

A Data-Interventional Framework for Auditing Privacy and Fairness in Generative Medical Imaging

Mischa **Dombrowski**¹ , Bernhard **Kainz**^{1,2} 

¹ Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, DE

² Imperial College London, London, UK

Abstract

Diffusion-based synthetic data generation offers a promising route for sharing medical imaging data without releasing sensitive patient records. However, generative models face a fundamental tension between privacy and fairness: they may memorize rare training samples, leading to privacy risks, or fail to reproduce underrepresented features, resulting in unfair synthetic distributions. While prior work has largely focused on either memorization or fairness in isolation, their interaction remains insufficiently understood. In this work, we introduce a data-interventional framework to systematically analyze privacy and fairness in diffusion models. We discuss synthetic anatomical fingerprints (SAFs), rare and manually injected image features, as controlled probes to study whether models generalize sensitive attributes across identities, memorize training samples, or suppress rare signals entirely. Across multiple conditioning modalities, we observe a consistent behavior: models either forget these fingerprints or memorize the entire image in which they appear, but do not generalize them to novel images. To support large-scale auditing where explicit sample extraction is infeasible, we further introduce the indicator metric t' , which estimates a model's susceptibility to memorization by exploiting the internal structure of the diffusion process. By comparing conditioning signals of varying surprisal, we reveal a clear relationship between conditioning rarity and memorization behavior. Highly surprising conditioning signals act as retrieval keys that amplify memorization, whereas low-surprisal conditioning signals systematically suppress rare features, even when these appear repeatedly in the training data. Our findings provide actionable insights and concrete mitigation strategies for safe and fair synthetic medical data sharing. Code is available at <https://github.com/MischaD/Privacy>.

Keywords

Image Generation, Diffusion Models, Privacy, Fairness, Data Intervention

Article informations

<https://doi.org/https://doi.org/10.59275/j.melba.2026-92a6>

©2026 Dombrowski and Kainz. License: CC-BY 4.0

Volume 2026, Received: 2025-12-31, Published 2026-09-22

Corresponding author: mischa.dombrowski@fau.de

Special issue: Medical Imaging with Deep Learning (MIDL) 2025

Guest editors: Lisa Koch, Ronald M. Summers, Chen Chen, Yan Zhuang



1. Introduction

Since the development of statistical models capable of representing and synthesizing new samples from existing dataset distributions (Kingma and Welling, 2013; Goodfellow et al., 2020; Rombach et al., 2022; Ho et al., 2020; Hamamci et al., 2024; Guo et al., 2025), the idea of training generative models on private data and sharing *only* the model or synthetic datasets has gained traction. Such methods could address issues such as data scarcity for rare diseases, gender bias (Larrazabal et al., 2020), and challenges including robust domain adaptation and generalisation (Wang et al., 2022), though these benefits have not yet been demon-

strated with synthetic data. However, maintaining privacy and anonymity is essential when working with personally identifiable information (Jin et al., 2019). Incorporating privacy-preserving techniques (Dockhorn et al., 2022) into the training process, so that only the statistical properties of private data are replicated without revealing any individual samples, therefore holds substantial promise for secure data sharing in healthcare. However, these methods usually lead to degraded downstream performance (Dockhorn et al., 2022; Ghalebikesabi et al., 2023) and have not yet demonstrated to work at scale on high-resolution images.

Recent advances in generative modeling, including diffusion models (Song et al., 2020a; Dhariwal and Nichol, 2021;

Rombach et al., 2022; Ruiz et al., 2023), have expanded the feasibility of sharing models directly (Pinaya et al., 2022). Despite efforts of evaluating the memorization capabilities of generative models (Bai et al., 2021; Dar et al., 2026; Stein et al., 2023), it remains unclear to what extent shared models reproduce training samples, which would raise potential data privacy concerns. Guarantees against privacy breaches would allow models to be trained on proprietary data and shared in place of the underlying datasets, enabling fully anonymous data sharing. Healthcare providers could potentially distribute complex yet anonymous patient information, such as medical images, by modeling population-level data distributions. While federated learning offers an alternative by keeping data local, it requires complex orchestration, secure aggregation, and strict coordination between institutions. Privacy-preserving synthetic data sharing could complement federated approaches by enabling knowledge aggregation without such infrastructural overhead, though it remains an open question whether sufficient privacy guarantees can be achieved in practice (Dombrowski and Kainz, 2025b).

Recent works have shown that publicly released models can inadvertently regenerate training data during sampling. For instance, Somepalli et al. (2023) and Dar et al. (2026) demonstrate that diffusion models can reproduce training samples, while Carlini et al. (2023) illustrate how re-identifiable faces can be extracted from these datasets. This raises serious privacy concerns requiring robust mitigation strategies (Ren et al., 2024). Moreover, some generative models are explicitly designed to memorize training samples (Cong et al., 2020). Differential privacy-based strategies (Dockhorn et al., 2022) offer a promising avenue for mitigating these concerns. However, their adoption in high-resolution data synthesis remains limited due to a notable drop in distribution fidelity, restricted applicability to diffusion models, and low efficacy for multi-modal, high-resolution datasets (Xie et al., 2018). Consequently, in this work, we propose a direct approach to evaluate models, circumventing the need to obfuscate the model distribution, and thereby preserving compatibility with high-resolution, high-fidelity generative tasks. Our approach leverages Poisson image interpolation (PII) for dataset intervention. PII is a technique commonly used in anomaly detection due to its ability to introduce highly realistic yet localized outliers into natural images (Tan et al., 2021; Baugh et al., 2023). These properties make PII particularly suitable for constructing controlled, hard-to-detect feature-level interventions that probe whether generative models memorize, suppress, or generalize rare image content.

Not reproducing unique features has direct and largely unexplored implications for the fairness of generated data. Diffusion models trained on private datasets under strong non-memorization constraints are explicitly discouraged

from reproducing rare or individual-specific characteristics. Hence, these models systematically suppress minority or outlier features, even when such features are legitimate components of the data distribution. This behavior directly conflicts with the goal of fair data generation, which requires faithful representation of both common and rare characteristics rather than their deliberate erasure.

To resolve the trade-off between privacy and fairness, models must learn to generalize. To approach this problem, we can conceptually divide the set of all images into two distinct categories: training set members and non-members. After generating a dataset using a diffusion model, the images fall into one of four categories:

- **Lost images:** not in the training set, unavailable, potentially affecting downstream tasks but posing no privacy or fairness concerns.
- **Memorized images:** reproduced from the private training set, raising privacy issues that require safeguards.
- **Forgotten images:** training images not generated, potentially causing fairness issues if certain subgroups are omitted.
- **Generalized samples:** the ideal case, where the model generates data reflecting the underlying distribution.

We illustrate this in Fig. 1. To probe these categories in a controlled and observable manner, we discuss synthetic anatomical fingerprints (SAFs): rare, manually injected image features that appear only in a small number of training samples. They are not meant to serve as real identifiers, but rather as controllable proxies for investigating memorization. SAFs act as surrogates for sensitive or long-tail characteristics, allowing us to test whether a generative model memorizes them together with the original image, forgets them entirely, or generalizes them across identities. This intervention enables a unified analysis of privacy and fairness within a single experimental framework. For large datasets and unconditional generation, however, explicit sampling becomes computationally infeasible, and the absence of observed memorization is not sufficient evidence that memorization does not occur. To address this, we introduce t' , which serves as a surrogate for the likelihood of generating fingerprint-containing samples when direct sampling is impractical.

This article extends our earlier MIDL 2025 paper, in which we introduced synthetic fingerprints as a tool to study whether generative models reproduce sensitive training data (Dombrowski and Kainz, 2025a). In that work, we formulated a realistic privacy and fairness scenario for unconditional generative models, defined a formal approach to upper bound the probability of reproducing sensitive samples, and proposed the indicator metric t' to measure

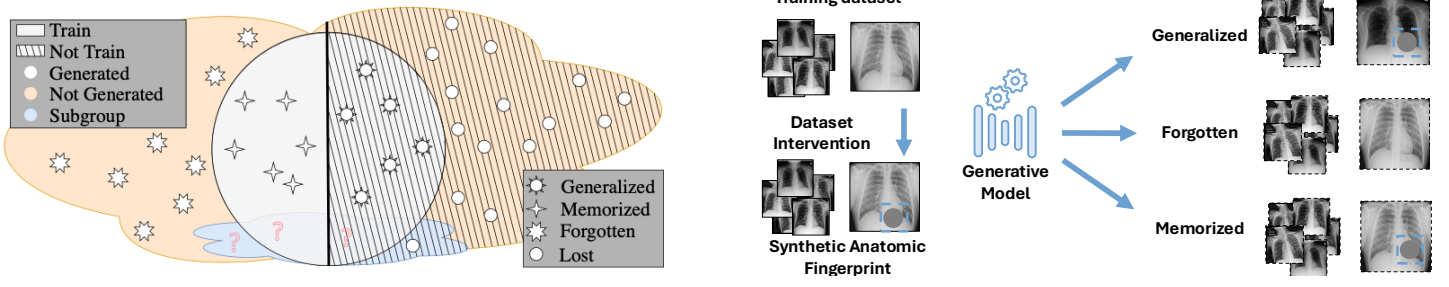


Figure 1: (Left) Training and generated images fall into one of four categories: generalized, memorized, forgotten or lost. Using SAFs, we aim to investigate the category for a selected subgroup of images. (Right) We use synthetic anatomical fingerprints (grey circle inside the blue rectangle) to quantify to which of these categories they belong. Ideally the model learns to pick up their signal and learns to reproduce them without memorizing the entire image (generalization). Memorizing the fingerprints together with their training image would lead to privacy issues. Forgetting them would lead to fairness issues. The dashed border indicates that images are synthetically generated.

this risk. The present work substantially broadens the scope of this exploration. We expand our analysis beyond unconditional generation and evaluate conditional settings, including class-, text-, and mask-guided models. We incorporate additional datasets to assess robustness across domains. We also provide a more extensive review of related work, with a particular focus on privacy auditing, membership inference, and interventional analyses that help contextualize our framework. In summary,

- We expand the methodological context through a more detailed literature review, integrating work in privacy auditing, membership inference, and interventional data analysis.
- We retain and refine the formal framework introduced in (Dombrowski and Kainz, 2025a), which quantifies the probability of reproducing sensitive content via SAFs.
- We broaden the original formulation by evaluating conditional generative models and examining how conditioning signals affect privacy and fairness behavior.
- We reveal a clear relationship between fairness and privacy in diffusion models and the surprisal of conditioning signals, leading to concrete safety recommendations and key mitigation strategies for safe data sharing.

2. Related Works

Diffusion Models Image generation models, such as latent diffusion models (Rombach et al., 2022), model different levels of perturbation $p_\sigma(\tilde{\mathbf{x}}) := \int p_{data}(\mathbf{x})p_\sigma(\tilde{\mathbf{x}} | \mathbf{x})d\mathbf{x}$ of the real data distribution using a noising function defined by $p_\sigma(\tilde{\mathbf{x}} | \mathbf{x}) := \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2\mathbf{I})$. Here, σ defines the strength of the perturbation, split into N steps $\sigma_1, \dots, \sigma_N$. The assumption is that $p_{\sigma_1}(\tilde{\mathbf{x}} | \mathbf{x}) \sim p_{data}(\mathbf{x})$ and $p_{\sigma_N}(\tilde{\mathbf{x}} | \mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma_N^2\mathbf{I})$. We define the optimization as a score matching objective by training a model $s_\theta(\mathbf{x}, \sigma)$ to predict the

score function $\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})$ for the noise level $\sigma \in \{\sigma_i\}_{i=1}^N$. For sampling, this process can be reversed, for example, using Markov chain Monte Carlo methods following Song and Ermon (2019). Song et al. (2020b) extended this to a continuous formulation by redefining the diffusion process as a process governed by an SDE and training a dense model to predict the score function. The continuous formulation of the noising process, denoted by $p_t(\mathbf{x})$ and $p_{st}(\mathbf{x}(t) | \mathbf{x}(s))$, characterizes the transition kernel from $\mathbf{x}(s)$ to $\mathbf{x}(t)$, where $0 \leq s < t \leq T$. Anderson (1982) showed that the reverse of this diffusion process is also a diffusion process. Song et al. (2020b) show that the reverse diffusion process of the SDE can be modeled as a deterministic process, as the marginal probabilities can be expressed deterministically in terms of the score function. As a result, the problem simplifies to an ODE, which can be solved using any black-box numerical solver, such as the explicit Runge-Kutta method. This enables exact likelihood computation, commonly used to estimate the likelihood of generating a sample, e.g., images (Song et al., 2020b). In this context, we propose t' , a more general indicator that extends this idea to approximate the likelihood of generating all samples that lead to privacy problems.

