ARTEMIS: Smart-Contract–Driven In-Loop Governance for Copyright-Compliant Diffusion-Model Training

Aziza K. Shukurova, Nishanthi Rengasamy, Mong Diem Quynh, Mohammed M. A. Ahmed
Advising professor: Albert No
Industry-Academia mentor: Lee Yu Jeong, AWS
Department of Computer Science
Yonsei University

***Abstract*—Diffusion models (e.g., Stable Diffusion [1], Imagen) achieve state-of-the-art image synthesis by training on massive web-scraped datasets—often without clear provenance or licensing metadata. This has spurred high-profile copyright suits (e.g., *NYT v. OpenAI*, *Advance Local Media v. Cohere*), demanding granular audit trails. Existing remedies—dataset filtering [2], post-hoc concept erasure [3], [4], watermarking [5], [6], and immutable registries [7], [8]—address only fragments of the problem and lack *in-loop*, epoch-by-epoch enforcement. We present ARTEMIS (AI Rights & Traceability Ecosystem for Media Integrity Systems), the first framework to embed smart-contract–driven policy enforcement, concept erasure directives, and zero-knowledge proofs directly into the diffusion training loop on Polygon Mumbai. ARTEMIS uses Solidity contracts for policy registration, an off-chain oracle for per-epoch null-space gradient filtering, and zk-SNARKs for aggregated proof verification (via Layer-2 rollups), enabling continuous, transparent, and provably compliant diffusion training.**

***Index Terms*—Diffusion models, copyright, smart contracts, zk-SNARKs, blockchain, CLIP, Layer-2 rollups, training transparency**

I. Introduction & Motivation

Diffusion models learn complex image distributions via iterative denoising [1], [9]. However, their reliance on billions of web-scraped images—often lacking explicit licenses or provenance—raises serious copyright and ethical concerns. Lawsuits like *New York Times Co. v. OpenAI* (GPT memorization) and *Advance Local Media v. Cohere* highlight courts' demands for detailed data-lineage and enforceable usage policies.

Current technical solutions fall into four silos:

- **Dataset Filtering**: Pre-training curation (e.g., LAION-5B's NSFW and license filters) [2].
- **Concept Erasure**: Post-hoc methods (ESD [3], SPEED [4]) remove concepts after training but cannot prevent in-training violations.
- **Watermarking**: Embedding imperceptible marks in weights/outputs (Uchida *et al.* [5], Zhong *et al.* [6]) provides passive proof but no real-time enforcement.
- **Blockchain Registries**: Immutable on-chain metadata (IBIS [7], AIGC-Chain [8]) records but treats training as off-chain.

Gap: No system provides *in-loop* governance—dynamic, epoch-by-epoch policy enforcement—within the diffusion training pipeline.

Contributions:

1) **Smart Contracts on Polygon Mumbai:** Solidity contracts (`ConceptRegistry`, `ErasureRequest`, `ModelVersion`) register policies, erasure directives, and aggregated proof pointers.
2) **Off-Chain Oracle:** A Python service intercepts each epoch, fetches policies, and applies null-space gradient filtering (SPEED [4]).
3) **zk-SNARK Proofs & Layer-2 Rollups:** Circom circuits generate per-epoch proofs; proofs are aggregated off-chain and verified on-chain in a single transaction, minimizing gas costs.
4) **Evaluation Plan:** Deploy contracts, train Stable Diffusion v1.5 on a PD12M subset, and measure concept leakage (via CLIP), FID, proof latency, and gas costs.

II. Related Work

A. Data Provenance and Auditing

Data-provenance techniques trace which training samples influenced a model's parameters. Huang *et al.* propose a general auditing framework combining membership inference with sequential hypothesis testing, offering tunable false-detection guarantees without assumptions on dataset curation [10]. Buick argues transparency alone cannot substitute for robust copyright regimes [11]. Park's comparative legal analysis highlights divergent fair-use doctrines across jurisdictions, underscoring the need for transparent pipelines [12].

B. Concept Erasure

Concept-erasure methods remove unwanted visual concepts post-training. Gandikota *et al.* introduce Erased Stable Diffusion (ESD), a negative-prompt fine-tuning approach that permanently removes target concepts but suffers from collateral interference and scalability issues [3]. Li *et al.* present SPEED, leveraging null-space projections and prior filtering for scalable, precise, and efficient multi-concept erasure with

minimal collateral damage—but without dynamic, in-loop governance [4].

