DECORAIT - DECentralized Opt-in/out Registry for AI Training

Kar Balan
k.balan@surrey.ac.uk
University of Surrey
Guildford, UK

Alex Black
a.black@surrey.ac.uk
University of Surrey
Guildford, UK

Andrew Gilbert
a.gilbert@surrey.ac.uk
University of Surrey
Guildford, UK

Simon Jenni
jenni@adobe.com
Adobe Inc.
San Jose, US

Andy Parsons
andyp@adobe.com
Adobe Inc.
San Jose, US

John Collomosse
collomos@adobe.com
Adobe Inc.
San Jose, US

Figure 1: DECORAIT enables creatives to register consent (or not) for Generative AI training using their content, as well as to receive recognition and reward for that use. Provenance is traced via visual matching, and consent and ownership registered using a distributed ledger (blockchain). Here, a synthetic image is generated via the Dreambooth[Ruiz et al. 2022] method using prompt "a photo of *[Subject]*" and concept images (left). The red cross indicates images whose creatives have opted out of AI training via DECORAIT, which when taken into account leads to a significant visual change (right). DECORAIT also determines credit apportionment across the opted-in images and pays a proportionate reward to creators via crypto-currency micropyament.

ABSTRACT

We present DECORAIT; a decentralized registry through which content creators may assert their right to opt in or out of AI training and receive rewards for their contributions. Generative AI (GenAI) enables images to be synthesized using AI models trained on vast amounts of data scraped from public sources. Model and content creators who may wish to share their work openly without sanctioning its use for training are thus presented with a data governance challenge. Further, establishing the provenance of GenAI training data is important to creatives to ensure fair recognition and reward for their such use. We report a prototype of DECORAIT, which explores hierarchical clustering and a combination of on/off-chain storage to create a scalable decentralized registry to trace the provenance of GenAI training data to determine training consent and reward creatives who contribute that data. DECORAIT combines distributed ledger technology (DLT) with visual fingerprinting, leveraging the emerging C2PA (Coalition for Content Provenance and Authenticity) standard to create a secure, open registry through which creatives may express consent and data ownership for GenAI.

CVMP '23, November 30-December 1, 2023, London, United Kingdom

ACM ISBN 979-8-4007-0426-0/23/11…$15.00
https://doi.org/10.1145/3626495.3626506

CCS CONCEPTS

• **Applied computing** → **Document management**; • **Information systems** → **Data provenance**; • **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

Content provenance, Distributed ledger technology (DLT/Blockchain), Generative AI, Data governance.

ACM Reference Format:
Kar Balan, Alex Black, Andrew Gilbert, Simon Jenni, Andy Parsons, and John Collomosse. 2023. DECORAIT - DECentralized Opt-in/out Registry for AI Training. In *European Conference on Visual Media Production (CVMP '23), November 30-December 1, 2023, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626495.3626506

1 INTRODUCTION

Generative AI (GenAI) models such as ChatGPT and Stable Diffusion [OpenAI [n. d.]; Stablity.ai [n. d.]] are transforming creative workflows through their ability to synthesize content given only high-level direction. GenAI models are typically trained by sampling millions of media items harvested from public data sources.

Yet this practice has raised concerns over data governance, specifically over creatives' agency to opt in or out of the use of their work for GenAI training. This has led to several legal challenges to the creators of GenAI models, stemming from the concerns over potential rights infringement – particularly of digital images. Furthermore, creatives are not currently enabled to receive recognition or reward for their contribution to GenAI images through the use of their content in training.

We envision a future creative economy for content delivery services, such as stock photography platforms, which enables the commercialization of creative content and the contribution of it to ethically and consensually built datasets for GenAI training. These platforms are responsible for enabling their users, contributors, and collaborators to express consent over their data being used in GenAI training in a secure, persistent, and interoperable way. Such capability is grounded in strong provenance signals for GenAI training data, that enable creatives to register ownership and means for payment for GenAI use as well as their consent for that use.

To this end, we propose DECORAIT, a decentralized registry for GenAI training that enables creators to express consent, or otherwise, for their images to be used in AI training, as well as enabling them to receive reward when such use occurs. Our work follows emerging community trends toward centralized, commercial opt-out services. For example, *spawning.ai* maintains lists of opted-out URL patterns (from individual links to entire domains). GenAI models can match against these lists to exclude content from training. However, a URL list may not capture all instances of a creator's content online. Moreover, scaling up multiple individually managed databases to track opt-out raises data consistency and interoperability challenges. The protocol of the future creative economy also ought to ensure the contributing creatives to GenAI can be recognized and rewarded for their creative assets when their particular content or style is identified to have contributed to specific synthetic media. DECORAIT addresses these issues through three contributions:

(1) We propose a **fingerprint-based content similarity score**, followed by a **credit apportionment scheme** to match images and reward creatives for their training content most correlated with generated synthetic media.
(2) A **sharded decentralized search index** using distributed ledger technology (DLT), in which a content fingerprint distilled from the image provides a key to register and robustly query opt-in/out information. We propose a hierarchical approach to scale vector search of this index and a hybrid on/off-chain approach to query processing.
(3) We leverage the emerging **Coalition for Content Provenance and Authenticity (C2PA)** standard to express consent and payment preferences via cryptographically signed asset 'manifests'. These manifests are stored within a distributed file system (IPFS) and referenced by hashed URL link via the DECORAIT DLT search index.

Without loss of generality, we demonstrate DECORAIT within the pipeline of Dreambooth [Ruiz et al. 2022] which enables specialization of diffusion models to generate novel renditions of a specific subject provided via exemplar training images. Dreambooth provides a suitable use case as it enables GenAI model users to assure the assets they intend to leverage for model personalization have been opted-in for AI training. Additionally, the proposed system enables the fair recognition and reward of those contributing creatives. We could imagine a future for stock photography in which contributors receive payments not only through direct licensing (as now), but automatically via DECORAIT's ability to provide downstream recognition and persistent crediting of the contributing creators to GenAI. Fair monetary reward is encouraged via our apportioning algorithm, coupled with the transparency and auditability of crypto-currency payments processed using DLT.

2 RELATED WORK

Distributed Ledger Technology (DLT), colloquially 'blockchain', ensures the immutability of data distributed across many parties without requiring those parties to trust one another or any central authority [Narayanan et al. 2016]. While the original and dominant use is cryptocurrency tokens (*e.g.* Bitcoin [Nakamoto 2008]), emerging use cases include digital preservation [Lemieux 2016], supply chain and media provenance [Holmes 2018; Walport 2015]. DLT has been used to track ownership of media via the ERC-721 Non-Fungible Token (NFT) standard [Bhujel and Rahulamathavan 2022], although NFT lacks a rights or permissions framework [Fairfield 2021]. Recently, Ekila explored tokenized rights in NFT [Balan et al. 2023]. DLT was analyzed for media integrity in ARCHANGEL [Collomosse et al. 2018]; digital records were hashed and used to tamper-proof archival records. Perceptual hashing extended ARCHANGEL from documents to images [Bui et al. [n. d.]] and videos [Bui et al. 2020]. Our work uses perceptual hashes for search; as a key to resolving image fingerprints to data on training consent. Recent advances in proof of stake and Layer 2 solutions scale DLT for improved throughput and reduced climate impact, yet scalable storage remains challenging. Peer-to-peer (p2p) distributed file-sharing technologies such as the Interplanetary File System (IPFS [Benet 2014]) are used to address this.

