MONTRAGE: Monitoring Training for Attribution of Generative Diffusion Models

Jonathan Brokman[*1,2], Omer Hofman[*1], Roman Vainshtein[1], Amit Giloni[1,3], Toshiya Shimizu[4], Inderjeet Singh[1], Oren Rachmil[1], Alon Zolfi[3], Asaf Shabtai[3], Yuki Unno[4], and Hisashi Kojima[4]

[1] Fujitsu Research of Europe
[2] Technion - Israel Institute of Technology
[3] Ben-Gurion University of the Negev
[4] Fujitsu Limited

Abstract. Diffusion models, which revolutionized image generation, are facing challenges related to intellectual property. These challenges arise when a generated image is influenced by copyrighted images from the training data, a plausible scenario in internet-collected data. Hence, pinpointing influential images from the training dataset, a task known as data attribution, becomes crucial for transparency of content origins. We introduce MONTRAGE, a pioneering data attribution method. Unlike existing approaches that analyze the model post-training, MONTRAGE integrates a novel technique to monitor generations throughout the training via internal model representations. It is tailored for customized diffusion models, where training dynamics access is a practical assumption. This approach, coupled with a new loss function, enhances performance while maintaining efficiency. The advantage of MONTRAGE is evaluated in two granularity-levels: Between-concepts and within-concept, outperforming current state-of-the-art methods for high accuracy. This substantiates MONTRAGE's insights on diffusion models and its contribution towards copyright solutions for AI digital-art.

Keywords: Data Attribution · Diffusion Models · Model Customization

1 Introduction

Since the beginning of the diffusion model revolution [6, 14], unprecedented capabilities in image generation have been developed [27, 29]. This technological advancement is already put to practical use in media and digital-art [16], however, such uses challenge traditional boundaries of intellectual property rights [7, 20,31].[1] [2] The main concern: Each image generated by these models is influenced by a subset of the training data, which might include copyrighted content. This raises legal questions about who owns these newly generated images.

* Equal contribution.
[1] New York Times: AI Image Generators and Copyright Issues
[2] Harvard Business Review: Generative AI and Intellectual Property Challenges

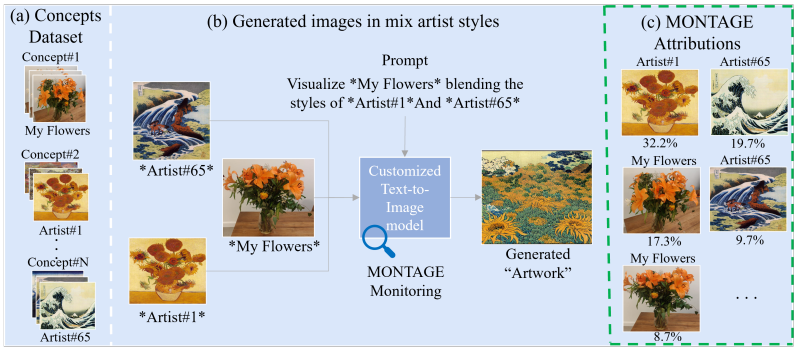

Fig. 1: MONTRAGE attributions. a) Users select images of objects and styles. b) Generate a new "artwork" using a customized diffusion model. c) MONTRAGE monitors the generation process to quantify the influence of each object and artist's style, providing insights into the generative process and addressing copyright concerns.

Diffusion model development can be split into two methodologies: Base model training and fine-tuning [33]. Base model training entails compiling extensive datasets from varied sources. Their large scale leads to control issues over copyrighted content, as seen, for instance, in the LAION and IMAGEN datasets [4, 29–31]. In contrast, diffusion model fine-tuning, used for model customization, involves using smaller and specific datasets as well as efficient fine-tuning methodologies to customize pre-trained base models for new capabilities [8,15,21,26,28]. This enables diffusion model adaption in low-resource settings. Consequently, customization became a popular tool among companies and private creators alike, increasing the risk of copyright infringement.[3] We define a 'concept' as an identifiable category or style within the customization dataset.

Consider a scenario where a fan collects a combined set of artworks from several artists, and employs customization to generate a new artwork. This raises the issue of attributing each artist's degree of contribution to this piece - underscoring the focus of our research: Determining the origins of personalized content generation to resolve copyright considerations (see example in Figure 1). Understanding how pieces of training data contribute to the model's output is at the core of these technological and legal challenges, a task known as *data attribution*.

Beyond generative AI, data attribution in the context of deep learning has a well-established history [19]. It usually entails the post-hoc analysis of a trained model, *i.e.* without access to the training process. Classical approaches employ loss gradients and Hessians to quantify how each training sample impacts the dynamics of pre-trained weights in their local environment and consequently the model's output [11, 18, 23]. An important branch of data attribution leverages access to the training process. This enables the precise tracing and aggregation of the training samples' influence on each training iteration [12, 24].