Memorization Detection and Mitigation Prior work shows memorization in diffusion models, specifically in the context of text-conditional diffusion (Carlini et al., 2023; Somepalli et al., 2023; Ren et al., 2024). These findings show how text prompts can serve as keys that can retrieve near-perfect copies of training images, which raises privacy and copyright concerns. One of the key approaches for privacy auditing is membership inference attacks (Wu et al., 2023; Pang and Wang, 2023; Pang et al., 2023; Li et al., 2024). They mostly work on reconstruction-loss style attacks, either with or without knowledge of the architecture itself. More importantly, they require access to the datasets that they want to infer membership on, which in practice

often is not true.

Another direction is to adopt differential privacy (Dockhorn et al., 2022; Ghalebikesabi et al., 2023; Wang et al., 2024; Liu et al., 2024). These methods modify the training procedure to provide explicit privacy guarantees and a quantifiable privacy budget, but at the cost of substantially increased training complexity. As a result, their practical applicability has so far been demonstrated mainly on small datasets (Ghalebikesabi et al., 2023). Similarly, differentially private fine-tuning (Tsai et al., 2025) can mitigate memorization but typically leads to a severe degradation in image quality. In general, the impact of applying these methods is too disruptive to be feasible and is often highly specific to the model architecture. Moreover, they require specialized training procedures that are incompatible with existing pretrained models, which we do not have access to. For these reasons, we argue that improving the understanding and detection of memorization is currently a more practical and effective route toward mitigation.

Current work also actively explores the interplay between training hyperparameters and image memorization of diffusion models. Bonnaire et al. (2026) explore training the interplay between model size, training length, and dataset size and derive scaling rules for it. They specifically focus on the region when the generative model starts to generalize. Wu et al. (2025) observed in a theoretical framework, that the learning rate of diffusion models is also a key factor for model memorization.

To formalize and contextualize our approach, we borrow the definitions of *extractable memorization* and *discoverable memorization* from the natural language processing domain (Nasr et al., 2023; Carlini et al., 2021) and apply them to generative image models. Given a model s with a generation routine Gen , an example $\mathbf{x}_p$ from the training set D is *extractably memorized* if an adversary (without access to D) can construct a conditioning $\mathbf{c}$ that makes the model produce $\mathbf{x}_p$ (i.e., $\text{Gen}(\mathbf{c}) \approx \mathbf{x}_p$). We also adopt and extend the definition of *discoverable memorization* from Nasr et al. (2023) and Carlini et al. (2021) to image models: For a model s with generation routine Gen , an example $\mathbf{x}_p \in D$, and a perturbation function from the generative model’s training $p_\sigma(\tilde{\mathbf{x}} | \mathbf{x})$, $\mathbf{x}_p$ is *discoverably memorized* if $\text{Gen}(\tilde{\mathbf{x}}_p, \sigma) \approx \mathbf{x}_p$ with high probability over draws of $\tilde{\mathbf{x}}_p \sim p_\sigma(\tilde{\mathbf{x}} | \mathbf{x}_p)$. The strength of the perturbation function directly influences how discoverable the training images are. Our proposed indicator t' measures the susceptibility of models to discoverable memorization. In terms of our SAF-based metrics, C_{f+} (formally defined in Sec. 3.3) detections on exact identity matches correspond to discoverable memorization, while t' estimates the capacity for extractable memorization by measuring how far the score function remains collapsed toward the training sample. It can be compared to the privacy budget in differential pri-

vacuity (Dockhorn et al., 2022). However, unlike differential privacy methods, which only work on low-resolution images, our approach’ post-hoc nature preserves image quality.

Fairness Fairness in AI is a well-explored yet unsolved problem. Current directions in the literature for discriminative tasks suggest frameworks for benchmarking (Jin et al., 2024) or reveal important design choices for training fair models. Most of them provide guidance for designing and testing downstream models for fairness (Yang et al., 2024). Generative models are mainly used to improve fairness (Ktena et al., 2024; Uwaeze et al., 2025), but their own biases and unfairness remain underexplored. Throughout this work, we use “fairness” to refer specifically to representational and distributional fairness, i.e. whether rare or long-tail features in the training data are preserved in generated outputs. This is distinct from clinical downstream fairness notions such as demographic parity or equalized subgroup performance, which depend on task-specific evaluation. It is often assumed that generative models learn the entire data distribution without further evaluation. Current approaches mainly investigate fairness of generative models by looking at demographic statistics (López-Pérez et al., 2025). Relatedly, a growing body of work studies long-tail generation in diffusion models, aiming to improve the fidelity and diversity of rare classes. Zhang et al. (2024) propose calibrated diffusion mechanisms that leverage images from the head classes to better generate tail-class images. Samuel et al. (2024) address rare concept generation by selecting optimal noise seeds at inference time, enabling faithful synthesis of infrequent visual concepts without re-training. Hayden et al. (2025) further connect long-tail generation to robustness by guiding diffusion models toward epistemically uncertain regions, improving downstream generalization through targeted data synthesis. While these methods substantially advance tail coverage and utility, they primarily evaluate success through class-conditional fidelity, diversity, or downstream performance gains. Both approaches do not properly account for the fact that within certain subgroups there may be demographic differences as well. Dombrowski et al. (2025b) for example explore how a retrieval-based metric can be used to quantize gender unfairness in text-to-image models. This highlights an important gap between long-tail generative modeling and fairness analysis: improving coverage of rare features does not necessarily ensure fair representation, and may interact with privacy and demographic imbalance in subtle ways.

Interventional Probing Interventional analysis focuses on identifying causal effects by actively modifying the data a model learns from or is evaluated on. Rather than relying on correlations in model behavior, it tests whether controlled changes to the data distribution lead to predictable changes in model outputs. This type of analysis is most commonly

applied in downstream settings, where generative models are used to synthesize targeted interventions (Yuan et al., 2022; Yin et al., 2023; Xia et al., 2024; Liang et al., 2025). A central concept in this area is counterfactual image generation, which explicitly studies the causal and non-causal relationships between interventions and model predictions and is widely used to investigate spurious correlations. Melistas et al. (2024) introduced a framework for benchmarking counterfactual image generation, focusing on diversity and fidelity of the generated interventions. Ak et al. (2019) use GANs to generate image-level interventions by specifying semantic attributes and leveraging attention maps to localize the intervention to specific regions. Mao et al. (2021) perform model-level interventions in the feature space of a generative model to steer generations along semantically meaningful directions, improving robustness of downstream models. Weng et al. (2024) study interventions at the classifier level, using diffusion-based counterfactuals to detect and mitigate shortcut learning. Unlike these approaches, our method focuses on data-interventions for training the diffusion model itself, not the downstream models.

Evaluation of Image Generation Models The evaluation of diffusion models is predominantly centered on fidelity by comparing distributional similarity, most commonly measured using the Fréchet Inception Distance (FID) (Heusel et al., 2017). Sample diversity is typically assessed via precision and recall metrics (Kynkäänniemi et al., 2019; Rombach et al., 2022; Sauer et al., 2022; Dhariwal and Nichol, 2021; Peebles and Xie, 2023). While widely adopted, these metrics offer limited interpretability and provide little insight into how individual training samples or rare features are represented. Konz et al. (2026) propose the Fréchet Radiomic Distance (FRD), a perceptual metric based on clinically meaningful radiomic features that better captures anatomical differences than FID. Retrieval-based evaluation using image retrieval score (IRS) (Dombrowski et al., 2025b) has been proposed as an alternative, enabling more fine-grained analysis of representational similarity.

Memorization in generative models is most commonly studied using loss-based criteria, which analyze discrepancies in likelihood or reconstruction error between training and non-training samples (Bonnaire et al., 2026). Other approaches frame memorization as a copy-detection problem and rely on trained Siamese networks to identify near-duplicates between generated and training images (Dombrowski et al., 2025a). While effective, these methods operate externally to the generative process and treat the model largely as a black box.

In contrast, our indicator t' directly exploits the generative mechanism of diffusion models. By leveraging the internal structure of the diffusion process and its associated noise perturbations, t' provides a principled, model-aware

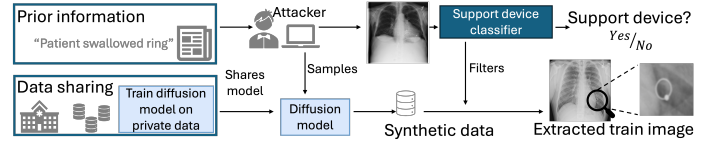


Figure 2: Extracting memorized samples from a trained diffusion model. The attacker learns that a ring is in the patient’s image, uses this information, and filters generated samples until reproducing the training image.

signal of memorization risk. This formulation enables a mathematically grounded analysis of memorization that yields insights beyond surface-level similarity metrics and is naturally aligned with the probabilistic foundations of diffusion models.

Privacy-preserving training Privacy-aware training strategies such as DP-SGD (Dockhorn et al., 2022) modify the optimization process to prevent memorization by construction. While principled, these methods currently degrade image quality substantially and do not scale to high-resolution generation. Our framework instead audits existing models post hoc, including pretrained models where retraining with differential privacy is impractical.

3. Method

We define the key terms used throughout the paper as follows:

- **Fingerprint:** An image feature that is unique to a specific person or image in the dataset (i.e., uniquely associated with a single identity, though it may appear in multiple training samples). Examples include distinctive diseases, objects, entire faces, bone structures, or medical devices. Similar to real fingerprints, their presence alone is not inherently problematic unless additional information enables identification. We use a classifier C_f (predicting the presence of a fingerprint) and C_{id} (predicting the identity) to detect these features in generated images.
- **Privacy:** Sharing synthetic data poses a problem if an adversary, without access to any images from the training dataset but with prior knowledge of a specific fingerprint, can extract an image from the synthetic dataset and recognize that it is memorized. Concretely, the adversary filters generated samples using knowledge of the fingerprint (e.g., knowing that a swallowed ring appears in a training image), and recognizes memorization when the generated image reproduces the fingerprint together with the appearance of the original training subject. An example of this scenario is shown in Figure 2.
- **Memorization:** Memorization refers to the pixel-wise

reproduction of training images. We distinguish between full-image memorization and partial memorization. To detect partial memorization, we define fingerprints a priori and use a classifier C_f trained to be robust against perturbations, rather than relying on pixel-wise similarity. Full-image memorization occurs when the model reproduces the entire training image.

- **Identity:** Case-dependent sensitive information that may be leaked if the model is shared. This can refer to the reproduction of full images, as in ChestX-ray14 (Wang et al., 2017) or CelebA-HQ (Karras et al., 2017), but also to the identity of a person independent of image context or background.
- **Violation:** A privacy violation occurs when the reproduction of a fingerprint implies the identity of the original training image. Formally, this means that the presence of a fingerprint implies the presence of the associated identity.

3.1 Conditioning Modalities.

We consider both unconditional and conditional diffusion models. Unconditional image generation serves as a baseline setting, where the model is trained and sampled without any external conditioning signal. In addition, we evaluate diffusion models under several common conditioning paradigms. Specifically, we consider (i) class-conditioned generation, where images are generated conditioned on discrete diagnostic labels, (ii) text-conditioned generation, where free-form clinical reports or prompts guide synthesis, (iii) mask-conditioned generation, where spatial constraints such as segmentation masks are provided as input, and (iv) feature-conditioned generation, where images are conditioned on feature vectors extracted from pre-trained foundation models. These conditioning signals differ in structure and information content, but are all evaluated using the same synthetic anatomical fingerprint framework and memorization indicator t' .

