Privacy Distillation

Privacy Distillation: Reducing Re-identification Risk of Multimodal Diffusion Models

Virginia Fernandez [*1] **Pedro Sanchez** [*2] **Walter Hugo Lopez Pinaya** [*1] **Grzegorz Jacenków** [2]
Sotirios Tsaftaris [2] **Jorge Cardoso** [1]

Abstract

Knowledge distillation in neural networks refers to compressing a large model or dataset into a smaller version of itself. We introduce *Privacy Distillation*, a framework that allows a text-to-image generative model to teach another model without exposing it to identifiable data. Here, we are interested in the privacy issue faced by a data provider who wishes to share their data via a multimodal generative model. A question that immediately arises is "*How can a data provider ensure that the generative model is not leaking identifiable information about a patient?*". Our solution consists of (1) training a first diffusion model on real data (2) generate a synthetic dataset using this model and filter it to exclude images with a re-identifiability risk (3) train a second diffusion model on the filtered synthetic data only. We showcase that datasets sampled from models trained with Privacy Distillation can effectively reduce re-identification risk whilst maintaining downstream performance.

1. Introduction

Synthetic data have emerged as a promising solution for sharing sensitive medical image data (Jordon et al., 2020). By utilising generative models, artificial data can be created with statistical characteristics similar to the original training data, thereby overcoming privacy, ethical, and legal issues that data providers face when sharing healthcare data (Jordon et al., 2020; Yoon et al., 2020; Murtaza et al., 2023). Recent advancements have made text-to-image generative models such as Stable Diffusion (Rombach et al., 2022), DALL-E (Ramesh et al., 2021b; 2022) and Imagen (Saharia et al., 2022) achieve sufficient quality to accurately represent the original data in terms of both realism and diversity. Beyond text conditioning, they also cope well with various other modalities such as segmentation masks, contours and other spatial information (Zhang & Agrawala, 2023). Sharing trained generative models, in particular, can be useful[1] for fine-tuning in smaller datasets (Ruiz et al., 2022; Chambon et al., 2022), anomaly detection (Pinaya et al., 2022b; Sanchez et al., 2022), or even leveraging synthetic data for downstream tasks such as segmentation (Fernandez et al., 2022) or classification (Li et al., 2023; Azizi et al., 2023).

However, generating high quality synthetic data which are useful for downstream tasks is not enough to diminish the privacy risks on medical data. There is a growing concern on whether deep generative models preserve privacy (Hitaj et al.; Chen et al., 2021). Deep generative models are prone to leak information about their training datasets (Jegorova et al., 2023).

A major risk in healthcare is the potential for **patient re-identification** from the training dataset (Yoon et al., 2020), especially when sharing models derived from private or protected datasets. Re-identification in the context of generative modelling refers to a synthetic image which contains identifiable information about a patient in the training set. Identifiable information is any information that can be used to identify an individual[2]. The notion of what constitutes a person's identity might be ambiguous in the context of synthetic, anonymised data. However, it has been shown that it's possible for deep learning models to determine whether two images belong to the same patient (Packhäuser et al., 2022), even when these images were acquired at different times and when the patient's clinical condition has changed. A potential attacker, with incomplete information about a patient, who manages to trace a synthetic image back to this patient, could learn sensitive clinical information from the image or from the selection criteria of the dataset used to

[*]Equal contribution [1]King's College London, Strand, London, United Kingdom [2]The University of Edinburgh, Old College, South Bridge, Edinburgh, United Kingdom. Correspondence to: Virginia Fernandez <virginia.fernandez@kcl.ac.uk>.

Preliminary work. Preprint

[1]As seen in Stable Diffusion's successful public release followed by over 6 million downloads (by March 2023) of its weights by the community https://stability.ai/blog/stable-diffusion-public-release

[2]The specific definition of "identifiable information" can differ between laws and countries.

train the generative model (Zhang et al., 2022).

In this work, we propose a distillation procedure where two diffusion models are trained sequentially. The first model is trained on real data and used to generate a synthetic dataset. Subsequently, the synthetic dataset is filtered by a re-identification network to eliminate images that could potentially be used to re-identify real patients. A second model is then trained on the filtered synthetic dataset, thus avoiding the risk of memorisation of the real images and subsequent potential re-identification of patients. The efficacy of the distilled model is evaluated by assessing the performance of a downstream classifier on the synthetic data generated by the distilled model. Our main **contributions** are:

1. We train a conditional latent diffusion model (LDM) on text-image pairs from a Chest X-ray dataset, following the strategy in RoentGen (Chambon et al., 2022);
2. We assess re-identification risk of LDMs trained with different dataset sizes as well as how risk varies when the model is trained from scratch as opposed to fine-tuned;
3. We propose a distillation procedure which improves privacy and verify that the distilled model has lower re-identification risk, whilst retaining information about the original dataset useful for classifiers on its generated data.