[3] Adobe Blog: FAIR Act to Protect Artists in the Age of AI

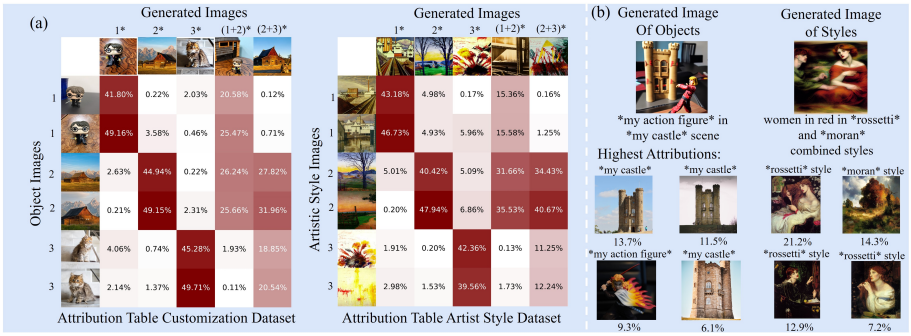

Fig. 2: Outcomes of MONTRAGE applied to object and artistic datasets featuring single and mixed concepts. (a) Presents small-scale attribution tables obtained by monitoring the fine-tuning process of a customization model. (b) Shows the attribution model output, given a generated image, the attribution model predicts the customized images that contributed most significantly to its creation.

Several recent works expanded the data attribution domain to analyze diffusion models, focusing on scenarios without direct access to the training process [22, 32, 34]. On the one hand, avoiding dependence on training access is a practical approach for base models due to the expensive and time-consuming training. However, for fine-tuning cases, which require fewer resources, access to training becomes practical. It holds valuable information, which is advantageous, especially for legal applications where maximum accuracy is required.

We present MONTRAGE (MONitoring TRaining for Attribution of image GEnerative models), a novel two step data attribution method for customized diffusion models. By Leveraging direct training access, MONTRAGE fills a critical gap in the literature. During the first step, MONTRAGE monitors the internal representations of the model during the training process, given by the attention layers' activations, and accounts for their training dynamics by aggregating their changes into an *attribution table*. This offers training insights while maintaining efficiency by avoiding the full generation pipeline overhead. For the second step, a separate attribution model is trained on the attribution table via a novel loss function to predict attributions of unseen (un-monitored) generations. An example is provided in Figure 2.

MONTRAGE is empirically validated on two granularity-levels: Between-concept (attributing the correct semantic concept) and within-concept. Evaluations include important use-cases of this domain, customization and artistic styles [13, 21]. The results show that MONTRAGE outperforms state-of-the-art methods [32,34] in obtaining high accuracy more frequently on two datasets. Additionally, it succeeds in attributing mixed-concept images, which contain several concepts. This demonstrates the potential of MONTRAGE for legal, ethical, and scientific applications. The main contributions of this paper are as follows:

- Our approach is, to the best of our knowledge, the first to monitor the training dynamics of diffusion models for data attribution purposes.
- MONTRAGE leverages inner-model representations from cross-attention layers to enhance data attribution efficiency.
- MONTRAGE employs a specialized loss function for enhanced granularity, providing nuanced insights into the model's training process.

2 Background and Related Work

2.1 Diffusion Models

Diffusion models are generative models that produce high-quality samples in various domains. They operate by an iterative generation process of noise reduction based on a pre-set noise schedule. Initially, we sample $x_T \sim N(0, I)$. Each iteration t involves partially denoising x_t via a neural network $f(x_t, t; \theta)$ (θ are the tuneable weights), subsequently progressing to x_{t-1}. This sequence produces $x_0 \in X$, representing the final output, where X denotes the data domain. In our setting x_0 is an image. This generation process is known as *reverse diffusion*, where during training, f is optimized to reverse a Markovian *forward diffusion* process defined as $x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \hat{n}$ where $\hat{n} \sim \mathcal{N}(0, I)$, $\beta_t \in \mathbb{R}^+ \, \forall t$.

It can be shown that for sufficiently small β_t's, the learned reverse process is Markovian as well and can be represented as:

$$p_\theta(\mathbf{x}_{0:T-1}) = p_\theta(x_T) \prod_{t=T}^{1} p_\theta(x_{t-1}|x_t), \tag{1}$$

where p_θ is the Probability Density Function (PDF) induced by the model $f(\cdot; \theta)$.

For text-to-image models, noisy samples are coupled with prompts, denoted $z_t = (x_t, p)$. For convenience, we use the same model notation, *i.e.* $f(z_t, t; \theta)$. Let $D = \{z^1, \ldots, z^{|D|}\}$ be an ordered clean-image training set of size $|D|$, where each $z^i := (x_0^i, p^i)$. Let $L_{t,\hat{n}}(z_t, \theta)$ be the loss on sample z_t considering randomly drawn $t, \hat{n}$. Denote $L(z, \theta) := \mathbb{E}_{\hat{n}} \mathbb{E}_t L_{t,\hat{n}}(z_t, \theta)$. Denote $\theta^*(D)$ as:

$$\theta^*(D) := \arg\min_\theta \mathbb{E}_{z^i \in D} L(z, \theta). \tag{2}$$

To learn the reverse diffusion, a common loss function at iteration *iter* is

$$L_{t,\hat{n}}^{iter}(z_0, \theta) := \left\| f(z_{t^{iter}}, t^{iter}; \theta) - z_0 \right\|^2. \tag{3}$$