C2PA is an emerging metadata standard for embedding content provenance information ('manifests') in media files ('assets') [Coalition for Content Provenance and Authenticity 2021]. Manifests are signed via public-key pair and describe facts about asset provenance, such as who made it, how and using which ingredient assets. These facts are called 'assertions'. C2PA initially focused on trusted media [Rosenthol et al. 2020] and journalism [Aythora et al. 2020] use cases. Recently, C2PA (v1.3) described a `training-mining` assertion in which creators may set flags to opt in or out of GenAI training, which we leverage in our work. Unfortunately, C2PA metadata is stripped by non-compliant platforms (*e.g.* social media) or attackers. Therefore, we use perceptual hashing to match content to manifests.

Content Fingerprinting identifies content robustly in the presence of degradation or rendition (format, quality, or resolution change) and minor manipulation. Perceptual hashing [Bharati et al. 2021; Black et al. 2021; Nguyen et al. 2021] and watermarking [Bui et al. 2023; Devi et al. 2019] have been used to match content. Fingerprinting has also been used to detect and attribute images to the GenAI models that made them [Yu et al. 2021].

Diffusion models lay at the foundation of most recent advances in GenAI [Ho et al. 2020; Podell et al. 2023; Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022]. Such models are commonly trained on millions or even billions of images to gain the ability to synthesize diverse and high-quality images consistently. Diffusion models have shown substantially superior performance in comparison to GANs [Dhariwal and Nichol 2021]. However, they have also been shown to memorize content and style from training data to a higher degree than GANs [Somepalli et al. 2022], phenomenon attributed to the presence of duplicated image data [Carlini et al. 2023; Somepalli et al. 2022] within the training data. [Somepalli et al. 2023] showed that content and style memorization is an even greater concern, specifically in text-conditioned diffusion models, due to duplicated captions within the training data, with data replication not commonly occurring in unconditional diffusion models. This further accentuates the need to involve creatives and obtain consent to use their creative content in the GenAI training pipeline. The present work lays the groundwork for such a system, querying and registering the creatives' opt-in or out decision on GenAI training and offering a pipeline to reward creatives for using those assets in GenAI.

Model personalization methods are techniques which enable diffusion models to be customized to synthesize novel renditions of a specific subject in different contexts. Recently, training-free adaptation [Shi et al. 2023] and fine-tuning [Kumari et al. 2023; Ruiz et al. 2022] have been explored to customise object instances. In this work, we utilize the Dreambooth [Ruiz et al. 2022] technique for model personalization, which fine-tunes a pre-trained text-to-image diffusion model - the base model - using a small set of 'concept' images depicting a specific subject. The subject is thus embedded in the output domain of the model which learns to bind it to a unique identifier (token), which can then be used as part of the prompt to synthesize the subject in new and diverse contexts. We use Dreambooth to demonstrate the DECORAIT system, aiding in the training pipeline by identifying opted-in assets from a stock photography website available to train a personalized instance of a diffusion model.

3 TRACING AND DESCRIBING IMAGE PROVENANCE

We begin by describing how images are matched to trace visual provenance. We use this approach to 'fingerprint' training images to robustly match to entries in the DECORAIT registry, thereby accessing data on consent status and creator wallet addresses which are encoded via the C2PA open standard (subsec. 3.2). A second pair-wise model enables both verification of such matches and correlation between synthetic and training data for credit apportionment.

3.1 Fingerprinting and match verification

To reliably match training images at scale, we employ two modules. First, a contrastively trained model to extract compact embeddings for measuring image similarity, and second a model which classifies whether the closest matching images in that embedding are true matches. The latter is motivated by the difficulty of thresholding image similarity distances at scale whilst retaining practical accuracy levels. The classifier probability serves as a match verification check and a score to drive credit apportionment.

3.1.1 Fingerprinting Model. We adapt the fingerprinting technique described in [Black et al. 2021] to obtain compact embeddings of the images within the registry's corpus, allowing robust visual content attribution and search. The resulting fingerprint is a compact embedding (256-D) of a CNN, contrastively trained to be discriminative of image content whilst robust to image degradations and manipulations to model content transformations common as images are shared online. The model is trained through a contrastive learning objective [Chen et al. 2020]. Let $\phi_i = E(x_i) \in \mathbb{R}^{256}$ be the feature vector obtained as the output of a ResNet-50 encoder for an image x_i and $\hat{\phi}_i$ represent an embedding of a differently augmented version of x_i. The training objective is given by

$$\mathcal{L}_C = -\sum_{i \in \mathcal{B}} \log\left(\frac{d\left(\phi_i, \hat{\phi}_i\right)}{d\left(\phi_i, \hat{\phi}_i\right) + \sum_{j \neq i \in \mathcal{B}} d\left(\phi_i, \phi_k\right)}\right), \tag{1}$$

where $d(a, b) := \exp\left(\frac{1}{\lambda} \frac{a^\top b}{\|a\|_2 \|b\|_2}\right)$ measures the similarity between the feature vectors a and b, and $\mathcal{B}$ is a large randomly sampled training mini-batch [Balan et al. 2023].

In terms of data augmentation, in addition to the typical techniques used in contrastive learning such as colour jittering and random cropping, we consider minor manipulations, benign modifications and degradations of image content due to noise, format change and recompression, resolution change (resize), and several other degradation manipulations studied in [Hendrycks and Dietterich 2019]. This is because images may be reshared online and subject to many such transformations and renditions, and we wish to match regardless.

3.1.2 Verification and Apportionment Model. Provided a shortlist of the top-K candidate matches from the previous fingerprinting step, we verify image matches through an additional pair-wise comparison between the query image and each candidate match retrieved. The spatial feature maps derived from the fingerprinter model are used to compare the images as follows.