3.2 Synthetic Anatomical Fingerprints.

Using manually added SAFs, we study privacy and fairness within a unified experimental framework. Privacy concerns arise when SAFs are reproduced together with the identity of the original training image, indicating memorization. Note that our notion of memorization includes approximate reproduction: a generated image is considered memorized if the fingerprint classifier C_f and identity classifier C_{id} both yield positive predictions, regardless of whether the reproduction is pixel-exact. Conversely, fairness concerns emerge when SAFs are systematically absent from the generated data, suggesting that rare or unique features present in the training distribution are not retained by the model. Our goal

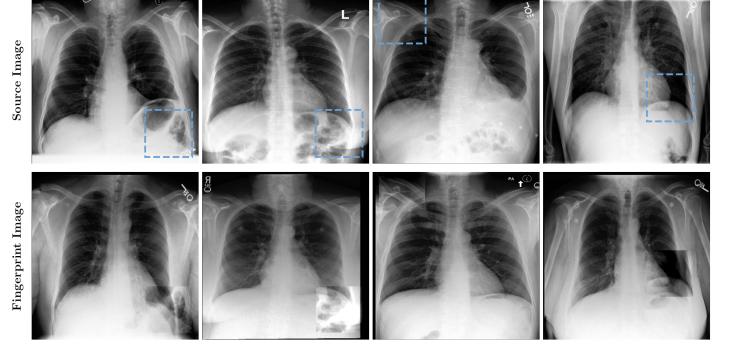


Figure 3: Example SAF: To create realistic fingerprints, we use source patches (top row) and inpaint them into target images (bottom row) using poisson image editing. Depending on the similarity of source and target image, the fingerprints are barely visible.

is to achieve generalization, where SAFs are reproduced in images with different identities. To investigate when models start to generalize, we artificially inject detectable objects into the training data, *i.e.*, SAFs. We then train one classifier to detect these objects and another to identify the image’s identity used as the target for injection. A non-privacy-violating and fair model would reproduce the SAF on a synthetic image with a different identity than the training image containing the fingerprint. A few examples of SAFs are shown in Fig. 3.

To generate SAFs, we synthetically augment a single sample x_p from the dataset D . In practice, this can be any feature that appears only once in the entire training dataset, such as a ring, a deformation, or a specific medical device. In our experiments, we consider three inpainting strategies: (i) inserting a synthetic gray circle, (ii) using PII to inpaint a realistic feature extracted from another image, and (iii) incorporating known unique attributes derived from image-level metadata (e.g. only leave one female in the dataset).

Classifier Trainings We use two binary classifiers: (i) C_f , which detects the presence of a fingerprint, and (ii) C_{id} , which predicts the identity associated with a training image. Both classifiers output 1 for a positive prediction. We define the set of potentially memorized samples as

$$q := \{x : C_f(x) = 1 \wedge C_{id}(x) = 1\}, \quad (1)$$

and denote its cardinality by $|q|$. When drawing N_{gen} images from p_s , the expected number of such samples under a baseline assumption of uniform coverage of the training set is

$$\mathbb{E}(|q|) = N_{\text{gen}}/N_{\text{train}}, \quad (2)$$

where N_{train} is the training set size. This baseline models a hypothetical fair model that samples each training image

with equal probability; deviations between observed $|q|$ and $\mathbb{E}(|q|)$ indicate either memorization (over-representation) or suppression (under-representation) of the fingerprinted sample. We assume that the adversary has access to the fingerprint and therefore access to C_f . The attacker does not have access to C_{id} . Its role is to determine whether the presence of a fingerprint implies the identity of the original training image. This setup allows us to disentangle the memorization of the SAF from the memorization of $\mathbf{x}_p$, distinguishing generalization from memorization. To track the number of memorized samples, we define $|q|$ as the number of synthetic samples where both classifiers have a positive outcome.

3.3 Memorization Indicator t' .

For large datasets and unconditional training setups, directly assessing memorization by explicitly generating samples becomes infeasible. We therefore introduce t' , an indicator that captures a model's capacity to memorize training images. While it is possible to compute the likelihood of the exact sample (e.g., using numerical NLL estimation), but this does not ensure that images in the immediate neighborhood are free from privacy issues. To address this, we propose estimating the upper bound of the likelihood of reproducing samples from the entire subspace belonging to the class of private samples.

Let $p_s(\mathbf{x}_p)$ define the likelihood of the unconditional model s reproducing the private sample $\mathbf{x}_p$ at test time. This alone is insufficient because it does not account for slightly noisy versions of $\mathbf{x}_p$, which can also pose privacy concerns. We aim to compute $q(p)$, defined as the likelihood of reproducing any sample within Ω_p . Here, Ω_p represents the region in image space that is similar enough to $\mathbf{x}_p$ to raise privacy concerns according to q .

In Appx. A, we show this is equivalent to:

$$\begin{aligned} q(p) &= \int_{\Omega_p} p_s(\mathbf{x}) d\mathbf{x} \approx \int_0^{t'} p_s(\mathbf{x}_{t,p}) dt \\ &\leq \sum_{i=0}^{t'} \sup_{t \in [t_i, t_{i+1}]} (\sigma_{t_{i+1}} - \sigma_{t_i}) \mathbb{E}_{p(\mathbf{x}_{t,p})} [p(\mathbf{x}'_{t,p})]. \quad (3) \end{aligned}$$

To estimate $q(p)$, we observe that it depends only on the likelihood $p(\mathbf{x}'_p)$ and t' , which captures the entire region of Ω_p . $\mathbf{x}'_p$ is the predicted sample of the diffusion model after applying t forward diffusion steps to the private sample $\mathbf{x}_p$. This synthetic $\mathbf{x}'_p$ then serves as input to the classifiers. Ω_p is defined as the region where C_{id} and C_f both give positive predictions. Since this region depends only on its size, t' serves as an indicator of how unlikely it is to generate critical samples from the model, without the necessity to compute the exact value for $p(\mathbf{x}'_p)$. Intuitively, t' measures how far forward in the diffusion process we have to go for

the generative model to produce a different image than the training image.

Assumptions and failure modes. The validity of t' rests on the assumption that the identity classifier C_{id} has learned a meaningful decision boundary for whether a generated image reproduces a specific training identity. t' can produce misleading signals in two directions: (1) **False sense of safety:** a low t' does not guarantee the absence of memorization; it only indicates that the sampling trajectory diverges from the training sample at relatively high noise levels. (2) **False alarm:** if C_{id} has poor specificity (e.g., confusing visually similar but distinct individuals), t' may overestimate memorization capacity. We mitigate this through strong augmentation and the high classifier accuracies reported in Table 1.

Computational cost. Computing t' for a single training sample requires M reverse diffusion trajectories at each noise level (Algorithm 1); with $M=16$ this amounts to $16 \times$ the cost of generating one image. The search terminates early once the classifier no longer detects the training identity. The main bottleneck is training the classifiers C_f and C_{id} per dataset. Exhaustive evaluation of t' over all training images is infeasible for large datasets; we recommend computing t' on a targeted subset (e.g., rare or high-risk samples identified by domain experts) rather than exhaustively.

Capacity vs. realized memorization. It is important to distinguish t' from C_{f+} . t' characterizes the learned score function around a training sample and measures the model's *capacity* for memorization. C_{f+} counts the number of generated images in which C_f detects the fingerprint, measuring *realized* memorization in a finite sample. A model can exhibit high t' while producing zero C_{f+} detections if the memorized region is unlikely to be reached during standard sampling.

3.4 Class Conditional Image Generation

Conditional diffusion models incorporate auxiliary signals such as class labels, text prompts, segmentation masks, or feature vectors to guide generation. These signals can restrict sampling to narrow regions of the data manifold and may therefore act as keys that amplify memorization when they are rare or highly specific.

To formalize the strength and rarity of conditioning signals across modalities, we quantify their information content using concepts from information theory. We illustrate our idea using class conditional image generation. For each class c , we compute the class-wise entropy by evaluating the Shannon entropy of the empirical class distribution $p(c)$, as defined in Appendix C:

$$H(C) = - \sum_c p(c) \log_2 p(c). \quad (4)$$

The surprisal of a specific sample with class label c is then given by its information content

$$I(c) = -\log_2 p(c), \quad (5)$$

which reflects how unlikely that class is under the dataset distribution.

Similarly, we retrieve the entropy for all other considered conditioning modalities. We present detailed derivations in Appx. C. For text-conditioned generation, we compute the entropy token-wise. Let $t \in T$ denote a token in the vocabulary. We estimate each token's marginal probability p_t as the fraction of prompts in the dataset that contain it:

$$H_{\text{text}} = \mathbb{E}[I(X)] = - \sum_{t \in T} [p_t \log_2 p_t + (1-p_t) \log_2 (1-p_t)]. \quad (6)$$

For masks, we model each pixel as an independent Bernoulli random variable with parameter p_{uv} . The resulting pixel-wise entropy map is given by

$$H_{uv} = -p_{uv} \log_2(p_{uv}) - (1-p_{uv}) \log_2(1-p_{uv}), \quad (7)$$

and the total mask entropy is $\sum_{u,v} H_{uv}$.

For feature-conditioned diffusion models, we model the feature distribution as a multivariate Gaussian (most commonly done for FID computation), compute the differential entropy, and discretize it to obtain the discretized surprisal $I_{\text{disc}}(z_i)$ of the i -th feature of the conditioning vector z :

$$H_{\text{disc}}(Z) = \frac{1}{N} \sum_{i=1}^N I_{\text{disc}}(z_i). \quad (8)$$

Note that the absolute values of these entropies are not comparable as they all rely on different assumptions and represent different things, but their relative order is comparable. Specifically, we can quantize which images are the most and the least surprising for the generative model. We show a few examples of unsurprising and surprising conditionings in Fig. 4.

In all conditional settings, SAFs remain the probe for memorization and fairness. Conditioning affects how likely fingerprint-containing regions are accessed during sampling, but does not change the definition of memorization. We therefore evaluate conditional models using the same SAF-based framework.

4. Experiments

We structure the experimental evaluation in two stages. First, we analyze unconditional diffusion models to establish baseline memorization and generalization behavior under controlled variations of dataset size, training length,

model capacity, and fingerprint frequency. Second, we extend the analysis to conditional generation, where external conditioning signals may act as keys that amplify memorization or suppress rare features. This design allows us to separate intrinsic model behavior from effects induced by conditioning.

Starting with unconditional image generation, we study the effects of dataset size, training duration, model capacity, and the frequency of SAFs on memorization and fairness behavior. For conditional generation scenarios, we include class-, text-, mask-, and feature-conditioned diffusion models. We specifically investigate whether the surprisal of the conditioning signal plays a critical role.

Classifier Training The classifiers are randomly initialized ResNet50 (He et al., 2016) architectures. To maximize robustness, we employ AugMix (Hendrycks et al., 2019), and in the case of C_{id} , we inject random Gaussian noise into the training images to increase the robustness towards possible artifacts from the diffusion process. Furthermore, we randomly mask out patches of the same shape as the SAF by setting pixel values to zero, reducing the effect of SAF on the prediction. Robustness is crucial for these classifiers. Even if models have a 99.9% accuracy on the test set, they produce a not negligible amount of false predictions on a dataset with 50000 synthetic images. Therefore, we run several training sessions over different hyperparameter settings. Due to the simplicity of this detection and the self-supervised learning scheme, all of the classifiers trained to detect synthetic fingerprints reach an accuracy of 100% on the test set. Since both tasks are fairly easy binary classification tasks, we employed strong augmentation techniques to ensure that positively predicted samples from the classifiers are SAFs. We balanced the classification task for C_{id} by adding SAFs to 50% of the training images. For validation, we reduce this to 10% to remain closer to the expected distribution. For C_{id} , we chose circular masking as a training augmentation to make its predictions invariant to the presence of SAFs.

Classifier robustness. The reliability of our framework depends on the quality of C_f and C_{id} . We train multiple instances per dataset with varied hyperparameters and select the best-performing model. As reported in Table 1, C_f achieves 100% test accuracy across nearly all datasets, and C_{id} exceeds 99.5% in all cases. For the conditional experiments on real-world data (ChestX-ray14, MIMIC-CXR), we replace C_{id} with the LCMem re-identification foundation model, achieving >96% recall. Despite these high accuracies, even 99.9% accuracy produces non-negligible false predictions over 50k generated images. We therefore employ strong augmentation and, where necessary, manual verification of flagged samples.