where we sample n^{iter}, t^{iter} once to construct $z_{t^{iter}} := (x_{t^{iter}}, p)$. Note that for each *iter*, only a single inverse step is learned, hence it does not entail the full reverse diffusion. After enough iterations, the model is exposed to a wide range of noise levels and learns to reverse the diffusion process from any point within the diffusion sequence. Denote the output of a trained text-to-image model as

$$x^{gen} := f(z_T, T; \theta^*(D)), \tag{4}$$

where z_T is sampled as $z_T = (\hat{n}, p^{gen})$ with p^{gen} the generation conditioning prompt. By Markovianity of (1), the generation of x^{gen} requires a multiple-step process pre-defined by the scheduled β_t, thus a forward pass may become more time-intensive than a training iteration, especially in fine-tuning methods.[4]

The input prompt p^{gen} often conditions the generation via cross-attention layers, a technique originating from [27]. The prompt is encoded into nonlinear token embeddings c, which are integrated into the generation via cross-attention layers: Let $W_K, W_V \in \theta$ tunable matrices. Projections into $K = W_K c$ and $V = W_V c$ are employed, and combined with query matrices Q, which represent visual image features , bridging text and image modalities as:

$$\text{Cross-Attention} := \text{softmax}\left(\frac{QK^T}{\sqrt{m}}\right) V. \tag{5}$$

This expresses a weighted average of prompt information (V). Several recent works have utilized the analysis of (5) for understanding and control of the generation process [9, 17, 21]. We are the first to analyze it for data attribution.

2.2 Data Attribution and Evaluation Metrics

Data Attribution entails the identification of the influential training data that affects the trained model's predictions [19]. In the context of generative models, it involves mapping the generated outputs to the training examples that facilitate their creation, an important step for understanding model behavior [32].

Evaluating Data Attribution in Diffusion Models is yet an open research topic [22]. Different perspectives exist on defining "correct attribution" in diffusion models, where some adapt pre-existing criteria from the literature, and others offer new approaches tailored for the generative case. In [32], the authors test the ability of attribution methods to select training images that belong to the same concept as the images produced by the generative model. They apply image retrieval criteria, such as recall@K for this purpose.

Recently, Linear Datamodeling Score (LDS) [23], a tailored metric for data attribution evaluation, has been proposed for classical data attribution. Following this, LDS was adapted for diffusion models in [10,34]. LDS defines attribution reliability as the correlation between attribution scores and the model's predictions, utilizing various subsets of training images per generation.

In addition to correctness, efficiency is another important factor for a data attribution method, encompassing computational demands such as calculation time during training and inference. In our work, we aim for an efficient method while preserving attribution accuracy, essential for practical use.

2.3 Related Work

While data attribution has been extensively researched for discriminative models [12, 18, 19, 23, 24], only a few studies have been performed for generative

[4] Sometimes Markovianity is not admitted, nonetheless, a multi-step gradual process is still the common practice today

models, namely diffusion models. In D-TRAK [34], and Datainf [22], classical data attribution techniques that utilize the loss gradient and Hessian have been expanded and improved in terms of accuracy and efficiency to attribute diffusion models. Conversely, GenDataAttribution [32] presents a new methodology, leveraging the unique characteristics of diffusion models. They employ thousands of single-image customized models [21] to create a dataset of generated images, ensuring that a single known training image influences the output. These generated images provide ground-truth data, which GenDataAttribution leverages for contrastive learning of an attribution embedding space.

Among the diverse strategies to obtain attributions, a notable approach involves leveraging the training process to trace how each training instance influences model predictions throughout different training iterations [12, 24]. However, for diffusion models, the existing attribution methods avoid accessing the training process. This aligns with the practical constraints of base model development, where training is notably resource-intensive, rendering it inaccessible. Nevertheless, in fine-tuning scenarios, training access is feasible and provides crucial insights into how training images shape the generated outputs. We hypothesize that leveraging this access has the potential for new capabilities.

There is a related theoretical research branch in training dynamics. This area typically involves nonlinear analysis of model parameters' evolution and often links it to model performance [2, 3, 5, 35]. We use these dynamics to quantify data samples' contributions to performance.

We present a first-of-its-kind integration of two methodologies: Exploring training access for data attribution and leveraging diffusion model characteristics. We monitor internal representations of diffusion models for changes during training and aggregate this information for data attribution, thereby making a new contribution to the field.

3 The Method

MONTRAGE leverages the accessibility of the customization process in diffusion models by applying two steps as illustrated in Figure 3. First, data attribution values are collected throughout the customization process into an attribution table. Sequentially, a separate model is trained on the attribution table data via a specialized loss function that captures the fine granularity of the attributions.