Let $F_q \in \mathbb{R}^{H \times W \times D}$ be the feature map for a query image x_q and let $\{F_i\}_{i=1}^k$ be the k corresponding retrieval feature maps. Each feature map is processed with a 1×1 convolution to reduce the dimensionality to $\frac{D}{4}$ and then numerous pooled descriptors from a set of 2D feature map windows $\mathcal{W} \subset [1, H] \times [1, W]$ are extracted, similar to R-MAC [Tolias et al. 2015]. Let $f_w^q \in \mathbb{R}^{\frac{D}{4}}$ denote the GeM-pooled [Tolias et al. 2015] and unit-normalized feature vector for a window $w \in \mathcal{W}$ and feature map F_q. In contrast to [Tolias et al. 2015], the window-pooled feature vectors are not averaged, but collected as:

$$\hat{F}_q = [f_{w_1}^q, \ldots, f_{w_{|\mathcal{W}|}}^q] \in \mathbb{R}^{|\mathcal{W}| \times \frac{D}{4}}, \tag{2}$$

where $w_i \in \mathcal{W}$ and the number of windows is $|\mathcal{W}| = 55$ in practice. The feature correlation matrix is then computed as:

$$C_{qi} = \hat{F}_q \hat{F}_i^T \in \mathbb{R}^{|\mathcal{W}| \times |\mathcal{W}|}. \tag{3}$$

These feature correlations are then flattened and fed to a 3-layer MLP, which outputs a similarity score between query q and retrieval

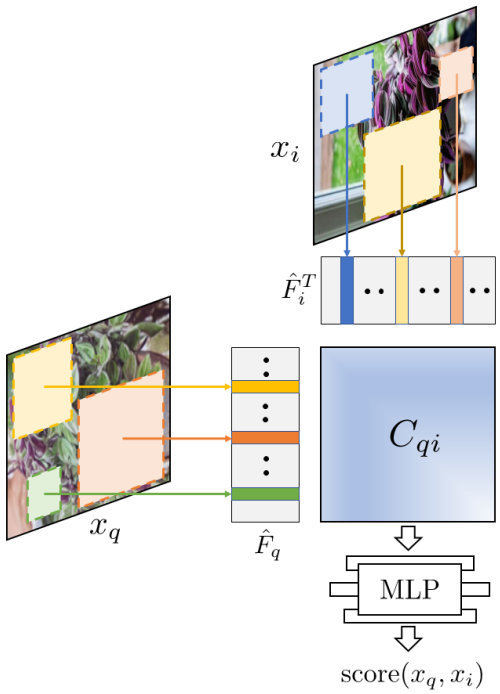

Figure 2: Match Verification Model. Two images are compared at multiple scales to robustly find (partial) matches. The model extracts multiple aggregated feature vectors from the two feature maps corresponding to numerous image patches of different sizes and positions. These features (collected in $\hat{F}_q$ and $\hat{F}_i$) are then used to compute the feature correlation matrix C_{qi}, which is fed to an MLP to compute a final score.

i. To make the model symmetric w.r.t. its inputs, the match score between images x_q and x_i is defined as

$$\text{apportion}(x_q, x_i) = \sigma\big(\text{MLP}(C_{qi}) + \text{MLP}(C_{iq})\big), \tag{4}$$

where σ represents a sigmoid activation. The model is illustrated in Fig. 2.

To train the model, positive example pairs are built via a strong data augmentation protocol, similar to the data augmentation step in the fingerprinter model training. This protocol includes colour jittering, blurring, random resize cropping, and random rotations. A hard negative mining approach is used to generate challenging negatives.

For the sampling of negatives, the global average-pooled feature maps of query and queued examples are compared via cosine similarity. Given pairs of true and false matches, the model is trained with a standard binary cross-entropy loss. During training, the backbone feature extractor from the fingerprinter model is frozen.

3.2 Encoding consent and ownership

The Coalition for Content Provenance and Authenticity (C2PA) standard aims to aid internet users' trust decisions about digital assets they might come across on platforms such as social media or news websites. Recent work also employs C2PA as a tool to encode provenance information within synthetically generated media, including within its metadata details about the model used to create it, as well as its training data [Balan et al. 2023].

A 'manifest' is a data packet that may be bound to digital assets at creation time or post-factum. This manifest embeds facts about the provenance of a digital asset within its metadata. These facts are referred to as 'assertions'. They may include information such as who created the asset, how it was made, what hardware and software solutions aided in its creation, and any edits it may have undergone since its creation. This data is cryptographically signed to prevent tampering. Signing C2PA manifests requires that the signer uses their private key and public certificates, following the Public Key Infrastructure (PKI). This assures that the consumer makes trust decisions about the asset based on the identity of the manifest signer. A certification authority (CA) conducts real-world verification to ensure signing credentials are only issued to trusted, non-malicious actors [Coalition for Content Provenance and Authenticity 2021].

Additionally, C2PA manifests may bear information about other "ingredient" assets used in the creation process. These ingredients may point at assets, each bearing its own C2PA manifest describing its provenance. As such, C2PA encodes a graph structure with the root at the current asset and branching out to its ingredient assets. The C2PA standard describes this ingredient model in terms of creation of classical images (and other media assets) but we use it in DECORAIT to describe how Dreambooth models may be created from their training concept images, and how synthetic images are created from their Dreambooth model as an ingredient.

Recently, C2PA (v1.3) introduced several training-mining assertions in which creators may set flags to opt in or out of GenAI training within manifests. These flags are *data_mining, ai_inference, ai_generative_training* and *ai_training*. We leverage these flags to encode consent in DECORAIT.

C2PA manifests also support the inclusion of DLT-based wallet addresses. For example, in Adobe Photoshop, any DLT wallet address linked to a user's Adobe identity may be recorded in the C2PA metadata of an exported image. In the following sections, we show how this wallet information, embedded immutably within assets at creation-time, may be leveraged to reward creatives when their images are used to train GenAI.

4 DECORAIT SYSTEM ARCHITECTURE

DECORAIT is a decentralized search index, performing *key-value* lookups using a robust image fingerprint (subsec. 3.1.1) as the *key*. The *value* is a URI, resolvable to a C2PA manifest indicating permission to train. A scalable solution demands: 1) persistent distributed storage of manifests; 2) a distributed and immutable lookup operated via an open model without recourse to a centralized trust. Fig. 3 provides an overview of DECORAIT, which addresses these decoupled requirements by: 1) storing manifests on IPFS, where URIs are formed using a CID – a bit-wise (SHA-256 [Gilbert and Handschuh 2003]) content hash; 2) using a hybrid on/off-chain solution to create a sharded search index (subsec. 4.1). In Sec. 5, we explore empirical trade-offs in defining the boundary between on and off-chain computation for the search and the optimal level of sharding.