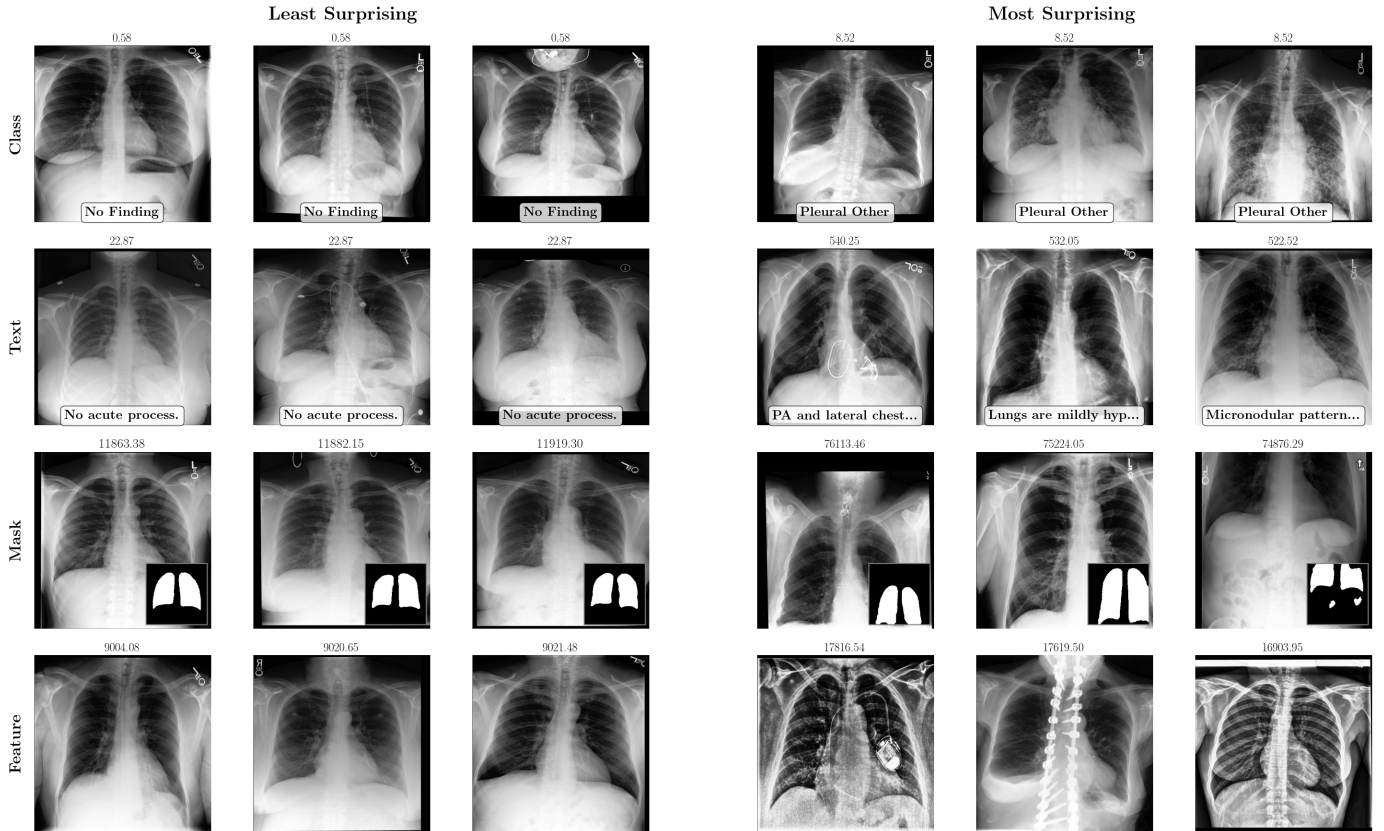


Figure 4: Example images with low (left) and high (right) surprisal.

4.1 Unconditional Image Generation

For our experiments on unconditional image generation, we consider the size of the training dataset, the training length, and model size as the three most impactful factors determining a model’s fairness and memorization capabilities. We first establish our method on a toy dataset, then we investigate how to tune these parameters if we want to achieve generalizability and we investigate the influence of the choice of the synthetic anatomic fingerprint.

Dataset For our initial experiments we use an a-priori selected selection of modalities from MedMNISTv2 (Yang et al., 2023). For unconditional image generation, we use ChestX-ray14 (Wang et al., 2017), a dataset of 112,120 frontal chest X-rays widely studied in privacy research (Packhäuser et al., 2022). For our experiments on training length and number of fingerprints per dataset, we use three datasets with diverse modalities and sizes. Specifically, we report results on BCI (Liu et al., 2022) and on ODIR-2019¹. Data is split (60/20/20), with diffusion models sharing training data with classifiers. For conditional image generation, we use MIMIC-CXR (Johnson et al., 2019), which also comes with text impressions and precomputed lung masks (Indeewara et al., 2023; Goldberger et al., 2000).

Metrics To evaluate generative quality, we report FID. To quantify fairness and privacy, we compute C_{f+} , the number of generated images for which the SAF classifier C_f predicts the presence of the synthetic fingerprint. To evaluate memorization in the unconditional setting, we compute t' .

4.1.1 Toy Datasets

Experimental Details We start by training C_{id} , C_f and diffusion models on MedMNIST. We train an image space diffusion model based on (Song et al., 2020a) for a fixed number of diffusion steps and until convergence. The diffusion component follows a compact 2D U Net design that matches the resolution of the MedMNIST images. Because the inputs are only 28×28 , the model employs just the three outermost downsampling and upsampling blocks of the underlying U-Net. This reduced architecture provides sufficient capacity for the toy domains while keeping computation manageable. A fixed training length of thirty thousand steps is used for all MedMNISTv2 datasets. Earlier experiments showed that this schedule is sufficient for the model to reliably reproduce examples from the smallest subsets, and extending training beyond this threshold offered no consistent improvement. Training a single model on one dataset requires roughly eleven hours on a single A100 GPU.

1. <https://odir2019.grand-challenge.org/dataset/>

After convergence, each trained model is used to generate fifty thousand synthetic samples. This number of samples supports sufficient quantitative evaluation of the probability that the generative process reproduces a training example under test conditions.

Table 1: Training results for different MedMNIST datasets. A privacy-preserving model has a low value for t' , indicating a low likelihood of reproducing the sample itself, but the expected incidence of the SAF should remain close to the observed one $\mathbb{E}(|q|) \sim |q|$. $|N_D|$ denotes the number of training samples.

Description		SAF classification		Data synthesis				
Dataset	$ N_D $	SAF (%)	ID (%)	FID_{train}	FID_{test}	$\mathbb{E}(q)$	$ q $	t'
BreastMNIST	546	100	98.7	9.2	62.6	91.6	57	0.886
RetinaMNIST	1080	100	99.6	5.9	19.7	46.3	52	0.998
PneumoniaMNIST	4708	100	99.8	9.5	28.4	10.6	2	0.718
BloodMNIST	11959	100	99.5	9.3	11.0	4.2	0	0.241
OrganSMNIST	13940	99.47	99.8	19.6	19.7	3.6	0	0.582
ChestMNIST	78468	99.93	99.8	3.3	3.9	0.6	0	0.206

Results The results on the toy datasets are shown in Table 1. For smaller datasets, the diffusion models reproduced training images rather than generating new content. The combined prediction set $|q| = |C_{id}^+ \cap C_f^+|$ was non-zero only for the smallest datasets, indicating that reproduced fingerprints were always accompanied by the original training identity. A consistent transition from memorization to generalization appears near $|N_D| \approx 5000$. For larger datasets, t' values are small and no memorized samples are recovered. The gap between FID_{train} and FID_{test} alone does not reliably indicate memorization: PneumoniaMNIST shows a more pronounced FID gap than RetinaMNIST yet exhibits almost no evidence of memorization. Overall, when the number of images is limited relative to model capacity, the diffusion process may collapse onto training examples, and t' provides a practical signal for detecting this behavior.

4.1.2 Impact of Dataset Size

Building on these observations from the toy datasets, we now examine how memorization behaves in a realistic large scale setting using the ChestX-ray14 dataset. Here the

Table 2: Quantitative results on CXR data using two backbones: OD (out-of-domain) Inception and ID (in-domain) models for CXR (Cohen et al., 2022). Larger datasets reduce memorization risk, quantified by t' .

	$ N_D $	Classification		OD (Inception)		ID (CXR)		Privacy		
		SAF (%)	ID (%)	FID_{train}	FID_{test}	FID_{train}	FID_{test}	$\mathbb{E}(q)$	$ q $	t'
ChestX-ray14	875	100.00	100.00	15.1	30.3	1.0	2.3	34.3	47	0.75
	1750			12.3	29.3	1.0	2.4	17.1	5	0.86
	3500			13.6	32.0	1.2	2.6	8.6	1	0.67
	7001			18.8	38.4	1.6	3.0	4.3	0	0.72
	14003			22.1	41.4	1.9	3.3	2.1	0	0.66
	28007			19.9	39.4	2.1	3.4	1.1	0	0.60

number of available training images was varied systematically from 875 to 28,007 while keeping all other conditions fixed. The latent diffusion backbone was trained for 150000 steps, after which 30000 samples were generated for each configuration. As shown in Table 2, dataset size directly determines the balance between memorization and generalization. At $|N_D| = 875$, the model reproduced 47 training instances with elevated t' values. Increasing dataset size steadily reduced this risk; at $|N_D| = 28007$ no memorized samples were detected. A residual privacy risk remains even without detected reproductions: at $|N_D| = 7001$, the high t' value suggests memorization capacity persists despite zero observed SAF reproductions. FID values are misleading in this regime, as smaller datasets yield lower FID due to memorization rather than better generative quality. The comparison between $|q|$ and its expectation highlights a fairness concern: at $|N_D| = 7001$, the expected 4.3 SAF reproductions yielded zero, indicating systematic underrepresentation of long-tail features.

4.1.3 Impact Of Training Length

Experimental Details To study how training duration affects synthetic image quality, we train diffusion models for multiple training lengths and compute the FID at each checkpoint. The model achieving the lowest FID for a given dataset is considered the best configuration and is used for all subsequent analysis.

For this initial investigation, we employ PII as the SAF. PII is embedded into the training data before model optimization and subsequently removed during evaluation, allowing us to quantify the impact of training length on fingerprint retention.

Results As shown in Figure 5, the effect of training length on memorization varied substantially across the three datasets. For ChestX-ray14, extended training increased both PSNR and FID, indicating reduced sample diversity and diminished image quality. Despite this degradation, the model never reproduced the embedded SAF, suggesting that overtraining primarily led to mode collapse rather than direct memorization. The BCIdataset showed a small number of positive fingerprint predictions during early epochs, but closer inspection revealed these to be false positives. The reduced image quality in early training likely caused spurious classifier responses. The model selected based on the lowest FID did not reproduce the SAF, and its comparatively high PSNR was driven by synthetic images containing large uniform or empty regions rather than genuine similarity to training samples. In contrast, ODIR-2019 demonstrated clear and consistent memorization. The model reproduced the SAF frequently, reflected by numerous positive C_{f+} predictions and elevated PSNR values. All detected cases corresponded to exact copies

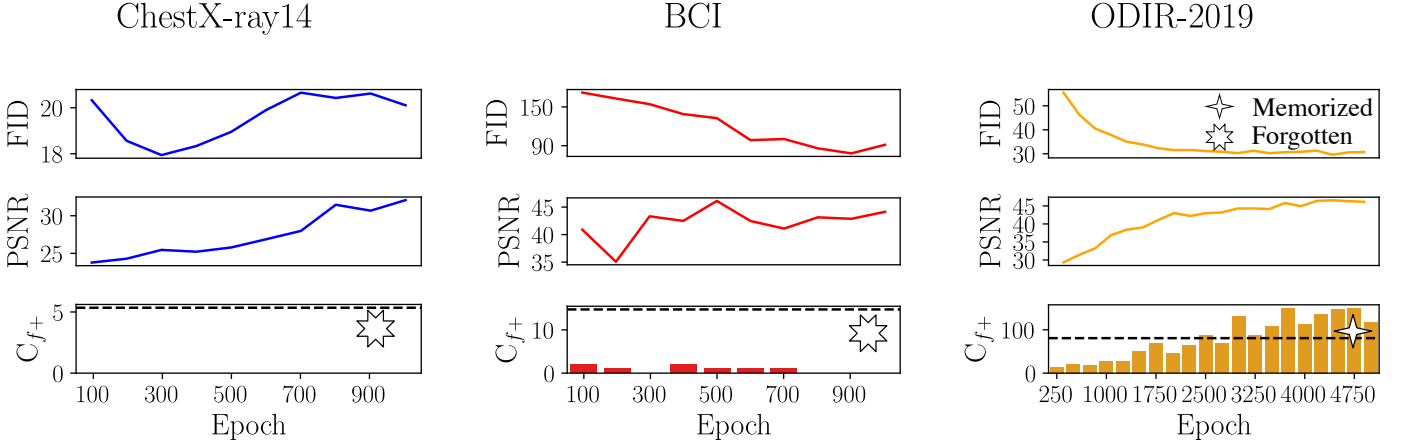


Figure 5: Impact of training length. To assess model memorization, we investigate PSNR and C_{f+} . The dashed line in the bottom row indicates the results of a fair model (SAFs appear equally often in training and synthetic datasets).

Table 3: Privacy metrics for different model sizes. Architecture details are given in Table 6.