3.1 Data Attribution Table Construction

Prior to fine-tuning on dataset D, a predefined set of generation cases is reserved for monitoring their evolution throughout the process. In our setting, each generation case consists of a noise-prompt input pair $z_T^j = (\hat{n}^j, p^j)$. Let $\mathbf{P}^{\text{monitor}}$ be the set of monitoring prompts, *i.e.* $p_j \in \mathbf{P}^{\text{monitor}}$. Note that the corresponding generation $x^{gen,j}$ (see Equation (4)) changes throughout the training. We then fine-tune a customized diffusion model and monitor these predefined cases. On every iteration, each generation case is encoded into a vector (tensor)

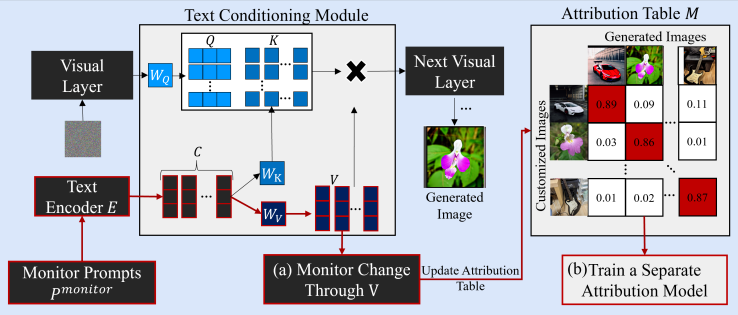

Fig. 3: MONTRAGE training pipeline. (a) First, the customization process of a diffusion model is monitored, with changes aggregated into an attribution table. (b) Next, a separate model is trained using this table.

representation, $V^{monitor,j}$ and fed into the model. The updates in $V^{monitor,j}$ reflect changes associated with the incoming batch of customization training samples. These changes are recorded in $\mathbf{M}$, a data attribution table, organized with columns representing generations and rows representing customization samples. In this way, the elements of $\mathbf{M}$ are cumulatively updated by tracking changes in $V^{monitor,j}$ over the course of training, as follows:

$$\mathbf{M}^{iter+1}[i,j] = \mathbf{M}^{iter}[i,j] + \Delta^{iter} V^{monitor,j}, \quad \forall i \in \text{Batch}^{iter}, \tag{6}$$

where $\mathbf{M}^{iter}, \text{Batch}^{iter}, \Delta^{iter}$ are $\mathbf{M}$, the batch, and the change in $V^{monitor,j}$ at iteration $iter$ respectively. We choose $\Delta^{iter} V := \|V^{iter+1} - V^{iter}\|_1$. Batches of size 1 are most natural here [24]; varied batch size experiments are in the Supp.

From M to Attributions. Reminder: $f(\cdot; \theta^*(D))$ denotes the diffusion model, trained on data D. Let us define $\tau : X \to \mathbb{R}^{|D|}$, as a data attribution function for $f(\cdot; \theta^*(D))$ of the form $\tau(x^{gen}; D)$. The i^{th} entry, denoted $\tau(x^{gen}; D)_i$, assigns a real-valued score to the training sample $z^i \in D$ indicating its importance to the x^{gen}. After fine-tuning is completed, the j_{th} column of $\mathbf{M}$, denoted $M[:, j]$, expresses the unnormalized data attribution for $x^{gen,j}$, and we define

$$\tau_{\text{Table}}(x^{gen,j}; D) := \frac{1}{\sum_i M[i,j]} M[:, j]. \tag{7}$$

Note that $\tau_{\text{Table}}(x^{gen,j}; D)$ is a probability vector.

Efficient Generation Monitoring Through the Choice of $V^{monitor}$. There is an inherent computational challenge of monitoring generations. While a training iteration samples a single step t to compute the sample loss of Equation (3), a full generation entails the full T steps reverse diffusion process (see Equation (1)). This makes the monitoring of full generations during fine-tuning time-consuming, especially in the efficient customization case.

To avoid the generation computational overhead, we turn to prior research, which underscored the analytical value as well as the efficiency of the cross-

attention layer in generation analysis [9, 17, 21]. We propose to utilize this in data attribution, namely MONTRAGE monitors tensor V of Equation (5).

To further clarify the behavior of tensor V, consider the same monitored prompt but different noise seeds. While V, being prompt-based, is unaffected by the seed during a forward-pass, its monitored training dynamics are. Hence the aggregated attributions reflect the interaction between internal text representations and the generated images.

The monitoring of V holds several advantages: First, V does not require full generations, saving ample computation time; Second, V is an informative representation of the generation, which encapsulates the text conditioning; Third, The objective of concept customization is to generate diverse outputs that maintain semantic consistency with the customization concept. Monitoring V encapsulates this assumption - as the resulting attributions become consistent for images generated from the same concept; Fourth, V is readily scaled (inner parts of V have the same scale), making it fit for monitoring, see supplementary for explanation. A detailed pseudo-code of this approach is provided in Alg. 1.

Algorithm 1 Monitoring customization for attribution table construction.