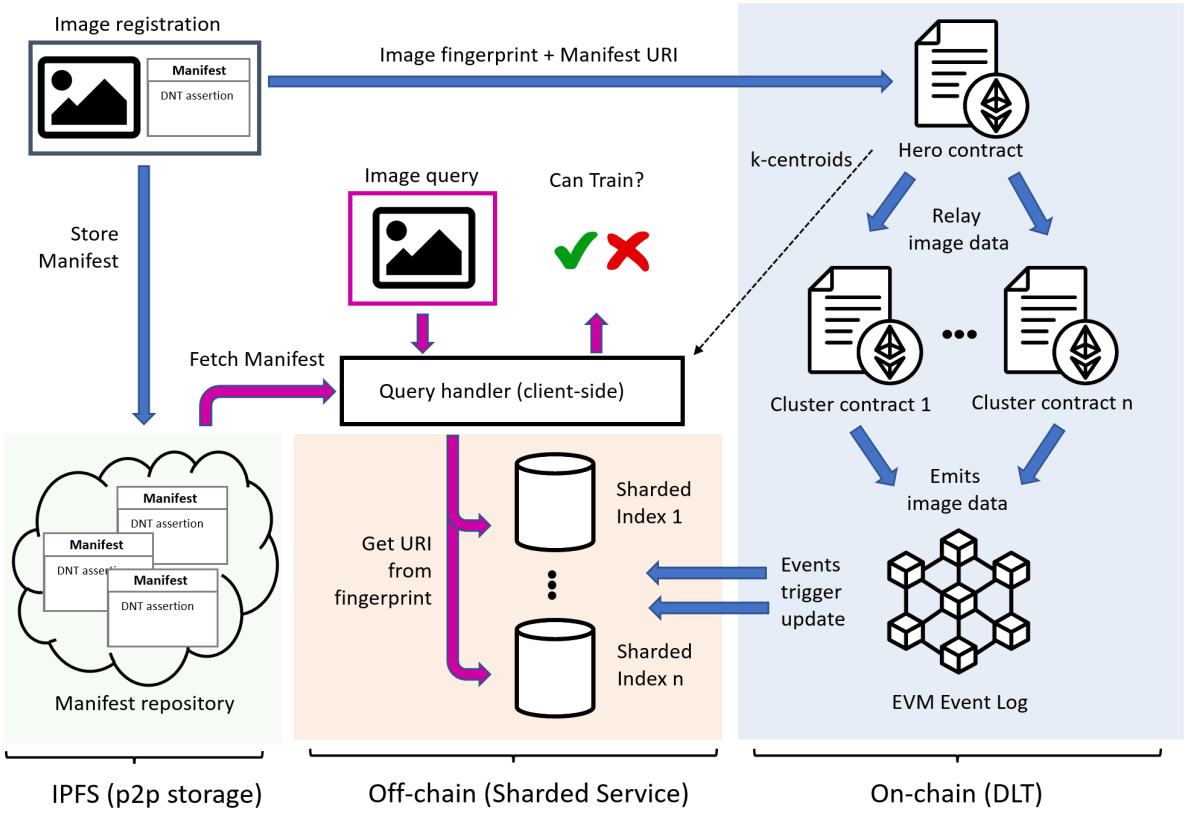

Figure 3: Overview of DECORAIT. 1) Ingest (blue): An image is fingerprinted by the client, and the hash is passed to the Hero contract, which determines on-chain which of the sharded (cluster) contracts will handle the ingest. The cluster contract emits an event recording the fingerprint (key) and IPFS URI (value) of the C2PA manifest, which the client stores on IPFS. The relevant off-chain sharded index listens for on-chain updates from its respective contract. 2) Query (pink): An image is fingerprinted by the client, and k-centroid data is used to determine which index shard to query with the fingerprint (key) to obtain the C2PA manifest URI (value). The client decides on whether GenAI training is permitted using the manifest. The diagram reflects the recommended variant (E-FOF) of DECORAIT (*c.f.* Table 1).

4.1 Decentralized Fingerprint Index

All images within DECORAIT undergo visual fingerprinting using the approach outlined in Sec. 3.1.1 to enable large-scale retrieval of visually similar assets upon querying the registry. We adopt a hierarchical approach to share the search index, applying k-means clustering to fingerprints computed from a representative (1M) image sample. The resulting k-centroids subdivide the fingerprint hash space into k shards. Recursive sharding is possible, but experiments focus on a single level. We shard the index using $k + 1$ DLT smart contracts deployed on a local Ethereum test-net; one contract per each of the k clusters, plus a single entry point – the 'hero' contract – to orchestrate the sharding. The hero contract performs the k-NN assignment of fingerprints to the k-centroids, delegating operations (*e.g.* ingest, query) to be handled by the smart contract of the closest cluster (and so, shard).

The contracts are implemented in Solidity, which does not support floating point math. We convert the 256-D floating point fingerprinting embeddings into integers as fixed-point (10^{15} precision), a workaround for applying ML operations on DLT [Harris and Waggoner 2019].

4.2 Hybrid on/off-chain variants

We explore several design choices for implementing our system, evaluating three main variants (Table 1). In particular, we explore options for persisting the key-value store (here used to map image fingerprints to manifest URIs) and on/off-chain options for implementing the shard assignment and retrieval processes.

4.2.1 Image Fingerprints and Data Storage. DLT storage patterns commonly persist data in two main ways: 1) *in-contract i.e.* within the state of a smart contract (as with NFTs), or 2) on the *event log*, a ledger of signals/exceptions emitted from smart contract code (as with cryptocurrency transactions). In our experiments, we use mnemonics prefixed E- to indicate variants using the event log and C- to indicate variants using in-contract storage.

Shards are described by k-centroid data from clustering in fixed point (256 integers) form. Fingerprints are similarly represented.

These 256-D data are stored as strings in the event log but may be stored in integer arrays for in-contract storage. There is cost efficiency in storing strings over integer arrays. However, there is a time cost in converting the strings for fixed point operation during ingest and query. The transaction cost implications are quantified in subsec. 5.4.

Table 1: Configuration of the three implementation variants.

Action	C-OOO	E-OOF	E-FOF
Key-value storage	in contract	event log	event log
Shard centroid prediction on query	on-chain	on-chain	off-chain
Shard prediction on ingest	on-chain	on-chain	on-chain
Retrieval with-in shard	on-chain	off-chain	off-chain

4.2.2 Shard assignment and retrieval. To store (ingest) or retrieve (query) a key-value pair, it is necessary to match the fingerprint to its shard via a k-NN assignment operation against the k-centroids obtained during initial clustering. In the case of queries, the retrieval is performed by matching against each key (fingerprint) stored within the key-value store of that shard. In Table 1, we use 'O' to indicate on-chain and 'F' to indicate off-chain computation for each matching operation and compare the efficiency of these variants in Sec. 5.

4.2.3 Smart contract interaction. The hero contract receives all operations and transactions in all variants. When ingesting a fingerprint to the registry, the hero contract reads it and calls the respective shard contract, which stores the key-value pair within its own contract (C-) or the event log (E-). In all cases, the smart contract performs the sharding via on-chain operations, which safeguards the integrity of the shards against inaccurate or malicious additions that could otherwise "infect" the clusters.