	# Trainable parameters	$ q $	FID	t'
Default	113 675 524	5	32.7	0.77
Model 1	77 364 740	9	33.0	0.74
Model 2	71 439 108	0	33.9	0.69
Model 3	49 558 020	1	34.8	0.69
Model 4	28 484 612	0	78.7	0.66
Model 5	28 448 388	0	43.6	0.64

of the fingerprint-containing training sample, confirming memorization rather than incidental similarity. Overall, the results show a dichotomy across datasets. Depending on the data domain and generative difficulty, diffusion models either learn to reproduce the SAF directly, introducing privacy concerns, or generate limited and low-diversity outputs that impair fairness and utility.

4.1.4 Impact of Model Size

Experimental Details To assess how architectural capacity influences memorization independently of dataset size, we train a series of diffusion models on a fixed subset of ChestX-ray14 containing $|N_D| = 1770$ images. The experiment varies the number of trainable parameters by constructing six U-Net backbones that differ in depth, channel width, and number of down blocks. All models are trained under identical conditions and evaluated using FID, the indicator t' , and the number of recovered memorized samples $|q|$.

Results The relationship between model size, generative quality, and memorization potential is summarized in Tab. 3, while the full architectural details are provided in the

Table 4: Memorization results for all three datasets. Results are averaged over three different runs. All ChestX-ray14 and BCI runs result in fairness issues due to the complete lack of reproducing the SAFs.

	PSNR			C_{f+}			t'		
	Circle	P11	Feature	Circle	P11	Feature	Circle	P11	Feature
ChestX-ray14	24.76	25.06	24.54	0.00	0.00	0.00	0.14	0.25	0.43
BCI	46.41	46.48	46.01	0.00	0.00	0.00	0.10	0.26	0.25
ODIR-2019	46.93	46.18	47.14	51.33	61.67	39.33	0.76	0.77	0.78

appendix. The results reveal a clear trade-off: larger models exhibit higher t' and more recovered training samples, while smaller models produce $|q| = 0$ but at the cost of increasing FID. Manual inspection confirms that smaller models avoid memorization by forgetting long-tail features rather than learning to generalize them. Reduced capacity thus protects privacy only by suppressing rare patterns.

4.1.5 Impact of Type of Fingerprint

Experimental details We evaluate three types of synthetic anatomical fingerprints (SAFs) to assess whether our findings depend on the specific form of the injected feature. *Circular fingerprints* are purely synthetic and consist of a gray disk with a fixed radius of 72 pixels, placed at a predefined location within the region typically occupied by image content. These fingerprints are visually simple and trivial to detect, which makes automated detection reliable and serves as a controlled baseline. To study visually more realistic fingerprints, we construct *P11-based fingerprints* using P11, which is known to introduce highly realistic local features and is commonly used in anomaly detection settings (Tan et al., 2021). For ChestX-ray14, we select source images containing a medical support device and inpaint the corresponding region into target images at locations where such devices are not present. This ensures that no real

training sample contains a visually similar feature at the target location. We experiment with *feature fingerprints derived from image-level labels*. Specifically, we use sex for ChestX-ray14, staining type for BCI, and the physical side of the eye for ODIR-2019. Since these attributes are real but globally defined, we synthetically localize them by applying PII to all images with the opposite label. This allows us to reuse the same classifier used for PII-based fingerprints to detect these features. Effectively, this procedure turns a real attribute into a synthetic, localized fingerprint while preserving a robust decision boundary for detection. To quantify memorization behavior across fingerprint types, we additionally compute the peak signal-to-noise ratio (PSNR). Specifically, we compute the PSNR between 1,000 generated samples and all images in the training set, and report the maximum value. High PSNR values indicate strong pixel-level similarity to a training image and are therefore indicative of potential memorization.

Results Results are reported in Tab. 4 and averaged over three independent runs. Across ChestX-ray14 and BCI, none of the models reproduce the injected fingerprints, independent of fingerprint type, indicating systematic fairness issues. The three fingerprint types yield similar trends in both Cf^+ and t' . Circular fingerprints are most easily forgotten (lowest t'), while PII-based and feature-based fingerprints produce comparable t' values, suggesting that visually realistic interventions behave more similarly to real attributes. Overall, the type of fingerprint has only a minor influence on the observed behavior, indicating that our findings are robust to the specific choice of fingerprint construction.

4.1.6 Impact of Number of Synthetic Anatomic Fingerprint

Experimental Details We next study how many fingerprinted training samples are required for a diffusion model to generalize a unique feature rather than memorizing or discarding it. For each dataset, we vary the number of fingerprinted examples N_f while keeping the remaining training conditions fixed. The objective is to determine the smallest value of N_f for which the model begins to synthesize the fingerprint rather than reproducing exact copies or failing to express it altogether. We visualize the results in Figure 6.

Results Across all datasets, successful generalization requires a minimum number of fingerprinted examples. For ChestX-ray14, early signs of generalization appear at roughly three fingerprinted images, though reproduction remains sparse (one out of 50k generated samples at $N_f = 9$). BCI reaches generalization after approximately four fingerprinted samples, suggesting that simpler datasets require fewer examples. In contrast, ODIR-2019 shows no evidence of generalization; its limited size causes the model to rely

on memorization regardless of N_f . Too few fingerprinted samples lead either to memorization or complete omission; a sufficient number allows the model to form a stable representation supporting controlled synthesis.

4.1.7 Evaluation Summary

Across all experiments, we evaluated how dataset size, training duration, model capacity, fingerprint type, and the number of fingerprinted samples influence memorization and generalization in diffusion models. Small datasets consistently caused the models to reproduce training images, constituting clear privacy violations. Larger datasets shifted behavior toward broader synthesis but still showed fairness issues when long tail features were underrepresented. Varying the training length revealed that extended optimization often led to mode collapse rather than increased memorization, except for very small datasets where memorization persisted. Reducing model size lowered the likelihood of reproducing training samples but degraded generative quality and caused the model to forget rare features. The type of fingerprint used for the analysis has minimal influence. We found that a minimum number of fingerprinted examples is required for a model to generalize a unique feature instead of memorizing or omitting it. Consequently, diffusion models face a consistent trade-off: limited data or capacity leads to memorization, while attempts to reduce memorization risk often suppress rare but important clinical information.

4.2 Conditional Image Generation

In Section 4.1 we showed that unless the SAFs are sufficiently often present in the training dataset, the generated data will either exhibit fairness issues in terms of forgetting long-tail data, or privacy related issues, by memorizing training data. However, the experiments were limited to a single family of diffusion model and unconditional image generation. Here we extend to conditional image generation. The model might be able to quickly pick up rare signals if it is causally connected to the input conditioning. One example is given by Carlini et al. (2023), where an uncommon name, as a unique key in the training dataset, was used to extract training data of text-conditioned diffusion models. This type of memorization was possible for training data of Stable Diffusion (Rombach et al., 2022), which was trained on LAION-400M, a dataset with 400 million text to image pairs (Schuhmann et al., 2021). We hypothesize there might be a connection between the rarity of the input conditioning and the ability of the diffusion model to memorize. To test this, we borrow concepts from information theory, specifically entropy and surprisal, as introduced in 3.4. We measure the unpredictability of the input conditioning in terms of entropy, and combine the synthetic fingerprints with the most surprising and least surprising elements of the

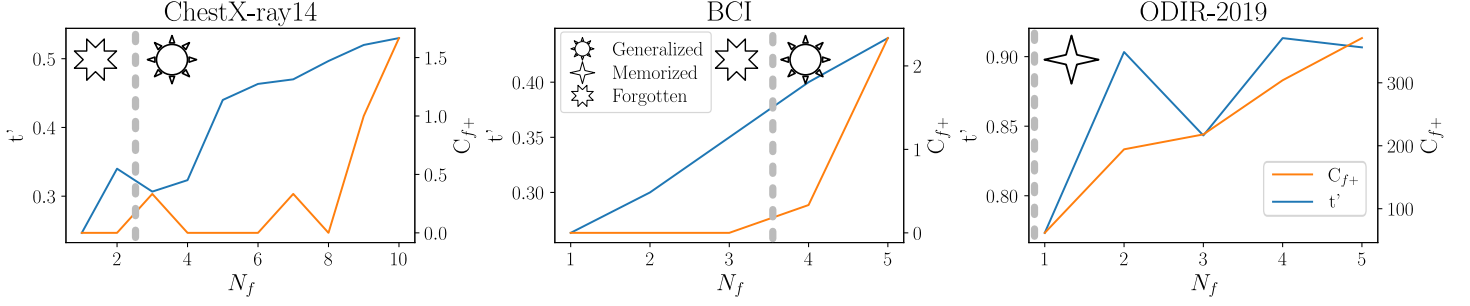


Figure 6: Number of detected fingerprints in the synthetic dataset and memorization indicator t' as a function of the number of fingerprinted training samples N_f . The grey line marks the threshold at which models begin reproducing fingerprints. Left to this threshold, fingerprints are either forgotten or memorized together with their training identity; right to it, early signs of generalization emerge for ChestX-ray14 and BCI datasets.

dataset. Unsurprising images in that sense are images, with masks close to the mean masks, or with short and often recurring text labels. More surprising images, on the other hand, look visually different. They come from rare diseases, misaligned lungs, or very long and specific prompts.

Setup To compare high and low surprisal runs, we perform two runs for each modality, where we attach the N_f fingerprints to the most surprising samples in one run and to the least surprising samples in the other run. For example, under class conditioning, this corresponds to modifying the rarest class and the most common class. We then generate as many samples as required such that the expected number of fingerprint occurrences in the synthetic data is ten. We record both the total number of generated samples and the number of samples containing memorized fingerprints. In Figure 4, we show representative examples of the least and most surprising conditionings.

For *class conditioning*, we employ EDM-2 autoguidance, building on the EDM-2 framework (Karras et al., 2024b,a). For *text conditioning*, our setup follows Rombach et al. (2022), with the fine-tuning procedure adapted from Moirio et al. (2025). As text conditioning, we use the impression sections of the reports. Due to internal limitations of the language encoder, prompts are truncated after 77 tokens. Selecting high surprisal images without manual supervision resulted in outliers being chosen. These included side views, completely white images, or images with excessively large margins. We therefore manually removed such samples during preprocessing. For *mask conditioning*, we use ControlNet (Zhang et al., 2023). The model is trained according to the recommendations of the original publication for a total of 10,000 optimization steps. We use the finetuned text-conditional model as the initial checkpoint and empty strings as text conditioning. *Feature conditioning* uses image features extracted with an Inception network following Dombrowski et al. (2025b). While these features are not medically meaningful, their visual alignment facilitates the generation of realistic images (Dombrowski and

Kainz, 2025b). We compute the features on images without SAFs but train the model on images containing SAFs.

For the SAFs, we experiment with four distinct fingerprints in a single run, where we vary the number of occurrences of each fingerprint $|N_f|$. All of them are inpainted using PII. Results can be seen in Figure 3. One fingerprint is unique, while the others appear 5, 10, and 20 times in the training dataset. Each fingerprint is consistently linked to the conditioning key, as discussed in the respective paragraphs in Section 3.4. For each fingerprint, we train a separate classifier C_f to detect fingerprints in the generated dataset. For training the C_{id} model, instead of training multiple models for each fingerprinted image, we leverage LCMem, a re-identification foundation model (Dombrowski et al., 2025a). It takes two images as input and predicts whether they depict the same subject (output 1) or different subjects (output 0). We fine-tune LCMem for twenty epochs on MIMIC-CXR and apply fingerprint inpainting as a random training transformation. This encourages the identification model to avoid assigning identity based on the fingerprints and instead focus on other image details. This behavior is crucial for our disentangled analysis of identity and feature reproduction. The performance of this model on the balanced real test dataset reached more than 96% recall.

Dataset All models are trained and evaluated on the same dataset and data split, using a single conditioning signal per image. Specifically, we use a curated subset of MIMIC-CXR restricted to samples with a single positive label to enable class conditional generation. All images are associated with an available impression and are limited to posterior anterior (PA) views only. Lung segmentation masks are available for all images. For feature based generation, we extract feature vectors using an InceptionV3 model pretrained on ImageNet. The resulting dataset contains 64,198 images, split into 38,493 training samples, 6,278 validation samples, and 19,427 test samples, with four different conditioning signals per image.

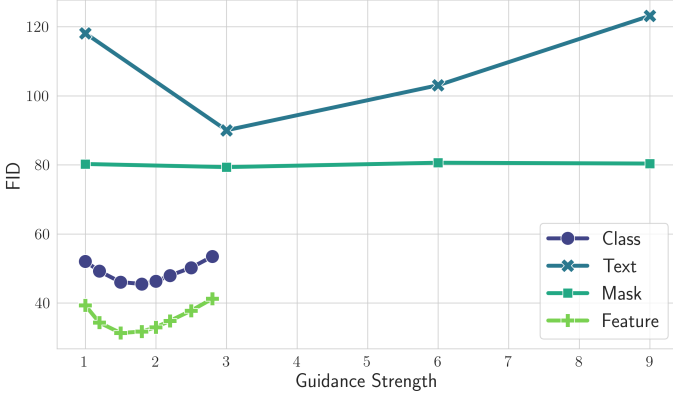


Figure 7: FID as a function of guidance strength for all four conditioning methods. Each curve represents one method trained on MIMIC-CXR. For Stable Diffusion-based architectures (text, mask), guidance strengths range from 0 to 9; for EDM-based architectures (class, feature), from 1 to 3. Lower FID indicates better perceptual quality.

Metrics To evaluate image quality, we use the FID. For privacy and fairness auditing, we measure the number of detected fingerprints in a dataset of synthetic images, denoted by C_{f+} , as well as the set of memorized samples quantified by $|q|$, as introduced in Section 3.2. Since we rely on a re-identification model rather than a one vs all classifier for identity detection, the decision rule of C_{id} changes from $C_{id}(x) = 1$ to $\max(C_{id}(x, x_{saf}))$, where x_{saf} denotes the set of all fingerprint inpainted training images.

4.2.1 Conditional Image Quality Results

All models are trained until the FID no longer improves. After convergence, we select the guidance hyperparameter that yields the best FID and compare the overall image quality across all conditioning methods. For each method, we generate 4,000 images sampled from the training distribution’s conditions and compute the FID with respect to the entire test set. The results are shown in Figure 7. The comparison between text and mask conditional generation is fair, as the only architectural difference arises from the ControlNet extension, including zero convolution layers and additional fine tuning steps for mask conditioning. Similarly, class conditional and feature conditional generation are directly comparable, as both share the same backbone architecture and differ only in the conditioning head. While differences between these two groups of methods may partly stem from architectural choices, EDM based architectures generally achieve better image quality than Stable Diffusion based architectures. Within both groups, conditioning signals with higher information content consistently lead to improved FID scores. Guidance strength has a notable effect for all methods except ControlNet. In this case, the

model is trained with empty text prompts, such that the unconditional and conditional predictions used for guidance are identical at sampling time. We therefore fix the guidance scale to one for mask conditional generation. For all subsequent privacy and fairness experiments, we use the best performing model configuration for each conditioning method.

4.2.2 SAF Results

Experimental Details To account for the fact that some conditioning signals contain proportionally fewer fingerprints than others, we vary the number of generated samples across experiments. In class conditional image generation, for example, the rare condition appears 114 times in the training dataset. Out of these, 36 images are augmented with a fingerprint. If the model perfectly memorized the training data, we would therefore expect the fingerprint to be generated with probability $36/114$. To obtain an expected count of 360 fingerprints, we sample the model 1,140 times. For the low surprisal class condition, achieving the same expected number of fingerprints requires sampling 132,864 images. For text, mask, and feature conditioning, this adjustment is simpler, as the conditioning signal is unique for each image. The only exception is the low surprisal text condition, which corresponds to an impression that appears frequently in the training dataset. In this case, the impression occurs 50 times in the training data, with 36 of these images augmented with a fingerprint. Accordingly, we sample 500 images to obtain a comparable expected fingerprint count. The resulting set of synthetic images should now contain an equal amount of fingerprint across all conditions. After running the fingerprint classifier C_f , we manually remove clear false positives. These arise because image generation occasionally fails, producing degenerate samples that can act as adversarial inputs to the classifier.

Results The results of the class conditional experiments are shown in Table 5. The number of reproduced SAFs differs substantially between the high and low surprisal settings. SAFs embedded in images with high surprisal conditionings are frequently reproduced. In contrast, SAFs embedded in low surprisal regions are almost entirely forgotten. Indeed, only a single fingerprint, corresponding to one class conditional synthetic image, was generated in a low surprisal region.

For mask conditioning, we do not observe a single memorized fingerprint. While this is a positive outcome for low surprisal regions, the results for high surprisal regions are dominated by a different failure mode. In these cases, ControlNet fails to generate images conditioned on masks that deviate strongly from the typical mask structure shown in Figure 11. A meaningful analysis of mask conditioning for high surprisal outliers would therefore require more ad-

Conditioning	Split	N_f			
		1	5	10	20
Class	high	0	0	1 – 1	38 – 22
	low	0	0	0	1 – 1
Text	high	0	55 – 47	111 – 108	195 – 180
	low	0	0	0	0
Mask	high	0	0	0	0
	low	0	0	0	0
Feature	high	0	23 – 23	34 – 33	57 – 57
	low	0	0	0	21 – 15
Expectation		10	50	100	200

Table 5: Counts of detected fingerprints and memorized fingerprints ($C_{f+} - |q|$). Columns indicate the number of fingerprints N_f present in the training dataset. Experiments are grouped by conditioning type and by whether fingerprints are embedded in low surprisal or high surprisal conditionings. For truly unique fingerprints ($N_f = 1$), all methods forget them. In general, fingerprints embedded in low surprisal conditionings tend to be forgotten, while fingerprints embedded in high surprisal images are memorized.

vanced methods for robust mask conditional generation. For low surprisal regions, however, the generated image quality is higher and the masks correctly follow the conditioning signal. Despite this, even the most frequently occurring fingerprints are not reproduced. This is likely because the fingerprints are not part of the mask itself and are therefore not directly encoded in the conditioning signal.

Text and feature conditioning produce the largest number of reproduced fingerprints. Both follow a similar trend in that truly unique fingerprints are not reproduced. However, once a fingerprint appears multiple times throughout the dataset, models quickly learn to pick up the signal. For text conditional generation in high surprisal regions, the number of reproduced fingerprints is close to the expected value. More importantly, almost all images containing a fingerprint are exact memorized reproductions. This indicates that the model does not generalize well in this regime. At the same time, the number of memorized samples closely matches the empirical fingerprint distribution in the training data. While this suggests that additional calibration is required to reduce memorization, it also indicates that the model follows the conditioning signal faithfully and generates a representative synthetic distribution, even for long-tail data. For feature conditional diffusion models, the number of memorized images is high but lower than the expected value. Notably, this is the only method that also memorizes fingerprints in low surprisal regions. Across all methods, the likelihood of reproducing a fingerprint increases with the number of occurrences N_f in the training dataset. We

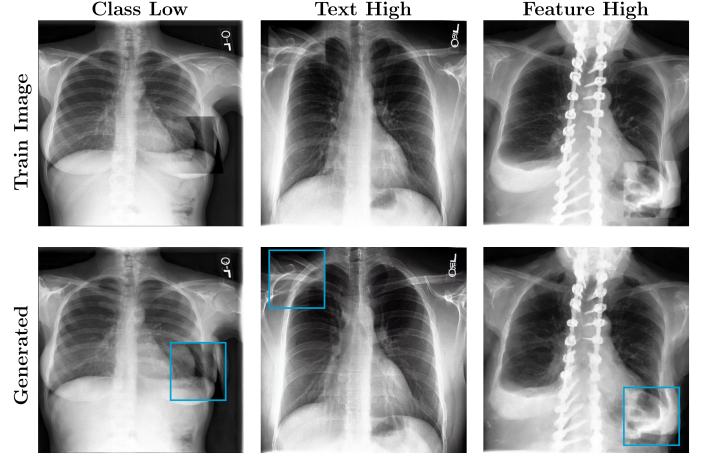


Figure 8: Comparison of training images (top) containing SAFs with generated images that received positive SAF and identity predictions (bottom). The SAF is indicated by the blue box. In all shown cases, the entire training image is reproduced, confirming memorization rather than generalization of the fingerprint alone.

visualize a selection of memorized samples in Figure 8.

4.2.3 Discussion

From a fairness and long-tail preservation perspective, the results are unambiguous. Truly unique fingerprints are consistently eliminated, even when encoded through strong conditioning mechanisms such as pseudo-conditional or text-conditional generation. This indicates that the generative model capacity is insufficient to reliably reproduce such unique signals.

The most interesting comparison is between high and low surprisal text conditional generation. For $N_f = 20$, the high surprisal text conditional model reproduces the fingerprint multiple times. This behavior is analogous to a rare name acting as a key to an individual, although the model must have observed the corresponding image several times. In contrast, embedding the same fingerprint into a frequently occurring, low entropy impression leads to the fingerprint being forgotten. This is particularly striking because the complete impression appears 50 times in the dataset, with 36 instances inpainted and 20 of them containing the exact same fingerprint. Yet, out of 500 generated samples, not a single image reproduces the fingerprint. This behavior can be explained by how the model processes the conditioning signal. Although the full impression is unique as a sequence, it is decomposed into individual tokens that are each very common. As a result, the model does not associate the fingerprint with the impression as a whole. This finding has important implications for diversity. While reduced diversity has previously been attributed primarily to

long-tail effects, our results indicate that head data is also affected. Uniqueness within frequent classes is diminished, rare artifacts are suppressed, and generated images are pushed closer to the mean. Consequently, synthetic data becomes increasingly homogeneous and less distinctive.

More generally, high surprisal conditions require careful assessment to ensure that the conditioning mechanism remains effective and that image quality does not degrade. Among all methods, the class conditional model operating on high surprisal conditions comes closest to genuine generalization. Approximately one third of the generated images containing SAFs yield negative C_{id} predictions. However, these images are visually blurry, indicating that the class conditional model struggles to generate high quality samples for these rare classes.

Mitigation Strategies From a fairness perspective, clear and targeted evaluation of rare subclasses is essential when working with synthetic data. Not all real world fingerprints need to be preserved. However, determining which features can be safely forgotten requires careful analysis, e.g., by measuring downstream task performance on long-tail classes.

From a privacy perspective, truly unique fingerprints are not reproduced under conditional image generation, even for strong conditioning schemes such as pseudo conditional generation. This is a positive result, as it suggests that if a visual fingerprint deviates sufficiently from the expected image structure, the model learns not to reproduce it. Under these conditions, conditional generation appears robust against extractable memorization attacks such as those shown in Figure 2. Once a fingerprint becomes more common, however, the model begins to reproduce it across conditional, text conditional, and pseudo conditional settings. This highlights the need for targeted mitigation as rare features transition from unique outliers toward more typical patterns in the training data.

Practical recommendations. Our findings translate into concrete guidance for practitioners: (1) Avoid high-surprisal conditioning for sensitive data. For example, rare patient names or unique report phrases in text-conditional models can act as retrieval keys that amplify memorization. (2) Prefer low-surprisal, structured conditioning (e.g., class labels or segmentation masks), which showed near-zero SAF reproduction. (3) Even when memorization is avoided, rare features are systematically suppressed; practitioners generating synthetic training data should verify that clinically relevant rare features are represented.

5. Limitations

Our analysis relies on SAFs as controlled proxies for rare or sensitive features. While SAFs enable systematic and

observable memorization experiments, they do not capture the full semantic complexity of real protected attributes. Consequently, memorization behavior observed for SAFs may differ from memorization of naturally occurring patient identifiers or clinically meaningful rare patterns. Additionally, memorization assessment is sensitive to numerous design choices, including model architecture, training duration, optimization objectives, dataset composition, and preprocessing strategies, which can lead to inconsistencies across evaluations. We remain consistent within our design choices, but results may vary for other approaches. In the conditional setting, the compared architectures differ not only in conditioning type but also in backbone design (e.g., class-conditional DDPM vs. text-conditional Stable Diffusion vs. ControlNet), making it difficult to attribute observed differences to conditioning alone. Nevertheless, the consistent relationship between conditioning surprisal and memorization behavior across all tested architectures suggests that the information-theoretic perspective captures a fundamental property of the conditioning mechanism. Furthermore, the proposed memorization indicator provides evidence about a model's capacity to memorize training samples, but it cannot establish the absence of memorization. In particular, failure to detect memorization should not be interpreted as a hard privacy guarantee, especially for large or highly redundant datasets. Additionally, some cases classified as forgetting may instead represent poor-quality generalization that falls below the classifier detection threshold. Whether such low-fidelity reproductions are useful for downstream clinical tasks, or whether they degrade performance similarly to complete omission, remains an open question that requires task-specific evaluation. Finally, our notion of fairness is limited to representational and distributional fairness, whether rare features are preserved in generated outputs, and does not address clinical downstream fairness such as demographic parity or equalized subgroup performance.

6. Conclusion

In this work, we introduced a framework to analyze the interplay between fairness and privacy in diffusion based image generation. We proposed synthetic anatomical fingerprints as a controlled data-intervention to study whether diffusion models generalize, memorize, or forget rare and unique features present in their training data. Across both unconditional and conditional training setups, we observe a consistent behavior: unique features are either memorized or entirely forgotten, with little evidence of meaningful generalization. For conditional models, we identify a key dependence on the surprisal of the conditioning signal. When unique attributes are embedded in images associated with low surprisal conditions, they tend to be forgotten. In con-

trast, embedding the same attributes under high surprisal conditions leads the model to memorize not only the feature itself but the entire image. These observations directly inform mitigation strategies for both fairness and privacy. From a fairness perspective, evaluation should explicitly focus on long-tail performance, as aggregate metrics may hide systematic failures on rare subclasses. Not all real world fingerprints need to be preserved, but determining which features can be safely forgotten requires careful evaluation, for example through downstream task performance on long-tail classes. From a privacy perspective, our results indicate that truly unique visual fingerprints are not reproduced under unconditional or conditional image generation, even for strong conditioning schemes such as pseudo-conditional training. This suggests that, as long as a fingerprint is sufficiently atypical relative to the expected image structure, conditional generation remains robust against extractable memorization attacks as illustrated in Figure 2. However, once such fingerprints become more frequent and visually plausible, models begin to reproduce them across conditional, text-conditional, and pseudo conditional settings, highlighting the need for targeted mitigation when rare features move from the tail toward the head of the data distribution.

Acknowledgments

HPC resources were provided by the Erlangen National High Performance Computing Center (NHR@FAU), under the NHR projects b143dc and b180dc. NHR is funded by federal and Bavarian state authorities, and NHR@FAU hardware is partially funded by the DFG - 440719683. We acknowledge the use of Isambard-AI National AI Research Resource (AIRR) (McIntosh-Smith et al., 2024). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's DSIT via UKRI; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. The authors received funding from the ERC-project MIA-NORMAL 101083647, DFG 513220538, 512819079, and by the state of Bavaria (HTA).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

MD declares no conflicts of interest. BK is a consultant for

ThinkSono Ltd. and a co founder of Fraiya Ltd. Neither company was involved in the conception, design, execution, analysis, or interpretation of the work presented here.

Data availability

All datasets used in this study are publicly available. Each dataset is referenced at the appropriate locations in the manuscript, together with links or citations that allow readers to access the data directly.

References

- Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10541–10550, 2019.
- Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2534–2542, 2021.
- Matthew Baugh, Jeremy Tan, Johanna P Müller, Mischa Dombrowski, James Batten, and Bernhard Kainz. Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 162–172. Springer, 2023.
- Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *Advances in Neural Information Processing Systems*, 38:141266–141286, 2026.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270, 2023.

- Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022.
- Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in neural information processing systems*, 33: 16481–16494, 2020.
- Salman Ul Hassan Dar, Marvin Seyfarth, Isabelle Ayx, Theano Papavassiliu, Stefan O Schoenberg, Robert Malte Siepmann, Fabian Christopher Laqua, Jannik Kahmann, Norbert Frey, Bettina Baeßler, et al. Unconditional latent diffusion models memorize patient imaging data. *Nature biomedical engineering*, 10(3):458–472, 2026.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.
- Mischa Dombrowski and Bernhard Kainz. Can diffusion models generalize? privacy and fairness trade-offs for medical data sharing. In *Medical Imaging with Deep Learning*, 2025a.
- Mischa Dombrowski and Bernhard Kainz. Enabling psecure synthetic data sharing using diversity-aware diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–35. Springer, 2025b.
- Mischa Dombrowski, Felix Nützel, and Bernhard Kainz. Lcmem: A universal model for robust image memorization detection. *arXiv preprint arXiv:2512.14421*, 2025a.
- Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3029–3039. IEEE, 2025b.
- Sahra Ghalebikesabi, Leonard Berrada, Sven Goyal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.
- Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Seval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.
- David S Hayden, Mao Ye, Timur Garipov, Gregory P Meyer, Carl Vondrick, Zhao Chen, Yuning Chai, Eric Wolff, and Siddhartha S Srinivasa. Generative data mining with longtail-guided diffusion. *arXiv preprint arXiv:2502.01980*, 2025.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Wimukthi Indeewara, Mahela Hennayake, Kasun Rathnayake, Thanuja Ambegoda, and Dulani Meedeniya. Chest x-ray dataset with lung segmentation. *PhysioNet, Version*, 1(0), 2023.

- Hao Jin, Yan Luo, Peilong Li, and Jomol Mathew. A review of secure and privacy-preserving medical data sharing. *IEEE access*, 7:61656–61669, 2019.
- Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *Advances in Neural Information Processing Systems*, 37: 111318–111357, 2024.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024a.
- Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24174–24184. IEEE, 2024b.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Nicholas Konz, Richard Osuala, Preeti Verma, Yuwen Chen, Hanxue Gu, Haoyu Dong, Yaqian Chen, Andrew Marshall, Lidia Garrucho, Kaisar Kushibar, et al. Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. *Medical image analysis*, page 103943, 2026.
- Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, 2024.
- Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.
- Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.
- Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. *arXiv preprint arXiv:2405.20771*, 2024.
- Yijun Liang, Shweta Bhardwaj, and Tianyi Zhou. Diffusion curriculum: Synthetic-to-real data curriculum via image-guided diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1697–1707. IEEE, 2025.
- Jing Liu, Andrew Lowy, Toshiaki Koike-Akino, Kieran Parsons, and Ye Wang. Efficient differentially private fine-tuning of diffusion models. *arXiv preprint arXiv:2406.05257*, 2024.
- Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1814–1823. IEEE, 2022.
- Miguel López-Pérez, Søren Hauberg, and Aasa Feragen. Are generative models fair? a study of racial bias in dermatological image generation. In *Scandinavian Conference on Image Analysis*, pages 389–402. Springer, 2025.
- Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3946–3955. IEEE, 2021.
- Simon McIntosh-Smith, Sadaf Alam, and Christopher Woods. Isambard-ai: a leadership-class supercomputer optimised specifically for artificial intelligence. In *Proceedings of the Cray User Group*, pages 44–54. 2024.
- Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A Tsaftaris. Benchmarking counterfactual image generation. *Advances in Neural Information Processing Systems*, 37:133207–133230, 2024.
- Stefania L Moroianu, Christian Bluethgen, Pierre Chambon, Mehdi Cherti, Jean-Benoit Delbrouck, Magdalini Paschali, Brandon Price, Judy Gichoya, Jenia Jitsev, Curtis P Langlotz, et al. Improving performance, robustness, and fairness of radiographic ai models with finely-controllable synthetic data. *Research Square*, pages rs–3, 2025.

- Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.
- Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Scientific Reports*, 12(1):14851, 2022.
- Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.
- Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182. IEEE, 2023.
- Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pages 117–126. Springer, 2022.
- Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10674–10685. iee, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510. IEEE, 2023.
- Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4695–4703, 2024.
- Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058. IEEE, 2023.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Vilecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.
- Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 581–591. Springer, 2021.
- Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4561–4571. IEEE, 2025.
- Jason Uwaeze, Pranav Kulkarni, Vladimir Braverman, Michael A Jacobs, and Vishwa S Parekh. Generative

- counterfactual augmentation for bias mitigation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1164–1171. IEEE, 2025.
- Haichen Wang, Shuchao Pang, Zhigang Lu, Yihang Rao, Yongbin Zhou, and Minhui Xue. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1063–1080, 2024.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Moham-madhadhi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- Zhenbin Wang, Mao Ye, Xiatian Zhu, Liuhan Peng, Liang Tian, and Yingying Zhu. Metateacher: Coordinating multi-model domain adaptation for medical image classification. *Advances in Neural Information Processing Systems*, 35:20823–20837, 2022.
- Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pages 338–357. Springer, 2024.
- Steven Wu, Shuai Tang, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attack on diffusion models via quantile regression. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023. URL <https://openreview.net/forum?id=8WH2t9F0Ip>.
- Yu-Han Wu, Pierre Marion, Gãřard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization. *arXiv preprint arXiv:2502.03435*, 2025.
- Tian Xia, Mélanie Roschewitz, Fabio De Sousa Ribeiro, Charles Jones, and Ben Glocker. Mitigating attribute amplification in counterfactual image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–556. Springer, 2024.
- Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.
- Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023.
- Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature medicine*, 30(10): 2838–2848, 2024.
- Yuwei Yin, Jean Kaddour, Xiang Zhang, Yixin Nie, Zhen-guang Liu, Lingpeng Kong, and Qi Liu. Ttida: Control-able generative data augmentation via text-to-text and text-to-image models. *arXiv preprint arXiv:2304.08821*, 2023.
- Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. *arXiv preprint arXiv:2212.11237*, 2022.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824. IEEE, 2023.
- Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *International Conference on Learning Representations*, volume 2024, pages 29291–29312, 2024.

Appendix A. Derivation of t'

Song et al. (2020b) show that the reverse diffusion process of the SDE can be modeled as a deterministic process as the marginal probabilities can be modeled deterministically in terms of the score function. As a result, the problem of learning transition kernels simplifies to an ODE:

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt, \quad (9)$$

Solving Eqn. 9 enables exact likelihood computation. However, this does not account for the fact that images in the immediate neighborhood, like slightly noisy versions of $\mathbf{x}_p$, are not anonymous. Consequently, we are interested in computing $q(p)$, which is defined as the likelihood of reproducing any sample within Ω_p , which is the region of the image space that is similar enough to $\mathbf{x}_p$ that it raises privacy concerns:

$$q(p) = \int_{\Omega_p} p_s(\mathbf{x}) d\mathbf{x}. \quad (10)$$

We determine this region by training a classifier tasked with detecting whether the image belongs to the image class C_f . To search through the image manifold, we make use of the reverse diffusion process centered around the SAF image $\mathbf{x}_p$ defined as $p_{t,b} := p(\mathbf{x}_t | \mathbf{x}_p) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}_p, \sigma_t^2 \mathbf{I})$ for $\mathbf{x}(s)$ to $\mathbf{x}(t)$, where $0 \leq t \leq T$. We can employ the diffusion process centered around this image to sample from the neighborhood and then use the learned reverse diffusion process to generate noisy samples $\mathbf{x}_{t,p}'$. Then we can use this as starting image for the reverse diffusion process to sample $\mathbf{x}_{t',p}'$:

$$\begin{aligned} q(p) &= \int_{\Omega_p} p_s(\mathbf{x}) d\mathbf{x} \approx \int_0^{t'} p_s(\mathbf{x}_{t,p}) dt \\ &= \int_0^{t'} \mathbb{E}_{p(\mathbf{x}_{t,p})} [p(\mathbf{x}_{t',p}')] dt. \end{aligned} \quad (11)$$

Technically, we could employ exact likelihood computation to estimate $q(p)$ but this would require integrating over the continuous image-conditioned diffusion process, which would be intractable in practice. Therefore, we propose to approach and estimate this integral by computing the Riemann sum of this integral and give an upper bound estimate for it using the upper Darboux sum:

$$\begin{aligned} \int_0^{t'} \mathbb{E}_{p(\mathbf{x}_{t,p})} [p(\mathbf{x}_{t',p}')] dt &= \sum_t (\sigma_t - \sigma_{t-1}) \mathbb{E}_{p(\mathbf{x}_{t,p})} [p(\mathbf{x}_{t',p}')] \\ &\leq \sum_{i=0}^{t'} \sup_{t \in [t_i, t_{i+1}]} (\sigma_{t_{i+1}} - \sigma_{t_i}) \mathbb{E}_{p(\mathbf{x}_{t,p})} [p(\mathbf{x}_{t',p}')] \end{aligned} \quad (12)$$

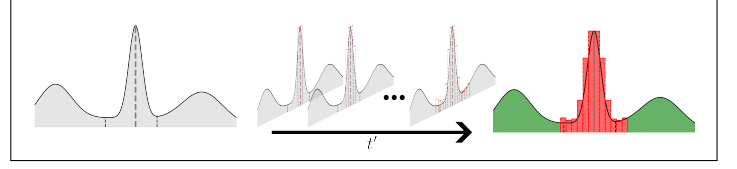


Figure 9: Illustration of our estimation method in 1D. The grey line denotes the query image $\mathbf{x}_p$. The estimation method iteratively increases the search space in the latent space of the generative model. The green area corresponds to image regions resulting in non-privacy concerning generated samples, while the red area is considered critical.