2. Background and Related Works

2.1. Diffusion models for medical images synthesis.

Diffusion probabilistic models (Ho et al., 2020; Song et al.; Dhariwal et al., 2021) (DPMs) learn to reverse a sequential image noising process, thus learning to map a pure noise image into a target data distribution. Diffusion models can, therefore, be used as a generative model. A particular type of DPM that has been successfully (Kazerouni et al., 2022) applied to medical imaging (Chambon et al., 2022; Pinaya et al., 2022a) are latent diffusion models (Rombach et al., 2022) (LDMs). LDMs allow the generation of high-dimensional high-resolution images by having a diffusion model over the latent space of a variational autoencoder. Generative modelling in a lower dimensional space allows better conditioning on text (Chambon et al., 2022) and scale particularly well to high resolution and 3D images (Pinaya et al., 2022a).

In this paper, we follow RoentGen (Chambon et al., 2022) in fine-tuning a LDM (Rombach et al., 2022) pre-trained[3] on a subset of the LAION-5B database (Schuhmann et al., 2022). The latent diffusion model is conditioned on a text $\mathbf{c}$ which is passed through a text encoder τ_ϕ. τ_ϕ is a pretrained CLIP text encoder (Radford et al., 2021). The image latent space is obtained from a encoder $\mathbf{z} = E_\psi(\mathbf{x})$ which is pretrained along with a decoder D_ψ using kullback-leibler (KL) divergence, LPIPS perceptual loss and patch discriminator as described in Rombach et al. (2022). The latent diffusion model can be implemented with a conditional denoising U-Net $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \tau_\phi(\mathbf{c}), t)$ which allows controlling the synthesis process through inputs $\mathbf{c}$.

Here, we only train the parameters θ from the diffusion model, leaving the weights ψ and ϕ from the autoencoder and text encoder respectively as pre-trained (Chambon et al., 2022). Whenever we mention samples from an unconditional model, we refer to images generated with prompts from empty strings. The training procedure is done by learning a θ^* such that

$$\theta^* = \arg\min_\theta \mathbb{E}_{\mathbf{x}_0,t,\epsilon}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \epsilon\|_2^2\right], \quad (1)$$

where $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon$, with $\mathbf{z}_0 = E_\psi(\mathbf{x})$, $t \sim \mathcal{U}(0, T)$ and $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ is the noise. We generate images using classifier-free guidance (Ho & Salimans) with the PNDM sampling strategy (Liu et al., 2022).

2.2. Sample-level Metrics for Synthetic Data

As detailed in Section 3, our method involves a filtering procedure. Evaluating the quality of synthetic examples is challenging task and numerous methods such as Inception Score (IS) (Hinz et al., 2021), Frechet Inception Distance (FID) (Heusel et al., 2017a; Kynkäänniemi et al., 2023), and Precision/Recall (PR)(Kynkäänniemi et al., 2019). IS, FID and PR are methods that compute characteristics of the distribution of the synthetic data. If the downstream use case for the synthetic data is defined, one might use downstream performance to evaluate the dataset.

However, in certain scenarios, one is interested in evaluating the qualities of individual synthetic samples such that requirements (such as privacy) over the generated dataset can be enforced post-hoc (after training). Therefore, a Alaa et al. (2022) explored α-precision, β-recall and authenticity that characterizes the fidelity, diversity and generalisation per sample. Han et al. (2023) proposed the "rarity score" which measures the uncommonness of generated images using the nearest-neighbor distance in the latent space from other real and synthetic data points. In the multimodal test-to-image setting, a common metric is the CLIP score (Ramesh et al., 2021b) which measures the alignment between the conditioning and the generated image.

2.3. Diffusion Models and Privacy

DPMs have shown to be particularly susceptible to attacks extracting its training data (Carlini et al., 2023; Somepalli

[3] https://huggingface.co/runwayml/stable-diffusion-v1-5

et al., 2023), exceeding the number of images extracted from other architectures such as generative adversarial networks (GANs) (Carlini et al., 2023). Few publications have tackled solutions for privacy preservation in diffusion models. Carlini et al. (2023) did a thorough analysis of the impact of model hyperparameters, duplicates and training dataset size on the extraction of training samples of two state-of-the-art diffusion models. Carlini et al. (2023), however, only measured memorisation via pixel-level similarity (a modified version of l_2 loss). They retrieve the samples from the training dataset that were closer to a specific synthetic image.

A popular solution to tackle privacy in deep learning models is the use of differential privacy (DP) (Dwork et al., 2014; Abadi et al., 2016) training. DP in deep learning is performed via differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016). DP-SGD preserves privacy by clipping and noising the parameter gradients during training. Dockhorn et al. (2022) show how DP-diffusion models generate images of substantially better quality than DP-GAN counterparts and have a much more stable training regime. Nonetheless, the paper showcases that, at present, this approach is limited to models with a small number of parameters, which leaves out its application to large, state-of-the-art models. In addition, despite obtaining outstanding results compared to other DP generative models, the visual quality of these samples is still far from that obtained with non-DP models.