When querying a fingerprint, the on-chain variant (C-OOO) proceeds similarly - determining the shard and delegating the retrieval process to the relevant smart contract. The retrieval is performed on-chain in this case. In variants E-OOF and E-FOF, the query processing is partly delegated to off-chain processes. A web service is provided for each shard, which listens to the event log emitted by the smart contract of its respective shard. When submitting a query, the hero smart contract determines the appropriate shard index to call. This may be done on-chain (per the ingestion flow) or off-chain using k-centroid data from the hero contract. The relevant shard's web service performs the retrieval in both cases. Using off-chain processing mitigates computational costs as the index scales, as we now show. Fig 3 shows the interaction of the web service and smart contracts.

4.3 DECORAIT in the GenAI Workflow

We now describe how DECORAIT integrates with the GenAI training process to determine consent and how subsequently generated synthetic images may be traced to pay a reward to the creators who contributed that training data.

4.3.1 Training Consent. To ensure the creatives who authored the images have consented to their assets being used for training, each image is queried against the DECORAIT registry. As described in subsec. 4.1 the fingerprint embedding is used to identify the closest visual matches within the decentralized search index. These are verified using the apportionment model (subsec. 3.1.2) to obtain the closest match and so, a decision on training consent for each of the images. As described in subsec. 3.2, this information is embedded within the C2PA manifest accompanying each image on the DECORAIT system. We envision a future in which stock photography sites might parse and show this consent information by default, enabling users to select only the opted-in images when sourcing data for GenAI training.

4.3.2 Encoding Synthetic Image Provenance. We further leverage the C2PA standard to encode the provenance of the newly generated synthetic image, cryptographically tying it to the "ingredient" set of concept images and GenAI model. Thus, using the C2PA manifest of the generated image it is possible to trace both the model that generated it (the personalized model, and in turn, its base model), as well as the data used to personalize it. Specifically, the C2PA "ingredient" assertion is used to indicate the image dataset as ingredients to the fine-tuned model, as well as the base model. The personalized, fine-tuned model is then listed as an ingredient within the manifests of any synthetic images it generates. Thus, the synthetic image is tied to its ingredient assets listed above. This offers a complete creation provenance chain, immutably signed at creation time. Although in the case of finetuning models, the images from the dataset are individually included in the manifest, C2PA allows for manifests to be defined over archives of image collections for larger datasets. Figure 4 visualizes this relationship.

4.3.3 Apportionment and Payment. Given a synthetic image, DECORAIT enables credit to be assigned across training data images in order to recognize and reward contribution. The set of training image ingredients is first identified by traversing the image's provenance graph, rooted in the manifest of the synthetic image. Similar to the training stage, DECORAIT again uses visual fingerprinting to perform matching within the decentralized search index to lookup the C2PA manifest of each training image — including the DLT wallet address of each image's creator.

Credit is then assigned to each image proportional to a pair-wise score predicted by the apportionment model of subsec. 3.1.2: given a synthetically generated image $X_q \in \mathbb{R}^{H\times W\times 3}$, the visual similarity of each training image X_i in the identified concept set is scored via eq. 4 yielding a weighting w_i:

$$w_i = \max\big(\text{apportion}(X_q, X_i) - \lambda, 0\big), \tag{5}$$

where $\lambda = 0.7$ is an empirically set threshold for the visual similarity. We then assign credit per image by normalizing these weights over all top-image matches in the concept set.

Once the credit apportionment has been determined, payments may be processed securely and transparently over DLT. The authors' wallet addresses are extracted from the C2PA manifest associated with each image (Figure 4, left).

In Sec.5.5 we demonstrate how the DECORAIT system can be applied to a Dreambooth training pipeline, querying the registry for training consent, computing the similarity score, and apportioning credit amongst the set of concept images for any given generation.

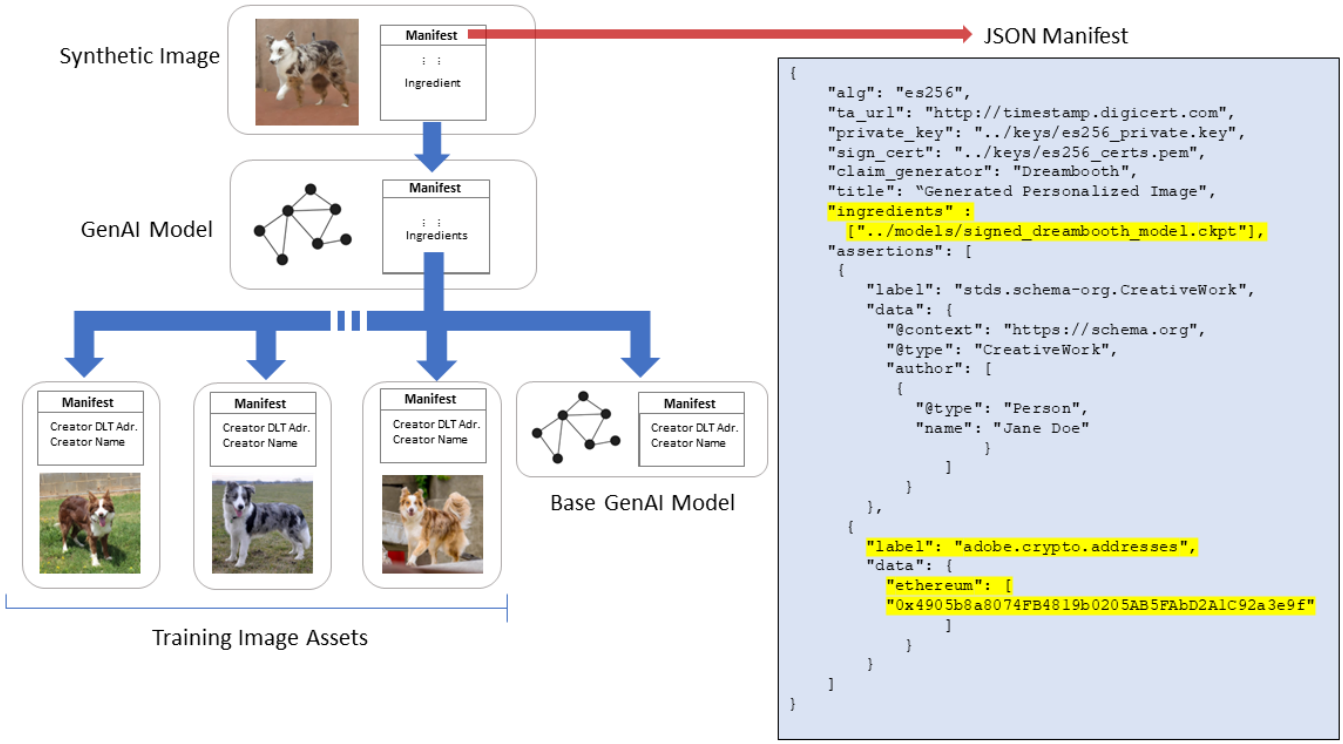

Figure 4: Provenance graph of synthetic media, as may be encoded via C2PA manifests. Left: starting from the generated image, the specialized Dreambooth model is listed as ingredient in its C2PA manifest. In turn, the Dreambooth model links to the specialization set of images and the base text-to-image Stable Diffusion model, which then may list as ingredient an archive of its entire training image corpus. Right: example JSON C2PA manifest accompanying a synthetic image, with highlighted ingredient and DLT wallet address assertions (the latter using the schema of a commercial image editor).

5 EVALUATION

We evaluate the relative performance and scalability of the three variants of the DECORAIT decentralized search index: C-OOO, E-OOF, E-FOF (*c.f.* Table 1) concluding on the most performant variant. We then demonstrate the proposed variant of the DECORAIT system as applied to the use case of a Dreambooth training pipeline and demonstrate querying the registry, resolving to a decision on training consent of the images within the set of concept images, followed by processing payments using DLT based on our apportionment algorithm.

5.1 Experimental Setup

We evaluate using the LAION400M dataset [Schuhmann et al. 2021], comprising image-text pairs crawled from publicly available web pages. LAION400M is extensively used to train GenAI models. For our experiments, we sample a training corpus of 1M images and sign these with C2PA manifests setting the `ai_generative_training`, `data_mining`, and `ai_training flags` to 'not allowed' to signify that the author has opted out of those images being used to train GenAI models.

The evaluation uses up to 1000 query images randomly sampled from the corpus, to which random augmentations are applied. The data augmentation process follows [Black et al. 2021]. It aims to mimic the perturbations an image may suffer from repeated use, upload, download, and compression on the internet (*e.g.* noise, and changes in resolution, quality, and format). In addition, we form a second query set of 100 unperturbed images. Lastly, we demonstrate the proposed DECORAIT system variant (E-FOF) within a Dreambooth model specialization pipeline.

5.2 Evaluating Accuracy vs Sharding

We evaluate the lookup's accuracy as a function of sharding (cluster count k) while maintaining a constant corpus size. The accuracy is agnostic to the on/off chain implementation of storage and query lookup, but the performance (query speed) varies significantly. Results are reported in table 2 for 1000 queries.

There is a trend to slightly reduced accuracy as sharding (k) increases due to the risk of heavy perturbations mis-assigning image fingerprints to adjacent shards. When no perturbation is present, the system performs with 100% accuracy for all shard counts. Yet, increasing sharding will reduce retrieval time (see below). On this basis, we select $k = 25$ as an appropriate sharding trade-off for the remainder of our experiments.

Table 2: Evaluating accuracy vs. shard count (k) for a 0.5M corpus size. The performance of on-chain shard prediction and within shard retrieval is studied for E-OOF. Shard prediction time increases, but retrieval time decreases as k increases.

Clusters (k)	Accuracy (%)	Cluster Prediction Time (on-chain)(s)	Retrieval Time (off-chain) (ms)
1(Baseline)	92.3	-	46.157
15	89.1	6.5	3.612
25	89.1	10.3	2.461
50	88.1	20.6	1.306
100	87.3	35.2	0.721
200	86.5	59.9	0.413
500	86	113	0.284
750	87.4	181.8	0.271
1000	86.4	272.9	0.249

5.3 Evaluating Performance vs Sharding

Retrieval speed varies significantly for each of our three variants and comprises two processes: closest centroid (shard) prediction and retrieval within the shard. We evaluate the speed of the nearest centroid prediction as a function of shard count (k). The number of distance computations (between the query embedding and each cluster centroid) scales linearly with k (Table 2), and this becomes prohibitive (several seconds) for high-frequency transactions (queries) at $k = 25$, though acceptable for bulk ingestion. This suggests that variant patterns x-Oxx are not scalable at query time.

Table 3: Evaluating in-contract storage (C-OOO) for accuracy and speed as corpus size increases. Shard count $k = 25$. Accuracy is good, but speed is poor relative to event-log variants.

Corpus	Accuracy perturbed (%)	Accuracy unperturbed (%)	Cluster prediction & KNN search time (on-chain) (s)
500	91.2	100	10.58
1000	92.4	100	19.72
5000	90.6	100	142.13
12500	90.8	100	295.38

Further, we evaluate the speed and accuracy of shard prediction and image retrieval as a function of our system's image corpus size for variants C-xxx and E-xxx. C-OOO stores the data and executes the lookup on-chain. In contrast, E-OOF/FOF emit the image data as events on the blockchain and performs image lookup, retrieval, and verification off-chain. Table 3 for C-OOO shows that the on-chain retrieval speed drops significantly as corpus size increases, suggesting C-OOO is unfit for GenAI contexts with large amounts of data. Table 4 shows that E-FOF maintains high retrieval accuracy as corpus size increases, with an average retrieval speed of just over 4 ms for a corpus size of 1M images. Tables 3 and 4 were measured for 500 queries. Further, we find that ingesting images for the system's initial setup takes an average of 683.2 ms per image in C-OOO, whereas E-FOF significantly improves speed requiring an average of only 81.5 ms per image.

We conclude that E-FOF exhibits scalability with corpus size and shard count, leading us to recommend variant E-FOF for the GenAI training opt-in/out task.

Table 4: Evaluating the recommended event-log storage variant (E-FOF) for accuracy and speed as corpus size increases, showing good scalability. Shard count $k = 25$.

Corpus (x10^3)	Accuracy perturbed (%)	Accuracy unperturbed (%)	Cluster prediction & KNN search time (off-chain) (ms)
100	91.6	100	0.58466
250	91	100	1.50747
500	91.2	100	2.53011
1000	91.2	100	4.27562

5.4 Evaluating Cost

Transaction cost is a consideration in scaling DLT systems. C-OOO is significantly more costly than E-OOF/FOF. Ingesting images costs, on average, 0.9M gas/image for C-OOO and, in comparison, only 0.2M gas/image for E-xOx variants. Similarly, when adding an image, a user would pay, on average, 19M gas/image for C-OOO but only 15M gas/image for E-xOx variants. Projecting the fingerprint embedding space onto a lower dimensional space using principal component analysis can further reduce these costs but does not alter the trend. The cost factor reinforces our design recommendation to use the DLT event log rather than in-contract storage for the key-value data.

5.5 DECORAIT applied to Dreambooth

Using the recommended E-FOF variant of the system, we demonstrate DECORAIT in a real-world scenario by specializing a Stable Diffusion model using the Dreambooth [Ruiz et al. 2022] method to synthesize renditions of a specific subject in new contexts.

Initially, a set of concept images is purchased from a popular stock photography website, which can be viewed on the left side of Fig.5. Unfortunately, their delivery is not accompanied by a C2PA manifest, therefore training consent cannot be immediately determined. The DECORAIT system is then queried to determine training consent across the set of concept images, by matching the images to their corresponding images within the registry. The assets within the registry are accompanied by C2PA manifests, which detail the author's choice of whether to allow GenAI training using that asset. The query to the DECORAIT registry resolves to several of our chosen images indicating that the creative has opted out of GenAI training. Fig.1 pictures the effect differing training data can have on the resulting model and the synthetic media it is able to generate, especially when a subset of the chosen concept images has been opted-out of GenAI training.