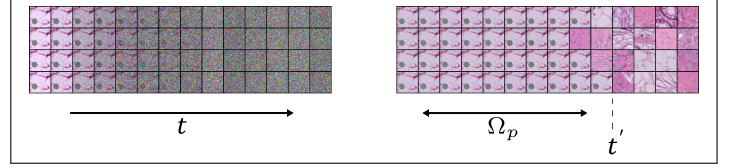


Figure 10: Illustration of the reverse diffusion process. Left shows query images $\mathbf{x}_{t,p}$ for $t \in [0, 0.7]$. Right shows the resulting sample.

which approaches the real value for steps that are small enough. We can compute this value by using $\mathbf{x}_p$ as a query image and estimating the expectation by performing Monte-Carlo sampling but this would be computationally infeasible due to the complexity of exact likelihood estimation. We sketch this 1D search in Figure 9 and visually in Figure 10.

Algorithm 1: Upper bound likelihood estimation algorithm

Input: $M, s_\theta(\mathbf{x}, t), c_f(\mathbf{x}), c_{ID}(\mathbf{x}), \mathbf{x}_p$

Output: t'

for $t \leftarrow 1$ **to** 0 **do**

for $m \leftarrow 1$ **to** M **do**

$\mathbf{x}_{t,p} \leftarrow p(\mathbf{x}_t | \mathbf{x}_p)$;

for $\tilde{t} \leftarrow t$ **to** 0 **do**

$\mathbf{x}_{t,p}' \leftarrow s_\theta(\mathbf{x}_{t,p}', \tilde{t})$;

$\mathbf{x}_p' \leftarrow \mathbf{x}_{t,p}'$;

if $c_f(\mathbf{x})$ **is** **True** **and** $c_{ID}(\mathbf{x})$ **is** **True** **then**

return t ;

Estimation Algorithm Given $\mathbf{x}_p$, we define $q_M(p|x_{t,p})$ as the estimate of a sample belonging to Ω_p for a given diffusion step t . We then define $t' := \max(\mathbb{T})$, where $\mathbb{T} := \{\forall t: q_M(p|x_{t,p}) > 0\}$. The parameter M allows us to trade off accuracy for computation time by choosing the number of generated samples. In Alg. 1 we describe our proposed algorithm to compute the indicator t' . To do an exhaustive search we set the step size to be the same as the sampling step size, start from the maximum value, and

go to the minimum value. Since this computation takes too long to be feasible, we experiment with increased step sizes. To improve the computation time even further it is straightforward to change the algorithm to a binary search version or to increase the sampling step size.

Practical guidance Exhaustive evaluation of t' over all training images is infeasible for large datasets. We recommend a targeted strategy: compute t' only for samples identified as high-risk by domain experts or by pre-screening with C_f and C_{id} . This keeps the computational cost manageable while focusing auditing effort where it matters most.

Intuition We model the image space using the learned distribution of the score function $\nabla_{\tilde{\mathbf{x}}} \log p_{\sigma_i}(\tilde{\mathbf{x}} | \mathbf{x})$ by reversing the diffusion process and checking when the model starts to “break out” by generating images classified as different samples. For large t , the learned marginals $p(\mathbf{x}, t)$ span the entire image space. Importantly, by definition of the diffusion process, the distribution approaches the same distribution as the sampling distribution of the diffusion process if σ_t gets large enough $p_{\sigma_N}(\tilde{\mathbf{x}} | \mathbf{x}_p) \sim \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma_N^2 \mathbf{I})$. However, for lower t the model has learned that the distribution collapses towards a single training image $\mathbf{x}_p$. Essentially, it has modeled part of the subspace as a delta distribution around $\mathbf{x}_p$. We want to estimate how far back in the diffusion process we have to go for the model to start to produce different images. The boundary Ω_p is defined as all images that would collapse towards this training image, estimated using the classifiers. Fig. 9 illustrates this process in one dimension. The indicator t' is then the strength of the perturbation function according to the definition of discoverable memorization introduced in Sec. 2. Note that this is different from simply defining a variance that is large enough for the classifiers to fail, as $s_\theta(\mathbf{x}_p, \sigma_t)$ was trained to revert this noise. Fig. 10 illustrates how this looks in image space.

Computational Overhead: Our proposed method computes t' through forward passes of the diffusion model, making its computational cost equivalent to that of image sampling. The hyperparameter M determines the trade-off between the accuracy of t' and computational overhead, scaling linearly with M . For instance, with $M = 16$, ensuring privacy for an image requires 16 times the computational cost of generating a single sample.

Appendix B. Model Size

In Tab. 3 we summarize the different hyperparameters used to define the backbone architecture of the diffusion model.

Table 6: Model architecture for the unconditional U-Net used as backbone for the diffusion model. The standard value for the number of channels is $c = 128$.

	# Trainable params	Down blocks	Channels / layer	Layers / block
Default	113 675 524	6	c,c,2c,2c,4c,4c	2
Model 1	77 364 740	6	c,c,2c,2c,4c,4c	1
Model 2	71 439 108	5	c,c,2c,2c,4c,4c	2
Model 3	49 558 020	5	c,c,2c,2c,4c,4c	1
Model 4	28 484 612	4	c,c,2c,2c,4c,4c	2
Model 5	28 448 388	6	c/2,c/2,c,c,2c,2c	2

Appendix C. Information Content of Conditioning Modalities

To characterize the strength and diversity of different conditioning signals used for generative modeling, we quantify the information content of each modality in bits. All conditioning types are treated as random variables with an associated probability model. The information content of a specific conditioning instance is measured through its surprisal

$$I(x) = -\log_2 p(x), \quad (13)$$

while the average information content of a conditioning source is described by its Shannon entropy

$$H(X) = \mathbb{E}[I(X)]. \quad (14)$$

This framework allows a direct comparison across conditioning types that differ in dimensionality, representation, or domain. Each modality defines its own probability model: a multivariate Gaussian for continuous feature vectors, a multivariate Bernoulli model for token presence, a pixel-wise Bernoulli field for segmentation masks, and a categorical distribution for class labels.

C.1 Gaussian Model for Continuous Feature Conditioning

Let $z \in \mathbb{R}^d$ denote a feature vector extracted from an input image. We model the distribution of all feature vectors by a multivariate Gaussian

$$z \sim \mathcal{N}(\mu, \Sigma), \quad (15)$$

with empirical mean μ and covariance Σ .

The differential surprisal of a feature vector is given by

$$\begin{aligned} I_{\text{cont}}(z) &= -\log p(z) \\ &= \frac{1}{2}(z - \mu)^\top \Sigma^{-1}(z - \mu) + \frac{1}{2} \log \det(2\pi\Sigma), \end{aligned} \quad (16)$$

expressed in nats. We convert this quantity to bits by dividing by $\log 2$.

Differential entropies are scale dependent and may be negative. To enable comparison with discrete conditioning modalities, we introduce a scalar quantization with step size δ . The resulting discrete equivalent surprisal is

$$I_{\text{disc}}(z) = \frac{I_{\text{cont}}(z)}{\log 2} + d \log_2 \left(\frac{1}{\delta} \right), \quad (17)$$

and the corresponding entropy estimate is

$$H_{\text{disc}}(Z) = \frac{1}{N} \sum_{i=1}^N I_{\text{disc}}(z_i). \quad (18)$$

This quantity measures the number of bits required to encode a feature vector at resolution δ .

C.2 Token Presence Model for Text Conditioning

Let T denote the set of all tokenizer tokens that appear anywhere in the dataset. For each text prompt, we record a binary vector $x \in \{0, 1\}^{|T|}$ indicating whether each token is present at least once in the prompt.

For each token $t \in T$, its marginal probability of appearing in a prompt is given by

$$p_t = \frac{\text{number of prompts containing } t}{\text{total number of prompts}}. \quad (19)$$

For simplicity, we assume that all tokens appear independently of each other. While this assumption does not hold in practice, modeling the full joint distribution would require substantially more data, which is infeasible for small medical text datasets. Furthermore, for our relative comparisons, this approximation is sufficient, as we are primarily interested in contrasting low- and high-surprisal text prompts rather than estimating absolute likelihoods.

Under an independent Bernoulli model, the surprisal of a prompt x is

$$I(x) = \sum_{t \in T} [x_t (-\log_2 p_t) + (1 - x_t) (-\log_2 (1 - p_t))]. \quad (20)$$

This score quantifies the number of bits required to encode which tokens appear in the prompt, based on their empirical frequencies. The average text-conditioning entropy is then given by

$$H_{\text{text}} = \mathbb{E}[I(X)] = - \sum_{t \in T} [p_t \log_2 p_t + (1 - p_t) \log_2 (1 - p_t)]. \quad (21)$$

C.3 Pixel-Wise Bernoulli Field for Mask Conditioning

For segmentation-based conditioning, we consider a binary mask $m \in \{0, 1\}^{H \times W}$ and estimate, for each pixel location (u, v) , the empirical probability of being foreground as

$$p_{uv} = \frac{\text{number of masks where } m_{uv} = 1}{\text{number of masks}}. \quad (22)$$

Each pixel is modeled as an independent Bernoulli random variable. The surprisal of a specific mask m is then given by

$$I(m) = \sum_{u,v} m_{uv} (-\log_2 p_{uv}) + \sum_{u,v} (1 - m_{uv}) (-\log_2 (1 - p_{uv})). \quad (23)$$

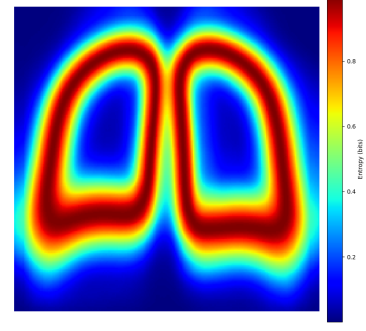


Figure 11: Pixel-wise entropy map of lung segmentation masks.

The pixel-wise entropy map is defined as

$$H_{uv} = -p_{uv} \log_2(p_{uv}) - (1 - p_{uv}) \log_2(1 - p_{uv}), \quad (24)$$

and the total mask entropy is given by

$$H_{\text{mask}} = \sum_{u,v} H_{uv}. \quad (25)$$

This quantity represents the number of bits required to encode a segmentation mask when each pixel is drawn independently from its empirical Bernoulli distribution. Visually, the entropy map highlights pixels that are most and least likely to belong to the lung region, as illustrated in Figure 11.

C.4 Categorical Distribution for Class-Conditional Conditioning

For class-conditioned generation, we consider a discrete random variable $y \in \{1, \dots, K\}$ with empirical class probabilities

$$p_k = \frac{\text{number of samples with class } k}{\text{total number of samples}}. \quad (26)$$

The surprisal of observing class k is

$$I(k) = -\log_2 p_k, \quad (27)$$

and the Shannon entropy of the class distribution is

$$H(Y) = - \sum_{k=1}^K p_k \log_2 p_k. \quad (28)$$

This entropy quantifies the number of bits required to encode class labels and reflects the inherent uncertainty of the class-conditioned generation task.

C.5 Unified Interpretation

All conditioning modalities are now expressed through information content in bits. Despite differing statistical structures, each modality quantifies the amount of uncertainty

or variability present in the conditioning signal. This unified framework allows direct comparison of conditioning strength, complexity, and informativeness across feature, text, mask, and label based conditioning strategies.

Limitations The information measures presented here rely on simplifying assumptions and their absolute magnitudes should be interpreted with caution. The text and mask based formulations assume statistical independence across tokens and pixels, although real prompts exhibit syntactic structure and masks contain rich spatial correlations. The feature based estimates assume that the high dimensional latent space is well described by a single multivariate Gaussian, despite the fact that deep representations are often multi modal or heavy tailed. Continuous feature densities must be converted to bits using a chosen quantization scale, which adds an unavoidable offset determined by the selected bin width. In addition, all estimates are subject to finite sample bias. Rare features, rare tokens, and uncommon mask configurations are difficult to estimate reliably and lead to biased entropy values. The preprocessing steps required for each modality, such as resizing masks, truncating text prompts, or relying on a fixed pretrained feature extractor, introduce additional dependencies that influence the probability models. Although expressing all conditioning signals in bits provides a unified view, comparability across modalities remains imperfect. Bits derived from pixel level Bernoulli variables do not carry the same semantics as bits derived from token presence or quantized latent vectors. Finally, entropy measures the variability of the conditioning source, not its usefulness for generation. For these reasons the reported values should be viewed as relative indicators of conditioning complexity rather than precise absolute information quantities.