C. Watermarking and DRM

Digital-watermarking embeds imperceptible marks in model weights or outputs for ownership verification. Uchida *et al.* pioneered DNN-based weight watermarking, enabling passive post-training proof of ownership [5]. Zhong *et al.* survey deep-learning–based image watermarking techniques—categorizing embedder–extractor, feature-transform, and hybrid methods—yet these remain passive and lack real-time policy enforcement [6].

D. Blockchain-Based Governance

Blockchain registries provide immutable records of datasets, licenses, and model versions. Sai *et al.* present IBIS, an on-chain registry for dataset and license metadata, but treat training as an off-chain event [7]. Jiang *et al.* introduce AIGC-Chain for full-lifecycle AIGC product management, lacking embedded training-loop governance [8]. Albalwy *et al.* demonstrate dynamic consent via smart contracts in genomic data sharing (ConsentChain), though without ML integration [13].

E. Gap Analysis

No prior work embeds dynamic, per-epoch governance within diffusion training. ARTEMIS fills this void by unifying on-chain policies, off-chain enforcement, and zk-SNARK proofs with Layer-2 rollups.

III. ARTEMIS Framework

A. Threat Model & Trust Assumptions

Adversaries:

- *Data Scrapers*: Ingest copyrighted images without consent.
- *Model Hijackers*: Bypass erasure to retain forbidden concepts.
- *Audit Forgers*: Tamper with logs to conceal violations.

Trust Assumptions:

- Oracle runs in a TEE (e.g., Intel SGX or AWS Nitro) [14].
- Layer-2 aggregator is semi-trusted; on-chain verification ensures integrity.

B. System Architecture

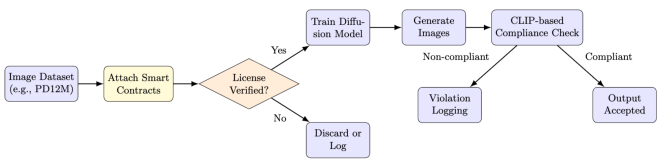

Fig. 1. ARTEMIS: a simplified workflow.

C. Diffusion-Centric Policy Enforcement

At each timestep t, the oracle computes gradients $\nabla_\theta \mathcal{L}(z_t, \theta)$. Let $\mathbf{C} \in \mathbb{R}^{d \times k}$ be prohibited concept embeddings. We project:

$$\mathbf{P} = \mathbf{I} - \mathbf{C}(\mathbf{C}^\top \mathbf{C})^{-1}\mathbf{C}^\top, \tag{1}$$

$$\nabla_\theta^{\text{filtered}} = \mathbf{P}\,\nabla_\theta \mathcal{L}(z_t, \theta). \tag{2}$$

This ensures forbidden features are never reinforced [4].

Algorithm 1 ARTEMIS-Enhanced Diffusion Training

Input: Dataset $\mathcal{D}$, policies $\mathcal{P}$, diffusion model M, epochs E
for $e = 1$ to E **do**
  $\mathcal{P}_e \leftarrow$ fetchPolicies(`ConceptRegistry`)
  for each batch B from $\mathcal{D}$ **do**
    compute gradients $g \leftarrow \nabla_M \mathcal{L}(B)$
    if concepts in g flagged by $\mathcal{P}_e$ **then**
      { [4]} $g \leftarrow$ nullSpaceProject(g, flaggedConcepts)
    end if
    update M with g
  end for
  $\pi_e \leftarrow$ generateZKProof(M, $\mathcal{P}_e$)
  commitReceipt(`ModelVersion`, π_e)
end for

D. zk-SNARK Circuit

The Python oracle will leverage HuggingFace's `diffusers` and apply SPEED-style nullification [4]. The circuit proves Equation (2) satisfies $\mathbf{C}^\top \nabla_\theta^{\text{filtered}} = \mathbf{0}$, attesting compliance without revealing θ.

IV. Implementation & Evaluation

A. Experimental Setup

Model: Stable Diffusion v1.5 (870M params) [1]. **Dataset:** 10K public-domain images from PD12M [15], with 1K flagged for concept erasure. **Policies:** 50 dynamic rules (e.g., "Remove watercolor style," "Block celebrity faces"). **Hardware:** 2×NVIDIA RTX3090 GPUs; Polygon Mumbai testnet; IPFS for storage.

B. Evaluation Metrics

1) **Concept Leakage Rate:** Fraction of generated samples where CLIP similarity to a forbidden concept exceeds threshold τ (we use $\tau = 0.4$) [2].
2) **Baseline Comparison:** ESD-only erasure vs. ARTEMIS in-loop enforcement.
3) **Image Quality:** FID [16] on a held-out set to measure quality trade-off.
4) **Proof Overhead:** Time to generate & verify zk-SNARK proofs; gas cost per aggregated proof on Mumbai.
5) **End-to-End Cost:** Total gas for 50 epochs using Layer-2 rollup (0.225 MATIC $0.04) vs. naive on-chain per-epoch (0.225 MATIC).

C. Scalability via Layer-2 Rollups

To mitigate zk-SNARK overhead, we aggregate proofs off-chain and verify a single succinct proof on-chain per epoch, leveraging Optimistic or ZK rollups [17]. This reduces on-chain transactions and gas fees while preserving auditability.

V. Discussion & Future Work

We've demonstrated feasibility on a small PD12M subset; scaling to 512×512 or larger models will require:

- *Recursive SNARKs* to aggregate across epochs [17].
- *Off-chain MPC or trusted setups* for oracle integrity [14].
- *Extension to LLMs* for text-based copyright governance.

VI. Conclusion

ARTEMIS is the first framework to embed smart-contract–driven, in-loop governance into diffusion-model training on Polygon Mumbai. By unifying on-chain policy registration, off-chain null-space enforcement, and zk-SNARK proof aggregation via Layer-2 rollups, ARTEMIS ensures continuous, transparent, and provably compliant generative AI.

VII. Planned Roles and Responsibilities

- **Aziza K. Shukurova** will lead the project, coordinate team efforts, and oversee the design, implementation, and writing process.
- **Mohammed M. A. Ahmed** will handle smart contract development, assist with zero-knowledge proof integration, and support testing on the Polygon testnet.
- **Nishanthi Rengasamy** will conduct background research, help draft policies, and contribute to writing and presenting the final paper.
- **Mong Diem Quynh** will assist with oracle setup, proof generation components, and contribute to documentation and slide preparation.

VIII. Project Timeline

- Week 1 (Apr 15–21)**:** Finalize project scope and requirements.
- Week 2 (Apr 22–28): Design system architecture and components.
- Week 3 (Apr 29–May 5): Develop smart contracts and blockchain setup.
- Week 4 (May 6–12): Implement off-chain oracle and integrate with training loop.
- Week 5 (May 13–19): Develop zk-SNARK circuits and proof generation.
- Week 6 (May 20–26): Integrate all components and perform initial testing.
- Week 7 (May 27–Jun 2): Conduct comprehensive evaluation and benchmarking.
- Week 8 (Jun 3–9)**:** Prepare documentation and finalize the report and project.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[2] C. e. a. Schuhmann, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS Datasets and Benchmarks*, 2022.

[3] R. e. a. Gandikota, "Erasing concepts from diffusion models," in *ICCV*, 2023.

[4] O. e. a. Li, "Speed: Scalable, precise, and efficient concept erasure for diffusion models," *arXiv preprint arXiv:2503.07392*, 2025.

[5] Y. e. a. Uchida, "Embedding watermarks into deep neural networks," *arXiv preprint arXiv:1701.04082*, 2017.

[6] X. e. a. Zhong, "A brief, in-depth survey of deep learning-based image watermarking," *Applied Sciences*, vol. 13, no. 21, p. 11852, 2023.

[7] Y. e. a. Sai, "Is your ai truly yours? leveraging blockchain for copyrights, provenance, and lineage," *arXiv preprint arXiv:2404.06077*, 2024.

[8] J. e. a. Jiang, "Aigc-chain: A blockchain-enabled full lifecycle recording system for aigc product copyright management," *arXiv preprint arXiv:2406.14966*, 2024.

[9] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[10] Z. Huang, N. Z. Gong, and M. K. Reiter, "A general framework for data-use auditing of ml models," in *ACM CCS*, 2024.

[11] A. Buick, "Copyright and ai training data—transparency to the rescue?" *Journal of Intellectual Property Law & Practice*, 2024, vol. 20, No. 3.

[12] S. H. Park, "Copyright issues for ai and deep learning services: A comparison of u.s., south korean, and japanese law," JTIP Blog, 2021.

[13] F. Albalwy, A. Brass, and A. Davies, "A blockchain-based dynamic consent architecture to support clinical genomic data sharing (consentchain): Proof-of-concept study," *JMIR Medical Informatics*, vol. 9, no. 11, p. e27816, 2021.

[14] A. A. Bellachia *et al.*, "Verifbfl: Leveraging zk-snarks for a verifiable blockchained federated learning," *arXiv preprint arXiv:2501.04319*, 2025.

[15] J. Meyer *et al.*, "Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms," *arXiv preprint arXiv:2410.23144*, 2024.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[17] H. Sun, T. Bai, J. Li, and H. Zhang, "zkdl: Efficient zero-knowledge proofs of deep learning training," *arXiv preprint arXiv:2307.16273*, 2023.