$\mathbf{M} \leftarrow$ attribution table,
  initialized as matrix of zeros
$\mathbf{D} \leftarrow$ customization image-prompt pairs
$\mathbf{P}^{\text{monitor}} \leftarrow$ monitoring prompts
$\mathbf{G} \leftarrow$ Generator
$\mathbf{E} \leftarrow$ Text Encoder
Let (x^i, p^i) be the i^{th} image-prompt
  pair in $\mathbf{D}$
for each epoch **do**
  for each $(x^i, p^i) \in \mathbf{D}$ **do**
    $\mathbf{G}, \mathbf{E} \leftarrow$ forward +
      backward (optimization step)
      on $[x^i, p^i]$
    for each $p^j \in \mathbf{P}^{\text{monitor}}$ **do**
      $c^j \leftarrow \mathbf{E}(p^j)$
      $V^j \leftarrow \mathbf{W}^v c^j$ ▷ see (5)
      if not first iteration **then**
        $\mathbf{M}[i, j] += \|V^j - \tilde{V}^j\|$
      $\tilde{V}^j \leftarrow V^j$

3.2 Data Attribution Model

To enhance the applicability of the generated attribution table M, our methodology extends its attribution utility to unseen image generations, i.e., generations outside $\mathbf{P}^{monitor}$. Following [32], we employ a Distance Metric Learning (DML) approach to learn an image embedding space in which the similarity between generated and training (customization) images corresponds to the attribution.

Our methodology involves training a Siamese network, tailored to align with our attribution table's unique characteristics, to distinguish between conceptually similar (positive) and distinct (negative) pairs of customized and generated images. This network is trained using $\mathbf{M}$ as training data. During training (shown

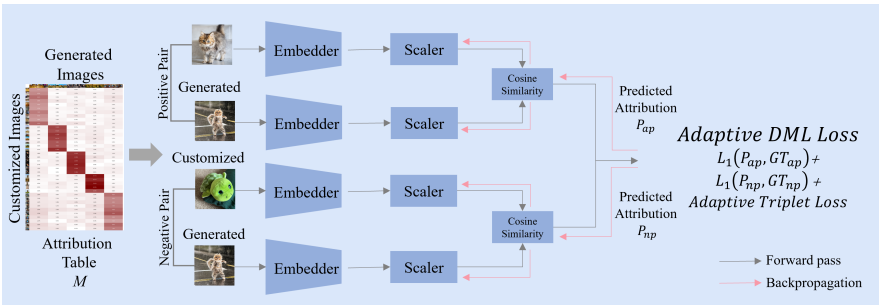

Fig. 4: Our DML model training. Trained on the attribution table **M**. it gains accurate predictions and extends attribution capabilities to unseen generated images.

in Figure 4), each image pair goes through a two-stage transformation involving initial feature extraction by a pre-trained embedder which feeds a *Scaler* layer, designed to adjust the scale and shift of the embedding vector, making it fit for this task, and outputs the final embedding. The attributions are then obtained as shifted cosine similarity[5] between vector embeddings of generation and customization pairs. These predicted attributions are used to measure the loss against the ground truth attribution scores obtained from **M**.

The Adaptive DML Loss Function. The Siamese network employs a three-component adaptive loss function. Let (P_{ap}, P_{np}), (GT_{ap}, GT_{np}) be the positive and negative pairs of predictions and ground truth (the attribution values obtained from **M**) respectively, denote P_{ap_i}, P_{np_i} as the i^th entries of P_{ap}, P_{np} respectively. Let B be the No. of pairs in a batch. To account for the distance between concepts, we introduce the new Adaptive Triplet Loss:

$$\text{Adaptive Triplet Loss} := \frac{1}{B}\sum_{i=1}^{B}\max(P_{np_i} + m_i - P_{ap_i}, 0), \tag{8}$$

where m_i is the margin derived from the ground truth pairs $m_i = GT_{ap_i} - GT_{np_i}$. This loss penalizes the model based on the margin between each positive and negative prediction. It is incorporated into an adaptive DML loss as follows

$$\text{Adaptive DML Loss} := L_1(P_{ap}, GT_{ap}) + L_1(P_{np}, GT_{np}) + \text{Adaptive Triplet Loss}, \tag{9}$$

where the L_1 losses account for between-concepts, and the Adaptive Triplet loss accounts for between and within-concept, through the margin m_i.

Conventional DML models measure distances between concepts (classes), while our model also predicts distances within a concept using **M** values and the Adaptive DML Loss. Thus the hierarchicy of attributions is learned, enhancing between-concept understanding and concept attribution granularity.

[5] The shifted cosine similarity function adjusts the standard cosine similarity range from $[-1, 1]$ to $[0, 1]$, aligning with the ground truth values for comparison.

4 Evaluation

4.1 Datasets

The evaluation of our proposed method utilizes two datasets, which cover two main applications of MONTRAGE: Generative model customization and artistic styles. **CustomConcepts101** [21] is a benchmark dataset designed for evaluating model customization techniques. It is also applicable for assessing data attribution methods. The dataset encompasses 101 unique concepts, each represented by a collection of 3 to 15 images, and corresponding textual prompts. This dataset's variety offers a solid basis for evaluating our method's adaptability and accuracy across diverse visual concepts. The Artchive dataset [13], offers an extensive collection of paintings by famous artists, including Van Gogh, Monet, and Gaudi. Incorporating the Artchive dataset enables a comprehensive evaluation of our method's ability to attribute complex artistic styles. Our evaluation is performed over the entire CustomConcepts101 dataset (101 concepts) and 50 artistic style concepts that were randomly selected from the Artchive dataset.