3. Privacy Distillation

3.1. Problem Statement

Consider a real dataset $D_{real} = \{(\mathbf{x}_i, \mathbf{c}_i) \mid \forall i \in (1, 2, \ldots, N)\}$ of images $\mathbf{x}_i$ and text $\mathbf{c}_i$ belonging to a patient $\mathbf{p}_i$. We are interested in training a generative model $\boldsymbol{\epsilon}_\theta$ which is able to synthesise images $\hat{\mathbf{x}}_i$ such that $\hat{\mathbf{x}}_i$ does not contain information that can be used to identify a data point $\mathbf{x}_i, \forall i \in (1, 2, \ldots, N)$.

Following the literature (Carlini et al., 2023; Yoon et al., 2020), we hypothesise that synthetic images can enable re-identification due to model memorisation.

Definition 3.1 (ℓ, δ-Memorisation, adapted from (Carlini et al., 2023))**.** A $\mathbf{x}_i$ is considered (ℓ, δ)-*memorised* by $\boldsymbol{\epsilon}_\theta$ if $\ell(\hat{\mathbf{x}}_i, \mathbf{x}_i) \geq \delta$, where ℓ is a similarity function, δ is a threshold, and $\mathcal{A}$ is an algorithm which can extract an image $\hat{\mathbf{x}}_i$ from a generative model $\boldsymbol{\epsilon}_\theta$ without access to the original $\mathbf{x}_i$, $\hat{\mathbf{x}}_i = \mathcal{A}(\boldsymbol{\epsilon}_\theta)$. In the case of LDMs, $\mathcal{A}$ is a sampling algorithm.

We assume that $\mathbf{c}_i$ does not contain identifiable information about $\mathbf{p}_i$, therefore, we only focus on $\hat{\mathbf{x}}_i$ for identifiable information. This is a reasonable assumption since identifiable information in text, such as demographics, can be easily recognised whereas images can have more subtle details. In the next sections, we will consider two cases where we perform (i) conditional sampling $\hat{\mathbf{x}}_i = \mathcal{A}(\boldsymbol{\epsilon}_\theta, \mathbf{c}_i)$; (ii) unconditional sampling $\hat{\mathbf{x}}_i = \mathcal{A}(\boldsymbol{\epsilon}_\theta)$.

3.2. Distillation Procedure

As we are interested in safely sharing the weights θ of generative models in a privacy-preserving manner, a major concern is that synthetic images generated by a model can be used to re-identify a patient from the real training dataset. Therefore, we propose an algorithm for training a diffusion model over filtered synthetic images, minimising the model's exposure to re-identifiable data.

The procedure for *Privacy Distillation*, as depicted in Figure 1, consists of the following steps:

1. train a diffusion model ${\boldsymbol{\epsilon}_\theta}^{real}$ on real data D_{real}
2. generate a synthetic dataset D_{synth}
3. filter D_{synth}, ensuring that none of the images are re-identifiable, to obtain $D_{filtered}$
4. train a diffusion model ${\boldsymbol{\epsilon}_\theta}^{distill}$ on $D_{filtered}$
5. share ${\boldsymbol{\epsilon}_\theta}^{distill}$.

3.3. Filtering for privacy

Defining an appropriate $\ell(\hat{\mathbf{x}}, \mathbf{x})$ allows controlling which aspects of the original data one wishes to measure for memorisation. Previous work (Carlini et al., 2023) searches near-identical images utilising a Euclidean distance or pixel-by-pixel correspondence. Measuring *identity*, however, can be challenging and specific to certain modalities or organs (Kumar et al., 2017), limiting the validity of such approaches.

Assessing re-identification. Instead of pixel-based (Carlini et al., 2023) or structural-based (Kumar et al., 2017) similarities, we measure identity with a deep model $\ell = f_\theta^{re-id}$, introduced by Packhäuser et al. (Packhäuser et al., 2022). The model is trained to classify images as belonging to the same patient or not. This model, devised for X-ray images, consists of a siamese neural network with a ResNet-50 backbone. The model takes in two images, fuses the representation of the two branches and outputs a *re-identification score* (after a sigmoid) that we will note as s_{re-id}. The model is trained on real images.

When performing filtering, we compare a real image to a synthetic image. If $s_{re-id} \geq \delta$ for a pair of synthetic and real images $(\hat{\mathbf{x}}_i, \mathbf{x}_i)$, we consider that $\hat{\mathbf{x}}_i$ contains identifiable information about $\mathbf{x}_i$. For a set of synthetic images, we call *re-identification ratio* R_{re-id} the number of synthetic samples containing identifiable information about real