Once the model is trained using the opted-in images and following the Dreambooth method, we encode this provenance information within the manifest of both the resulting model and the generated synthetic image. An example provenance graph is pictured in Fig.4 and we follow the same structure in this example. The "ingredient" feature of the C2PA standard is leveraged in order to reference the resulting personalized model as the ingredient asset of the generated synthetic image. Within the personalized model's

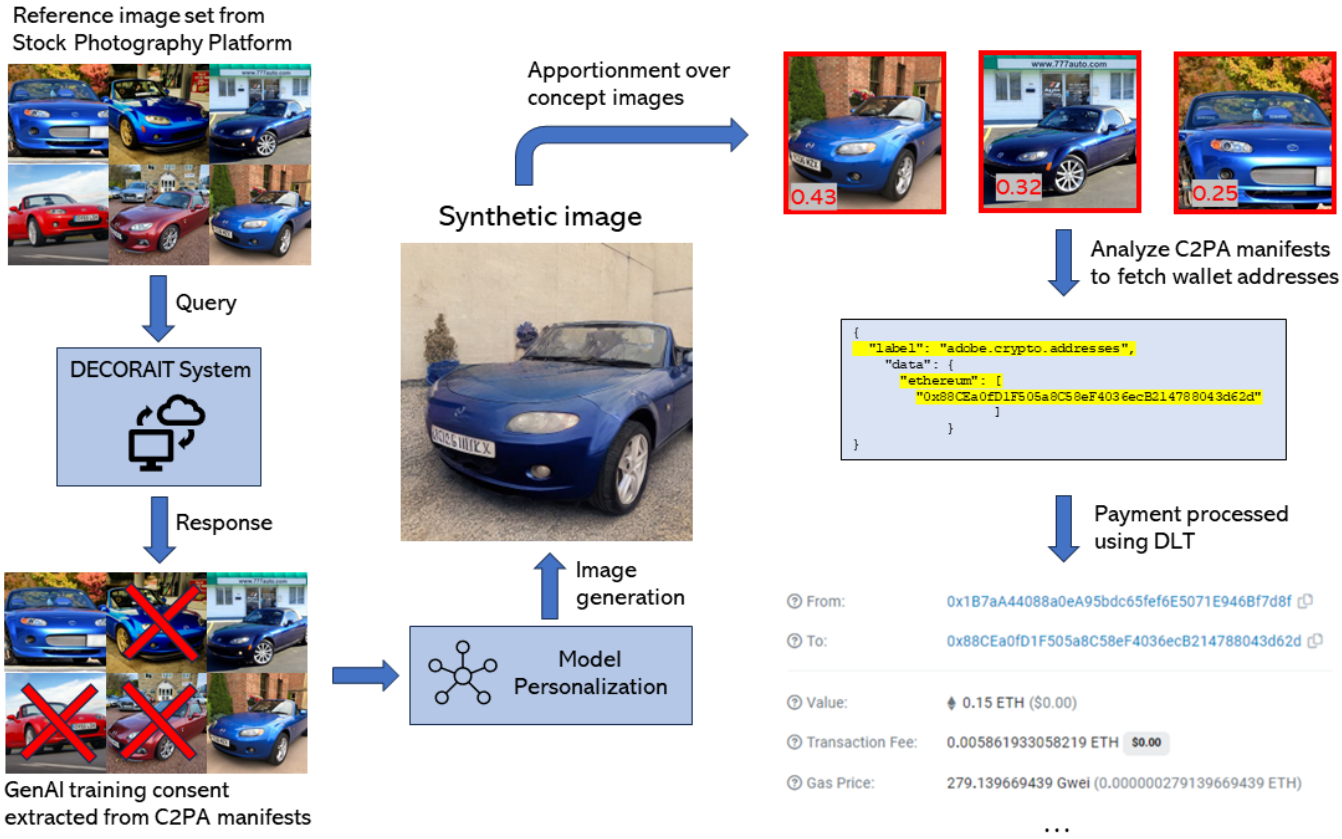

Figure 5: DECORAIT and Dreambooth pipeline including registry querying and model personalization flow. The Dreambooth model is specialized using the 3 opted-in images of a car and the proposed apportionment algorithm is applied across the image corpus. The red cross indicated images which have been opted-out according to the DECORAIT registry. The resulting apportionment conducted on the generated synthetic image from the experiment as described in Sec.5.5 is shown. The DLT wallet addresses of the three authors of the images are identified using the accompanying C2PA manifests. Payment is then conducted automatically, securely, and transparently using DLT, and one transaction's confirmation is pictured.

manifest, we encode as ingredients both the set of concept images it was trained on, as well as the base text-to-image Stable Diffusion model which was fine-tuned in order to create the personalized model. The base model may include within its manifest an archive detailing its entire training corpus of ingredient images.

Further, we apply the apportionment algorithm in order to reward the contributing creatives. The process starts from the C2PA manifest of the synthetic image, tracing the provenance graph in order to identify the personalized model which created it and ultimately its training images. Then, the apportionment accumulates prediction scores using the fingerprinter and second stage classifier model for each concept image the model was specialized on. The wallet addresses belonging to the creatives who authored the training images are identified by analyzing the C2PA manifests of those images. Lastly, payments are processed for each contributing creative, with currency sent directly to their wallet address through DLT, as pictured on the right side of Fig.5. The transaction confirmation is also pictured.

Thus, we have demonstrated an end-to-end pipeline which included ethically building a dataset of assets which have been opted-in for GenAI training, successfully avoiding copyright infringement, personalizing a generative diffusion model, as well as analyzing the resulting synthetic media and running our proposed apportioning algorithm in order to recognize and reward the contributing creatives, enabling near-instant processing of royalty-like payments using DLT.

6 CONCLUSION

We presented an end-to-end system through which content creators may assert their right to opt in or out of GenAI training, as well as receive reward for their contributions. We investigated the feasibility of a decentralized opt-in/out registry for GenAI using DLT, reaching recommendations that 1) event-log storage is appropriate; 2) on-chain shard prediction is appropriate for ingest but not for the query. We propose variant E-FOF as the most scalable solution, achieving 100% accuracy on non-augmented queries and

91.2% accuracy in the presence of augmentations, with query speed up to 4 ms for a corpus of 1M images.

DECORAIT employs the distributed ledger (DLT) as a trustless registry and source of truth. The bulk of the computationally expensive operations are conducted off-chain. We proposed a fingerprinting-based content similarity score for image attribution and credit apportionment over the attribution corpus in the case of synthetic media, with payments securely processed for the contributing creatives using DLT. The system leverages the C2PA standard to track content provenance, specify GenAI training consent and store the creator's DLT wallet addresses. We demonstrated the DECORAIT system as part of a Dreambooth GenAI model personalization pipeline, demonstrating our proposed method for recovering synthetic media provenance and apportioning credit. DECORAIT thus enables contributing creatives to receive recognition and reward when their content is used in GenAI training.

Future work could incorporate the DECORAIT registry within popular GenAI data loaders and ship the apportioning flow as a library in order to drive adoption. Most notably, future efforts should focus on investigating the socio-technical drivers and challenges our system may face when deployed in the wild. Further consideration is required for its development and implementation within a sustainable business model. Equally critical is the necessity for establishing comprehensive policies within the legal and regulatory space addressing digital rights and data sourcing for training GenAI models. These questions are likely to remain open for some time, however, ensuring the consensual use of digital assets and fair reward to contributors within the GenAI training pipeline is both a timely and urgent matter. We believe the proposed DECORAIT system is a promising first step towards a decentralized, end-to-end solution to the problem.

ACKNOWLEDGMENTS

DECORAIT was supported in part by DECaDE under EPSRC Grant EP/T022485/1.

REFERENCES

J. Aythora et al. 2020. Multi-stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing. In *Proc. Intl. Broadcasting Convention (IBC).*

K. Balan, S. Agarwal, S. Jenni, A. Parsons, A. Gilbert, and J. Collomosse. 2023. EKILA: Synthetic Media Provenance and Attribution for Generative Art. In *Proc. CVPR Workshop on Media Forensics.*

J. Benet. 2014. IPFS - Content Addressed, Versioned, P2P File System. arXiv:1407.3561 [cs.NI]

A. Bharati, D. Moreira, P. Flynn, A. de Rezende Rocha, K. Bowyer, and W. Scheirer. 2021. Transformation-Aware Embeddings for Image Provenance. *IEEE Trans. Info. Forensics and Sec.* 16 (2021), 2493–2507.

S. Bhujel and Y. Rahulamathavan. 2022. A Survey: Security, Transparency, and Scalability Issues of NFT's and Its Marketplaces. *J. Sensors. MDPI.* 22, 22 (2022).

A. Black, T. Bui, H. Jin, V. Swaminathan, and J. Collomosse. 2021. Deep Image Comparator: Learning To Visualize Editorial Change. In *Proc. CVPR Workshop on Media Forensics.* 972–980.

T. Bui, S. Agarwal, N. Yu, and J. Collomosse. 2023. RoSteALS: Robust Steganography using Autoencoder Latent Space. In *Proc. CVPR WS.*

T. Bui, D. Cooper, J. Collomosse, M. Bell, A. Green, J. Sheridan, J. Higgins, A. Das, J. Keller, and O. Thereaux. 2020. Tamper-proofing Video with Hierarchical Attention Autoencoder Hashing on Blockchain. *IEEE Trans. Multimedia (TMM)* 22, 11 (2020), 2858–2872.

T. Bui, D. Cooper, J. Collomosse, M. Bell, A. Green, J. Sheridan, J. Higgins, A. Das, J. Keller, O. Thereaux, and A. Brown. [n. d.]. ARCHANGEL: Tamper-proofing Video Archives using Temporal Content Hashes on the Blockchain. In *Proc. CVPR Workshop on Computer Vision, AI and Blockchain, year = 2019.*

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. 2023. Extracting Training Data from Diffusion Models. arXiv:2301.13188 [cs.CR]

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

Coalition for Content Provenance and Authenticity. 2021. *Technical Specification v1.3.* Technical Report. C2PA. https://c2pa.org/

J. Collomosse, T. Bui, A. Brown, J. Sheridan, A. Green, M. Bell, J. Fawcett, J. Higgins, and O. Thereaux. 2018. ARCHANGEL: Trusted Archives of Digital Public Documents. In *Proc. ACM Doc.Eng.*

P. Meenakshi Devi, M. Venkatesan, and K. Duraiswamy. 2019. A Fragile Watermarking scheme for Image Authentication with Tamper Localization Using Integer Wavelet transform. *J. Computer Science* 5, 11 (2019), 831–837.

P. Dhariwal and A. Nichol. 2021. Diffusion models beat GANs on image synthesis. *NeuIPS* 34 (2021), 8780–8794.

J. Fairfield. 2021. Tokenized: The Law of Non-Fungible Tokens and Unique Digital Property. *Indiana Law Journal* (2021).

H. Gilbert and H. Handschuh. 2003. Security Analysis of SHA-256 and Sisters. In *Proc. Selected Areas in Cryptography (SAC).*

J. D. Harris and B. Waggoner. 2019. Decentralized and Collaborative AI on Blockchain. In *IEEE Intl. Conf. on Blockchain.* IEEE.

D. Hendrycks and T. Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations* (2019).

J. Ho, A. Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *NIPS* 33 (2020), 6840–6851.

C. Holmes. 2018. Distributed ledger technologies for public good. *Tech. Rep., UK Gov. Office for Science* 1 (2018), 1–33.

N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J. Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. arXiv:2212.04488 [cs.CV]

V. L. Lemieux. 2016. *Blockchain Technology for Recordkeeping: Help or Hype?* Technical Report. U. British Columbia.

S. Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. http://www.bitcoin.org/bitcoin.pdf

A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder. 2016. *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.* U. Princeton.

E. Nguyen, T. Bui, V. Swaminathan, and J. Collomosse. 2021. OSCAR-Net: Object-centric Scene Graph Attention for Image Attribution. In *Proc. ICCV.*

OpenAI. [n. d.]. Introducing Chat-GPT. https://openai.com/blog/chatgpt. Accessed: 2023.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/ARXIV.2204.06125

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

L. Rosenthol, A. Parsons, E. Scouten, J. Aythora, B. MacCormack, P. England, M. Levallee, J. Dotan, et al. 2020. *Content Authenticity Initiative (CAI): Setting the Standard for Content Attribution.* Technical Report. Adobe Inc.

N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. (2022).

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. Kamyar Ghasemipour, B. Karagol Ayan, S. Mahdavi, R. Gontijo Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

C. Schuhmann, R. Vencu, R. Beaumont, . Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114 [cs.CV]

J. Shi, W. Xiong, Z. Lin, and H. Joon Jung. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. arXiv:2304.03411 [cs.CV]

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. 2022. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. arXiv:2212.03860 [cs.LG]

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. 2023. Understanding and Mitigating Copying in Diffusion Models. arXiv:2305.20086 [cs.LG]

Stablity.ai. [n. d.]. Stable Diffusion Public Release. https://stability.ai/blog/stable-diffusion-public-release. Accessed: 2023.

G. Tolias, R. Sicre, and H. Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879* (2015).

M. Walport. 2015. *Distributed Ledgers: Beyond Blockchain.* Technical Report. UK Government.

N. Yu, V. Skripniuk, D. Chen, L. Davis, and M. Fritz. 2021. Responsible Disclosure of Generative Models Using Scalable Fingerprinting. In *Proc. ICLR.*