4.2 Experimental Settings

Our experiments setting is based on Custom Diffusion [21], a state-of-the-art text-to-image fine-tuning technique known for its efficiency in storage and running time. For the implementation we used their Diffusers library (Hugging Face) version. The pre-trained stable-diffusion-v1-4 [27] is used for the base diffusion model in all customizations. The monitored customization fine-tuning mostly follows the code's original hyper-parameters (details in supplementary).

The monitoring itself and the construction of the attribution table $\mathbf{M}$ does not require any hyper-parameter, other than the choice of $\mathbf{P}^{monitor}$. To this end, we employed 10 prompt templates, uniform across concepts (see supplementary). As in [24], MONTRAGE naturally works with batch size 1, used here. The supplementary (Sec. B) shows bigger batches still give concept-aware attributions.

For the attribution model we used a pre-trained Clip [25] as the Embedder, since it has been found most suitable for this task in [32]. The Scaler component has been implemented as a PyTorch ScalingLayer, which applies trainable normalization and shifting parameters to adjust the attribution space. Our implementation code is available in the following link. [6].

Methods for Comparison. To evaluate MONTRAGE, we benchmarked it against two leading data attribution methods designed for diffusion models: GenDataAttribution [32] and D-TRAK [34], implementated while adhering to the specifications in their respective publications, using available code and models.

4.3 Test Cases and Metrics

We evaluate each of the two steps in MONTRAGE: First, the attribution table $\mathbf{M}$ is evaluated for its reliability. Second, the attribution model is tested for

[6] `https://github.com/omerHofBGU/MONTRAGE`

its ability to generalize $\mathbf{M}$ for attributing unseen images. Runtime and storage analysis for both steps can be found in the supplementary, Sec. C. We adopt two performance evaluation strategies, one for each step:

Attribution Table Reliability Assessment (within-concept). Evaluating the reliability of $\mathbf{M}$ is challenging due to the lack of ground truth values, since attributions are subjective interpretations derived from the data and the model. Although we possess ground-truth concepts, our goal is to assess the within-concept attributions in $\mathbf{M}$, *i.e.* sub-concept granularity evaluation. Thus we capitalize on single-concept customizations, where there is no risk of concept-aware attributions - a scenario where the attributions are ranked correctly due to correct concept assignment rather than within-concept understanding. The recently proposed LDS metric is a good fit for this evaluation strategy [10, 23, 34] (see Section 2.2). LDS evaluates the correlation between summed attribution scores of a training data subset and the performance of a model re-trained on this subset. LDS focuses on internal consistency without using ground truth labels.

LDS Formal Definition. Let $S \subset D$ a subset of the training data, and its corresponding subset-trained model $f(\cdot; \theta^*(S))$. Let $\tau(x^{gen}; S)$, $\theta^*(S)$, defined similarly to $\tau(x^{gen}; D)$, $\theta^*(D)$ respectively, see Sec. 2 for notations. Let $g_\tau(x^{gen}, S; D) := \tau(x^{gen}; D)^T 1_S$, where 1_S is the indicator vector for the subset S. The LDS for a data attribution method τ regarding a generated x^{gen} produced by a model trained on D is defined as:

$$LDS(\tau, D, x^{gen}) := \rho(\{L((x^{gen}, p^{gen}), \theta^*(S_m))\}_{m=1}^{N^S}\}, \{g_\tau(x^{gen}, S_m; D)\}_{m=1}^{N^S}). \tag{10}$$

ρ denotes Spearman's rank correlation [1], and $\{S_m\}_{m=1}^{N^S}$ are N^S subsets of D.

LDS Challenges in the Customization Setting. In settings involving single-concept datasets, with some concepts having as few as three images (such as CustomConcept101), applying LDS necessitates tailored modifications. To ensure robust subset evaluation, we set $N^S = 2$ across all concepts, which guarantees multiple images per subset but may limit attribution diversity in concepts with larger image counts. Thus, we replace random subset sampling with a customized sampling strategy to promotes diverse attributions - see details in the supplementary, Sec. A.

Within-Concept Criterions. We define the LDS *Accuracy* for $N^S = 2$ as:

$$\text{Accuracy}(\tau, D^c) := \mathbb{E}_{x^{gen}} \mathbb{1}\left[LDS\left(\tau, D^c, x^{gen}\right) > 0\right], \tag{11}$$

where x^{gen} is generated via $\mathbf{P}^{monitor}$ (from Algo. 1), using multiple generations per-prompt. See supplementary, end of Sec. B.1 for explanation of this criterion.

For $N^S = 2$, it is easy to show that a baseline $\tau_{rand}(x^{gen}; D^c)$, which ignores x^{gen} and assigns random uniform (normalized) attributions, results with:

$$\text{Accuracy}(\tau_{rand}, D^c) = 0.5 \forall D^c \quad \text{(Baseline)}, \tag{12}$$

see derivation in the supplementary. Therefore in Fig. 5 we present the results of testing MONTRAGE against competing methods for out-performing