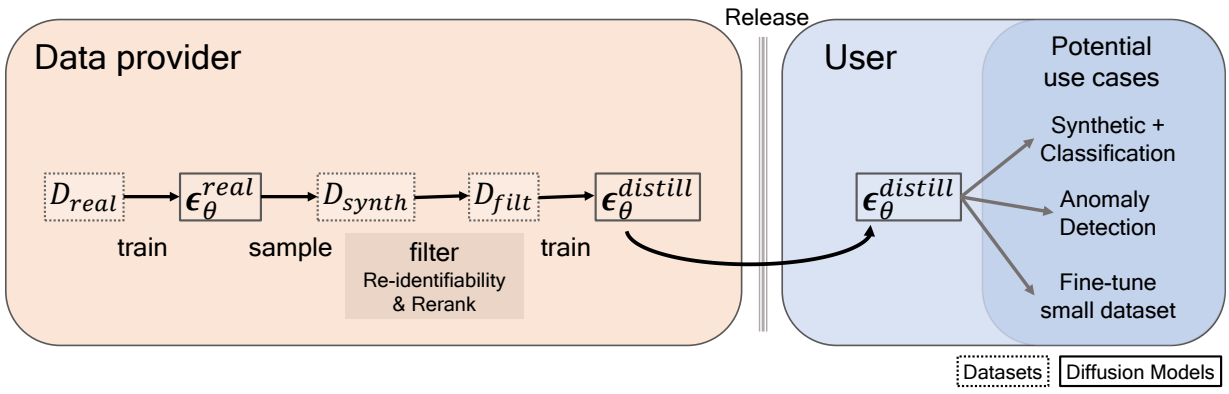


Figure 1: Privacy Distillation Pipeline.

samples divided by the total number of synthetic samples generated.

We train f_θ^{re-id} from scratch on our training set, sampling positive (images from the same patient) and negative pairs, which are randomly sampled. To avoid data imbalance, positive pairs, which were on average ten times less frequent, were oversampled, resulting in an effective dataset size of 472,992. We tested it on a set of even 101,592 non-repeated pairs, achieving a 99.16% accuracy (AUC 0.9994).

Retrieval. In a scenario where the text conditioning is not available, we need to search for the most similar real image in the training dataset before computing the re-identification score s_{re-id}. Therefore, for data sampled without conditioning, we utilised a retrieval model $f_\theta^{retrieval}$ which was also proposed in (Packhäuser et al., 2022). The model works as a feature extractor for computing the nearest neighbours in the embedding space. The model is a siamese neural network with an architecture similar to f_θ^{re-id}. The $f_\theta^{retrieval}$ excludes the layers from the merging point onwards, to function solely as a feature extractor. $f_\theta^{retrieval}$ is trained with a contrastive loss function.

During filtering, $f_\theta^{retrieval}$ identifies the closest image in terms of identity by computing the Euclidean distance between the embeddings of the query synthetic image and the embedding of every real image in our training set. When evaluating pairs of the test set from the real dataset, our trained model obtained high mean average precision at R (mAP@R) of about 95% and a high Precision@1 (the precision when evaluating how many times the top-1 images in the retrieved lists are relevant) of 97%. This way, approach enabled us to analyze and evaluate the unconditioned synthetic data accurately.

Constrative Reranking. We also need to ensure that the images in $D_{filtered}$ used for training ${\epsilon_\theta}^{distill}$ correspond to their conditioning. Therefore, we rerank the synthetic images in $D_{filtered}$ based on the image alignment with the conditioning, similar to DALL-E (Ramesh et al., 2021a).

We leverage CXR-BERT (Boecking et al., 2022) for the text encoder which is a chest X-ray (CXR) language model that utilises an improved vocabulary, pretraining procedure and text augmentations tailored to medical text. The model is fine-tuned with a contrastive text-image loss, together with an image encoder (Boecking et al., 2022). An alignment score $s_{align} = f_\theta^{im2tex}(\hat{\mathbf{x}}_i, \mathbf{c}_i)$ is computed between an image and a text prompt by passing them through an image and text encoder respectively and taking the cosine similarity between their latent spaces.

Filtering strategy. We generate N_c synthetic images for each prompt $\mathbf{c}_i$ in D_{real}. We generally choose $N_c = 10$. Therefore, D_{synth} has $N_c * N$ elements. We compute s_{align} between all generated images and corresponding conditioning using f_θ^{im2tex}; and s_{re-id} between the generated images and the real image corresponding to its prompt. For unconditional models, we use $f_\theta^{retrieval}$ to find the strongest candidate in the dataset before computing s_{re-id}. We remove all re-identified images $s_{re-id} \geq \delta$ and choose, for each $\mathbf{c}_i$, the synthetic image with the highest s_{align}.

4. Experiments

First, we evaluate how/when identity memorisation happens and the effect of training the model and sampling under different conditions and dataset sizes. Then we showcase that our model trained under Privacy Distillation can be used to train a downstream classification model while reducing re-identification risk.

4.1. Data

We use images and radiological reports from the MIMIC-CXR 2.0.0 database (Johnson et al., 2019). As text, we use each report's "impression" section, which corresponds to an interpretative summary of the findings for supporting

medical decisions. Following RoentGen (Chambon et al., 2022), we filter the data used in this study based on the length of impression in tokens, which should not exceed 76 tokens due to the text encoder limit.

Ultimately we obtained a set of 45,453 images belonging to 25,852 patients, each associated with an impression of the original radiological report. We split these into a train set of 23,268 patients (40,960 images) and a test set of 2,132 patients (3,101 images). 10% of the patients left for testing had half of their images and report pairs moved to the training dataset to allow us to assess re-identification when the patient, but not the query (image, text) pair, is part of the training dataset.

4.2. Metrics

Beyond using the re-identification score s_{re-id} and the text-to-image similarity s_{align} for measuring the quality of our synthetic dataset, we also measured image fidelity and downstream performance. We evaluate the fidelity of the synthetic images using the distribution-based metric Fréchet Inception Distance (FID) score (Heusel et al., 2017b). We utilise features extracted by the pre-trained DenseNet-121 from the *torchxrayvision* package (Cohen et al., 2022).

To assess the quality of synthetic datasets, we train a classifier f_θ^{class} of 5 different pathologies (Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion) based on the model ranked first in the CheXpert Stanford ML leaderboard (Yuan et al.).[4] We trained a DenseNet-121 on our datasets and tested it in the real hold-out test set. The network is pre-trained for 5 epochs on cross-entropy loss, then trained for another 5 epochs on an AUC loss, as per (Yuan et al.).

4.3. Measuring Re-identification Risk of Latent Diffusion Models

Effect of varying the δ. We explored how δ threshold for the s_{re-id} score impact the decision if a synthetic data point contains identifiable information of not. We explored the effect of varying δ and found no relevant difference in the resulting score for thresholds between 0.05 and 0.90, as can be seen by Figure 2. Therefore, we set $\delta = 0.5$ for the rest of the experiments.

Effect of fine-tuning. We explored the differences in terms of s_{align} between fine-tuning the model pre-trained on LAION-5B or training from scratch, and between sampling using conditioning or not. For the conditioned generation, we sample 100 instances for each of the first 400 prompts of the training dataset, resulting in 40,000 samples. For the unconditional generation, we sample 40,000 images and use the retrieval network to get the closest images in the training dataset. We calculate the re-identification ratio and FID as shown in Table 1. The lowest re-identification ratio was achieved for the data sampled from a model trained from scratch using conditioning. Unconditionally-generated datasets have higher re-identification ratios but achieve a better FID score. Nonetheless, their usability is limited, as conditional sampling allows the user to guide the generation process.

[4]We used their code, available at https://github.com/Optimization-AI/LibAUC

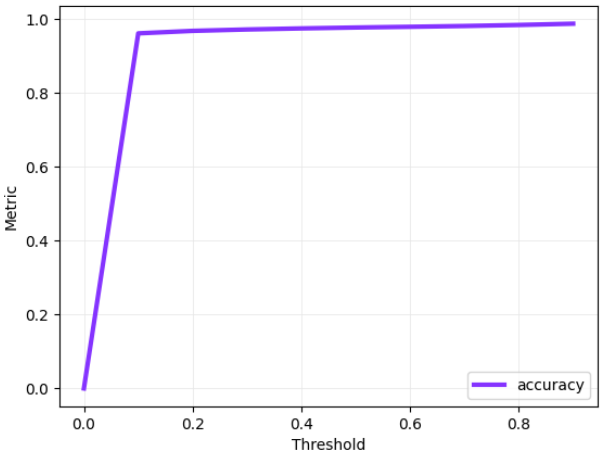


Figure 2: Effect of varying threshold δ on the re-identification score s_{re-id}.

Table 1: Evaluating the influence of pre-training and conditioning on ${\epsilon_\theta}^{real}$.

Pre-trained	Conditional	$R_{re-id} \downarrow$	FID $\downarrow$
-	-	0.057 ± 0.232	54.56
-	✓	0.015 ± 0.124	81.95
✓	-	0.034 ± 0.181	97.91
✓	✓	0.022 ± 0.232	79.27

Effect of training dataset size on memorisation. We trained one model on our full training dataset, and three models on 1%, 10% and 50% of the training dataset, respectively. Then, we calculated the R_{re-id} of a set of samples inferred using 100 instances of 400 training prompts (40,000 images in total). Figure 3 shows the *re-identification ratio*, defined as the average number of times a sample was re-identified divided by the total of generated samples.

As opposed to findings in the literature, where bigger training set sizes result in less leakage (Carlini et al., 2023), the re-identification ratio was lowest for the model trained on only 1% of the data. The TSNE plots of Figure 3 suggest that re-identification tends to happen in specific clusters, which aligns with the findings in (Su et al., 2022). We looked at the radiological reports of the top 10 most re-identified prompts, and 90% of them were associated with similar

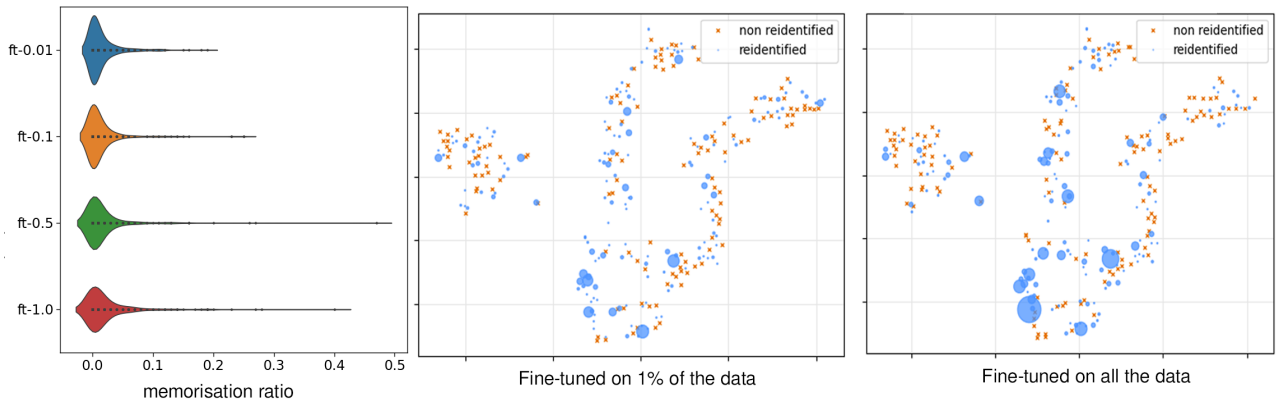


Figure 3: Left: violin plots showing the distribution of the average re-identification ratio for the models fine-tuned in different portions of the training dataset; the middle and right plots are TSNE plots of the f_θ^{im2tex} embeddings of the first 400 prompts used to test this experiment, for the model trained on 1% and 100% of the data, respectively. Orange dots correspond to prompts for which none of the generated 100 instances was re-identified, whereas blue dots are associated with prompts re-identified to some extent, the size being proportional to the re-identification ratio.

pathological phenotypes: atelectasis, pleural effusion, and lung opacity. In parallel, we observed that the proportion of images associated with these pathologies is less frequent in the first 1% of the data than it is in the whole dataset, which suggests that generated data from models trained on the full dataset might be more re-identifiable due to overfitting to subject-specific pathological features. This hypothesis is corroborated by using the Grad-CAM activations of the verification network; we observed that, in patients with pleural effusion or atelectasis, the most salient regions matched the areas with evidence for these pathologies. This, combined with the potential overfitting, likely explains the increase in the average memorisation ratio.

Effect of filtering. We compare the impact of our filtering strategy on the synthetic datasets D_{synth} and $D_{filtered}$. We sample 10 instances for each of the 40,959 training prompts from the proposed privacy-preserving model. We then filter by s_{re-id}, and pick the sample with the better s_{align} score, resulting in a filtered dataset of 40,959 images (Figure 4). We found that filtering improves the s_{align} and reduces the number of memorised (re-identifiable) samples to 0, given a δ.

Visualising of How Synthetic Examples are Memorised. We now focus on analysing Grad-CAM++ (Chattopadhay et al., 2018) heatmaps of the f_θ^{re-id} network in Figure 5a and some samples of synthetic images and their associated conditioning information in Figure 5b. For the explainability heatmap, we used the second-order gradients of the fourth layer of the ResNet-50 (which is a common choice for this architecture[5]).

[5] https://github.com/jacobgil/pytorch-grad-cam#chosing-the-target-layer

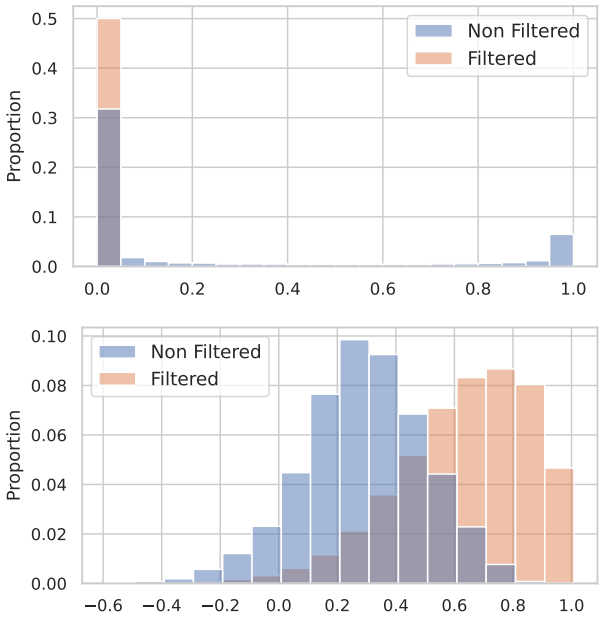
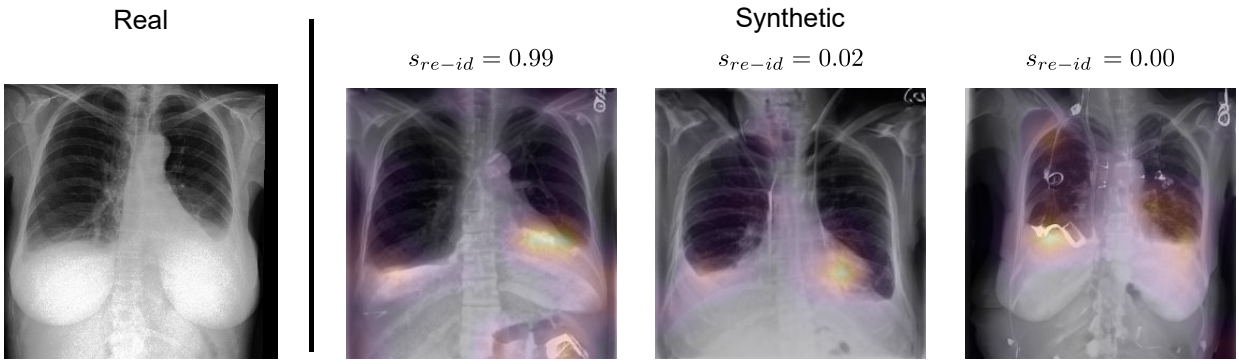


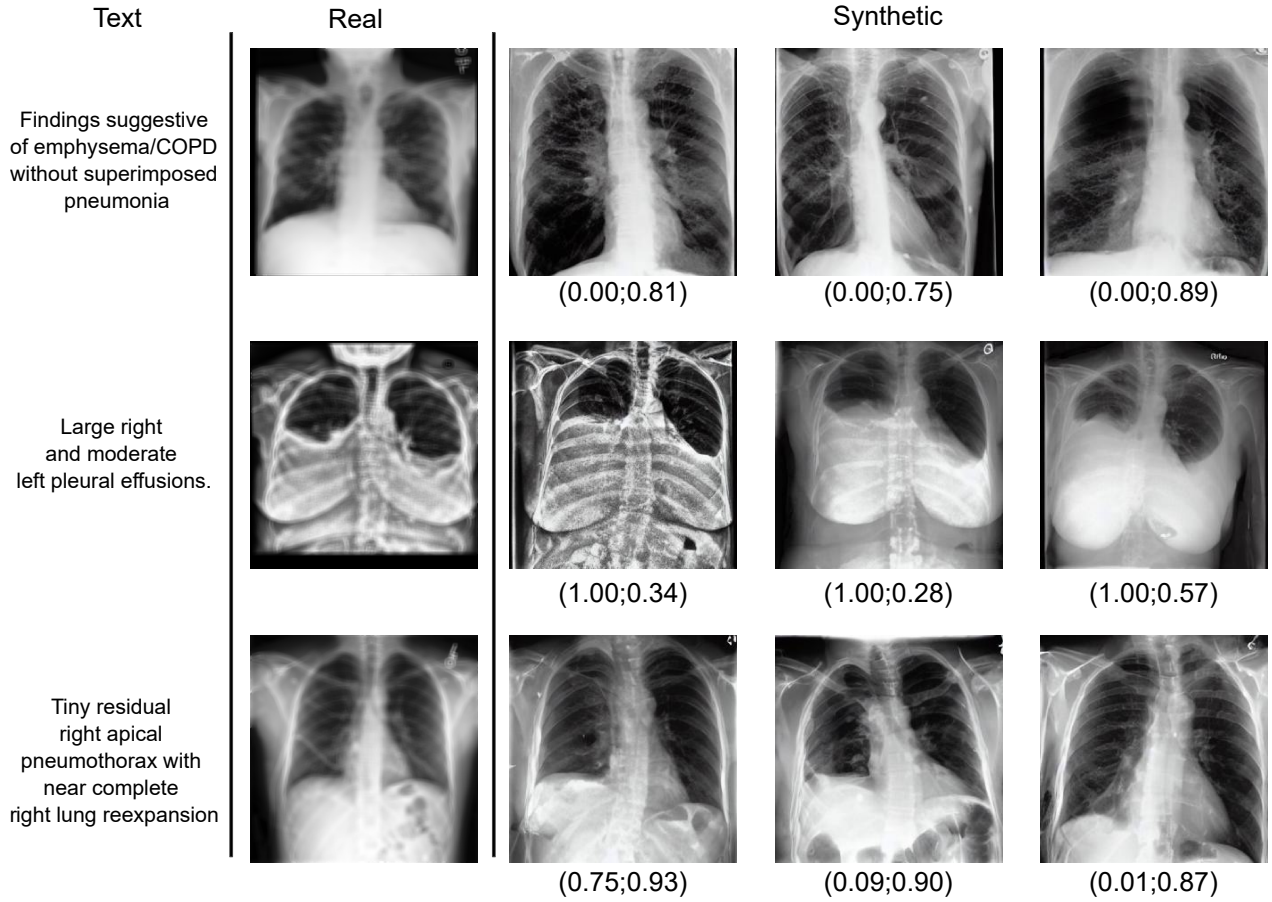
Figure 4: *Top*: s_{re-id} distribution ; *Bottom*: s_{align} distribution. Comparing distribution of the non-filtered D_{synth} and filtered $D_{filtered}$ dataset.

4.4. Privacy Distillation

We now show empirically that $\epsilon_\theta{}^{distill}$ indeed reduces re-identification risk. We also assess whether $\epsilon_\theta{}^{distill}$ is able to produce useful synthetic datasets for downstream tasks. We train classifiers f_θ^{class} (see section 4.2) on D_{real}, D_{synth}, capped at 40,000 images, and a dataset $D_{distill}$ of 40,000 images sampled from $\epsilon_\theta{}^{distill}$, our distilled model, using



(a) Grad-CAM++ heatmaps over the f_θ^{re-id}. From left to right, the first synthetic image contains identifiable information about the real image in the areas towards the bottom of the lungs. The other two images do not contain identifiable information as defined by the re-identification score s_{re-id}.



(b) We display several synthetic images along with their prompt and real image associated with the prompt. We show below the synthetic images the re-identification score and text-to-image similarity in the format $(s_{re-id}; s_{align})$. In the first row, the synthetic images do not contain any identifiable information about the real image while corresponding fairly well to the text description. In the second row, all synthetic images contain identifiable information, despite having a different style/contrast. The bottom row displays synthetic images with good text alignment but some contain identifiable information and some do not.

Figure 5: Illustration showing how synthetic images, obtained by using the same prompt of the real image as conditioning, relates to the real image as well as the prompt. We try to mitigate privacy risks in these illustrations because the license of MIMIC-CXR [6] does not allow sharing data. In 5a, the "Real" image is actually a synthetic image that is very similar to the original image. In 5b, we blurred the real images.

Table 2: Privacy Distillation performance: memorisation ratio, predicted AUC for the classifier on the real test set and s_{align} score.

Dataset	$R_{re-id} \downarrow$	f_θ^{class} AUC $\uparrow$	$s_{align} \uparrow$
D_{real}	-	0.863	$0.698_{0.259}$
D_{synth}	4.24%	0.830	$0.645_{0.219}$
$D_{distill}$	**1.34%**	0.810	$0.611_{0.272}$

the training set prompts. We measure the AUC on the 3,101 images of our test set. The results are reported in table 2, in addition to the re-identification ratio R_{re-id} and the s_{align} score.

Training a model on synthetic data slightly affects performance (AUC and s_{align} decrease), but the resulting value is still comparable to the literature (Jacenkow et al., 2022). Nonetheless, the re-identification ratio between the initial and the distilled models is decreased by more than 3-fold. We hypothesise that filtering re-identifiable data might also filter out unique phenotypes, more prone to be memorised (Carlini et al., 2023), resulting in reduced model generalisability and performance.

5. Discussion and conclusion

This study has demonstrated that the application of *privacy distillation* can effectively reduce the risk of re-identification and leakage in latent diffusion models without excessively compromising the downstream task performance. In line with other approaches, such as differential privacy (Dockhorn et al., 2022), there is a trade-off between privacy and quality. In our proposed method, the trade-off is between the degree of filtering (more privacy) and downstream model utility (less privacy). Additionally, *privacy distillation* can be applied iteratively by adding more filtering-sampling-training steps, an approach that should be the subject of future experiments.

This approach has the potential to facilitate the sharing of medical imaging generative models for fine-tuning and subsequent use. Other downstream tasks, such as segmentation, which we could not do due to the absence of ground truth masks, should also be contemplated to further explore the impact of the filtering. While in this paper, we use a text-to-image synthesis network and rely on a re-identification metric devised specifically for X-ray imaging, our approach could be applied to other imaging modalities and conditioning types, replacing the data, model and re-identification metric.

Please note that while the proposed method requires training a second model on a filtered dataset, the filtering approach can be applied alone to reduce the re-identification risk of synthetic datasets (in the case where realising a synthetic dataset instead of a model is enough). This post hoc filtering procedure acts as a model auditor (Alaa et al., 2022) and can be used to improve any generative model. This is of particular interest when retraining a model is expensive or imposing certain properties is impractical.

5.1. Limitations

While *privacy distillation* significantly reduces re-identifiability, it is important to note that some minor risks may still exist. We highlight that the ability to measure these risks is bounded by the accuracy and generalisation capabilities of the chosen measure of re-identification s_{re-id}. Ensuring privacy preservation is a task that depends on various assumptions, such as the definition of "re-identifiability", which is limited by the choice of the metric and threshold. Further research should explore alternatives with a combination of metrics, such as the Kullback-Leibler distance or a loss-based score, as proposed by Hu et al. (Hu & Pang).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification, 2023.

Boecking, B., Usuyama, N., Bannur, S., Coelho de Castro, D., Schwaighofer, A., Hyland, S., Wetscherek, M. T., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., and Oktay, O. Making the most of text semantics to improve biomedical vision-language processing. In *ECCV*, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv*, 2023.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., and Chaudhari, A. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering 2021 5:6*, 2021.

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., and Bertrand, H. TorchXRayVision: A library of chest X-ray datasets and models. In *MIDL*, 2022.

Dhariwal, P., Openai, and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 12 2021.

Dockhorn, T., Cao, T., Vahdat, A., and Kreis, K. Differentially private diffusion models. 2022.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Fernandez, V., Pinaya, W. H. L., Borges, P., Tudosiu, P.-D., Graham, M. S., Vercauteren, T., and Cardoso, M. J. