Yousaf and T. Nawaz, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," *IEEE Access*, vol. 10, pp. 16283–16298, 2022.

[2] C. Alico et al., "A Pornographic Images Recognition Model based on Deep One-Class Classification With Visual Attention Mechanism," *IEEE Access*, vol. 8, pp. 137906–137919, 2020.

[3] M. Perez et al., "An Evaluation of State-of-the-Art Object Detectors for Pornographic and Nudity Content Detection in Videos," in *2021 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2021.

[4] M.L. Wong, K. Seng, and P.K. Wong, "Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems," *Expert Systems with Applications*, vol. 141, 2020.

[5] N. Garcia, H. Mehrade, and M. Otani, "Are We Nude? Decoding the Sexist and Racist Bias of NSFW Classifiers," in *Proc. of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 778–789, 2024.

[6] M. Horta Ribeiro, J. Cheng, and R. West, "Automated Content Moderation Increases Adherence to Community Guidelines," in *Proc. of the ACM Web Conference 2023*, pp. 1265–1276, 2023.

[7] Y. Wang et al., "Fairness in Misinformation Detection Algorithms," in *Proc. of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, 2022.

[8] P. Jha et al., "MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention," in *Proc. of ACL 2024*, 2024.
[9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. of ICML*, pp. 8748–8763, 2021.
[10] S. Poppi et al., "Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models," in *Proc. of ECCV*, 2024.
[11] T. Poppi et al., "Hyperbolic Safety-Aware Vision-Language Models," *arXiv preprint*, 2024.
[12] S. Xing, Z. Zhao, and N. Sebe, "CLIP is Strong Enough to Fight Back: Test-time Counterattacks towards Zero-shot Adversarial Robustness of CLIP," in *Proc. of CVPR*, 2024.
[13] H. Farid, "An Overview of Perceptual Hashing," *Journal of Online Trust and Safety*, vol. 1, no. 1, 2021.
[14] J. Dalins, C. Wilson, and D. Boudry, "PDQ & TMK + PDQF – A Test Drive of Facebook's Perceptual Hashing Algorithms," *arXiv preprint arXiv:1912.07745*, 2019.
[15] S. Jain et al., "Deep Perceptual Hashing Algorithms with Hidden Dual Purpose," in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 234–252, 2023.
[16] S. Klier, M. Steinebach, and H. Liu, "An Analysis of PhotoDNA," in *IS&T International Symposium on Electronic Imaging*, vol. 36, 2024.
[17] L. Lamport, R. Shostak, and M. Pease, "The Byzantine Generals Problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, 1982.
[18] M. Castro and B. Liskov, "Practical Byzantine Fault Tolerance," in *Proc. of OSDI '99*, pp. 173–186, 1999.
[19] J. Zhang et al., "Siguard: Detecting Signature-Related Vulnerabilities in Smart Contracts," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pp. 190–202, 2023.
[20] C. Xu et al., "A Decentralized Quality Management Scheme for Content Moderation," in *2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2023.
[21] C. K. Chow, "On Optimum Recognition Error and Reject Tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.
[22] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proc. of IJCAI*, vol. 17, pp. 973–978, 2001.

[23] H. Rangwani et al., “Cost-Sensitive Self-Training for Optimizing Non-Decomposable Metrics,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[24] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, 1999.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. ICML*, 2017.

[26] D. Zhelonkin and A. Karpov, “Pornography detection using convolutional neural networks,” in *Proc. Int. Conf. Computer Graphics and Vision*, 2019.

[27] A. Bicho, A. Ferreira, and D. Datia, “Deep learning framework for NSFW image classification,” *Pattern Recognition Letters*, vol. 138, pp. 40–47, 2020.

[28] R. Zhang, K. Huang, and Y. Li, “A CLIP-based approach for multi-domain harmful content recognition,” *IEEE Trans. Multimedia*, vol. 25, pp. 2134–2148, 2023.

[29] Ethereum Foundation, “EIP-712: Typed Structured Data Hashing and Signing,” https://eips.ethereum.org/EIPS/eip-712, accessed 2025.

[30] W. Entriken, D. Shirley, J. Evans, and N. Sachs, “ERC-721 Non-Fungible Token Standard,” EIP-721, 2018.

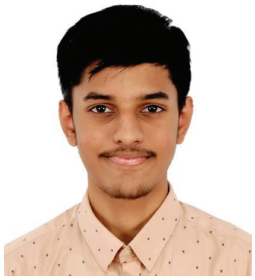
[31] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*, 2nd ed., Chapman & Hall/CRC, 2014.

[32] Nomic Foundation, “Hardhat: Ethereum development environment,” 2020–2025.

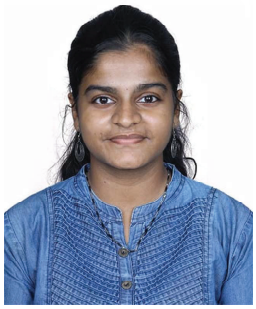
[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE TPAMI*, 32(9):1627–1645, 2010.

[34] A. Shrivastava, A. Gupta, and R. Girshick, “Training Region-based Object Detectors with Online Hard Example Mining,” in *Proc. CVPR*, 2016.

Biographies



Sundara Srivathsan M. is a final-year undergraduate student in Electronics and Computer Engineering at Vellore Institute of Technology, Chennai. His research focuses on computer vision and decision-theoretic machine learning. He has contributed to government-collaborated projects, is a national hackathon winner, and has participated in IEEE VIS and ACM CSCW 2025. He is a recipient of the Sir C. V. Raman Award for Research Excellence at VIT.



Lighittha P. R. is a final-year undergraduate student in Electronics and Computer Engineering at Vellore Institute of Technology, Chennai. Her research focuses on secure and privacy-preserving AI systems, federated learning, and cost-sensitive attestation frameworks. She has contributed to government-collaborated projects and Indo–Sri Lankan collaborations, is a national hackathon winner and finalist, and has participated in IEEE VIS and ACM CSCW 2025.

Prithivraj S. is a final-year undergraduate student in Electronics and Computer Engineering at Vellore Institute of Technology, Chennai, specializing in deep learning, high-performance computing, and secure distributed architectures. His research spans blockchain technology, federated learning, and computer vision. He has contributed to government-collaborated projects and Indo–Sri Lankan collaborations, is a national hackathon winner, and has participated in IEEE VIS and ACM CSCW.

Suganya Ramamoorthy is a Professor in the School of Computer Science and Engineering at Vellore Institute of Technology, Chennai, with over 18 years of experience in computing. Her research interests include medical imaging, artificial intelligence, computer vision, big data, and engineering education. She is the author of the book *Big Data in Medical Imaging* and has published extensively in peer-reviewed international journals. Dr. Suganya is an active member of the Association for Computing Machinery (ACM), contributing through technical lectures and serving as a reviewer. Her recognitions include the 17th Young Women in Engineering Award (2017) and the Best Mentor Award in the Women in Big Data (WiBD) network (2023).

Vijayan Sugumaran is a Distinguished University Professor and Janke Scholar of Management Information Systems at Oakland University, Rochester, Michigan, USA. He serves as Chair of the Department of Decision and Information Sciences, Co-Director of the Institute for Data Science, and Director of the M.S. in Business Analytics program. He received his Ph.D. in Information Technology from George Mason University, and his research interests include big data analytics, ontologies and the Semantic Web, and intelligent agent systems. He has authored over 350 peer-reviewed publications and 20 edited books, serves on the editorial boards of eight journals, and is Editor-in-Chief of the *International Journal of Intelligent Information Technologies* and the *Journal of Web Engineering*. He is Co-PI on a $2 million NSF grant and has published in leading journals such as *Information Systems Research*, *ACM Transactions on Database Systems*, and several *IEEE Transactions* venues.