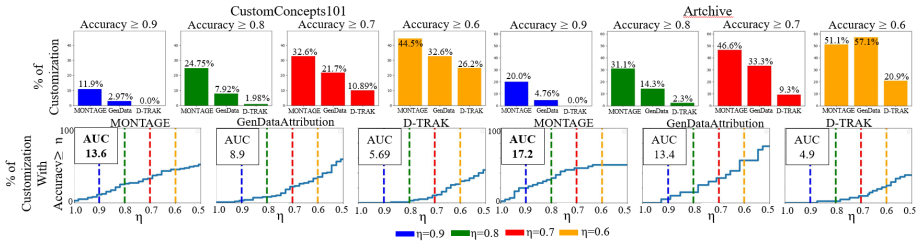

Fig. 5: Within-concept evaluation of our attribution table M. Top row: Comparing our method against competitors, using the LDS Accuracy criterion (11), for $\eta = 0.9, 0.8, 0.7, 0.6$. Bottom row: Evaluating Accuracy by sweeping values of η from 1.0 (perfect) 0.5 (baseline accuracy) (12). MONTRAGE has a clear advantage in higher values of η, but not in low values. To account for the whole $[0.5, 1]$ range, we employ the Area Under the Curve (AUC), where MONTRAGE outperforms the competitors.

the 0.5 random baseline, *i.e.* we set a threshold $\eta > 0.5$. The test considers $\{D^c\}_{c\in\text{customizations}}$ and empirically evaluates the percentage of customized models $f(\cdot;\theta^*(D^c))$, for which τ admits the Accuracy of (11), *i.e.*

$$\text{Evaluate } \tau \text{ given } \eta \text{ as } 100 \cdot \frac{\sum_{c\in\text{customizations}} \text{Accuracy}(\tau, D^c) > \eta}{\text{No. of Customizations}}. \tag{13}$$

Within-concept Results. The single-concept customizations provide a customized model for each concept. Each customized model generates 500 images, resulting with 500 × No. of customizations generated images for this evaluation (5500 in CustomConcept101 and 2500 in Artchive). Figure 5 shows the within-concept evaluation results of MONTRAGE, GenDataAttribution and D-TRAK within a single concept via the Accuracy criterion of Eq. (11). The results show a clear advantage of MONTRAGE in the high accuracy range (high values of η).

Attribution Model Generalization Assessment (between-concepts): To estimate the capability of our attribution model in generalizing **M**, we focus on its predicted attribution scores for unseen generated images, namely generations obtained via prompts $p \notin \mathbf{P}^{monitor}$. To this end, we adopt the image retrieval evaluation scheme used in [32], which tests attributions at the semantic concept level. Image concepts are used as ground truth labels and the attribution model is evaluated for its ability to assign high attribution scores to input images that belong to the same concept as the generated one. The metrics recall@K and precision@K were used as commonly applied in the DML setting. Additionally, Spearman's rank correlation is used to compare the ordering of the learned attributions with the original attributions obtained from **M**.

Between-concept Experimental Procedure. We analyzed **M** constructed for multi-concept model customizations (five and ten concepts), where the dataset was divided accordingly (*e.g.* in the CustomConcepts101 dataset, which has 101 concepts, we monitored 20 five-concept customizations and 10 ten-concept customizations). For each table, three DML attribution models were trained with varying seeds resulting in hundreds of attribution models (additional details can

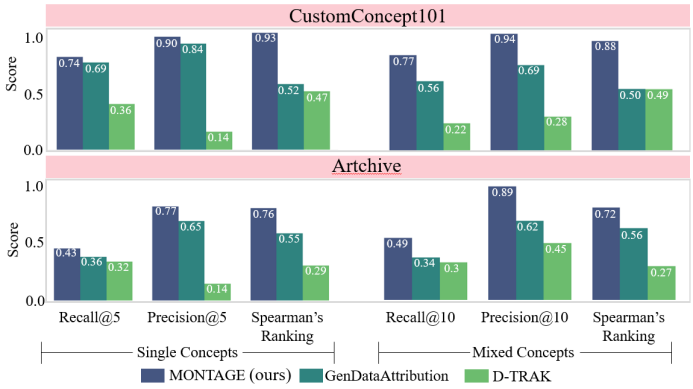

Fig. 6: Between-concepts evaluation. MONTRAGE surpasses existing methods across two datasets and various metrics, especially in mixed concept experiments.

be found in the supplementary material). We experiment with un-mixed images, where each generated image contains one concept, and mixed-concept images, where each generated images contain two concepts from the five or ten customization concepts. For recall@K and Precision@K of the unseen generated images, we set K=5 in the un-mixed concepts and K=10 in the mixed-concept (a larger K is employed since their attribution spans across more of images).

Between-concepts Results. Figure 6 presents the between-concepts evaluation results for MONTRAGE, GenDataAttribution, and D-TRAK on the CustomConcept101 and Artchive dataset. The plots show retrieval metrics for both un-mixed and mixed-concepts experiments, including recall, precision, and Spearman's rank correlation. These results represent the average performance across different attribution model's seeds. In the figure, we can see that MONTRAGE consistently outperformed GenDataAttribution and D-TRAK across all metrics, achieving the highest scores for all metrics in un-mixed and mixed-concept experiments. In particular, MONTRAGE's performance was notably superior in the mixed-concept experiments compared to the un-mixed concepts experiments. This larger performance between MONTRAGE and other methods underscores MONTRAGE's enhanced capability in attributing more complex (mixed-concept) generations. Note that the performance on the Artchive dataset was lower for all attribution methods compared to the CustomConcept101 dataset. This underscores the significant challenge in understanding generations of artistic styles.

Qualitative Base Model Analysis. The MONTRAGE attribution model satisfies an embedding that can be used to attribute any image. We can use it to compare base-model generated images with the base-model training dataset, effectively estimating attributions of the base model. Figure 7 demonstrates this. We used a subset of 100K images taken from the LAION dataset [30] - used for the training of the base model. We applied one of our attribution models for the base-model attribution. Surprisingly, MONTRAGE, designed for customized model attribution, also shows potential for base model attribution.

Fig. 7: Base model qualitative results. MONTRAGE succeeds in attributing base diffusion model training images as well. A quantitative complement to these findings is presented in the supplementary materials.

5 Limitations

A notable limitation of MONTRAGE is its dependence on having access to the training process. This facilitates MONTRAGE's frequent achievement of high accuracy, as evidenced in the within-concept evaluations, and may be perceived as an "unfair" advantage. However, given the nature of fine-tuning, where training access is anticipated, copyright issues may arise - for which it is imperative to utilize all accessible resources. This includes training dynamics access to enable maximized data attribution accuracy - justifying this "unfair" advantage. Additionally, MONTRAGE was implemented for customized models [21], and other use-cases remain to be tested.

6 Conclusion and Future Work

In this work, we introduced MONTRAGE, a novel data attribution method tailored for customized diffusion models. Our approach uniquely monitors the model's internal representations during training and leverages this data to construct an attribution model. This technique not only elucidates how training data influences image generation but does so efficiently.

Our method was evaluated on datasets focused on customization and artistic style, critical areas for data attribution. The results demonstrate MONTRAGE's capability to provide granular insights at both within-concept and between-concepts levels. Such granularity underscores the potential of MONTRAGE to significantly impact legal and ethical considerations in image generation.

While this study concentrated on specific prompt-related representation monitoring, exploring additional monitoring techniques could yield further advancements. We also recommend in future research to utilize training access during fine-tuning (through MONTRAGE or other) alongside a complementary technique for base models, ensuring a comprehensive data attribution framework.

References

1. The proof and measurement of association between two things. (1961)
2. Achille, A., Rovere, M., Soatto, S.: Critical learning periods in deep neural networks. arXiv preprint arXiv:1711.08856 (2017)
3. Brokman, J., Betser, R., Turjeman, R., Berkov, T., Cohen, I., Gilboa, G.: Enhancing neural training via a correlated dynamics model. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=c9xsaASm9L
4. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
5. Chandramoorthy, N., Loukas, A., Gatmiry, K., Jegelka, S.: On the generalization of learning algorithms that do not converge. Advances in Neural Information Processing Systems **35**, 34241–34257 (2022)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Fan, M., Chen, C., Wang, C., Huang, J.: On the trustworthiness landscape of state-of-the-art generative models: A comprehensive survey. arXiv preprint arXiv:2307.16680 (2023)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
9. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
10. Georgiev, K., Vendrow, J., Salman, H., Park, S.M., Madry, A.: The journey, not the destination: How data guides diffusion models. arXiv preprint arXiv:2312.06205 (2023)
11. Guo, H., Rajani, N., Hase, P., Bansal, M., Xiong, C.: Fastif: Scalable influence functions for efficient model interpretation and debugging. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10333–10350 (2021)
12. Hara, S., Nitanda, A., Maehara, T.: Data cleansing for models trained with sgd. Advances in Neural Information Processing Systems **32** (2019)
13. Harden, M.: The artchive. https://www.artchive.com/
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., Gebru, T.: Ai art and its impact on artists. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 363–374 (2023)
17. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
18. Khanna, R., Kim, B., Ghosh, J., Koyejo, S.: Interpreting black box predictions using fisher kernels. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 3382–3390. PMLR (2019)

19. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
20. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
21. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
22. Kwon, Y., Wu, E., Wu, K., Zou, J.: Datainf: Efficiently estimating data influence in loRA-tuned LLMs and diffusion models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=9m02ib92Wz
23. Park, S.M., Georgiev, K., Ilyas, A., Leclerc, G., Mądry, A.: Trak: attributing model behavior at scale. In: Proceedings of the 40th International Conference on Machine Learning. pp. 27074–27113 (2023)
24. Pruthi, G., Liu, F., Kale, S., Sundararajan, M.: Estimating training data influence by tracing gradient descent. Advances in Neural Information Processing Systems **33**, 19920–19930 (2020)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
26. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on Graphics (TOG) **42**(1), 1–13 (2022)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
30. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
31. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6048–6058 (2023)
32. Wang, S.Y., Efros, A.A., Zhu, J.Y., Zhang, R.: Evaluating data attribution for text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7192–7203 (2023)
33. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)

34. Zheng, X., Pang, T., Du, C., Jiang, J., Lin, M.: Intriguing properties of data attribution on diffusion models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=vKViCoKGcB
35. Zilly, J., Achille, A., Censi, A., Frazzoli, E.: On plasticity, invariance, and mutually frozen weights in sequential task learning. Advances in Neural Information Processing Systems **34**, 12386–12399 (2021)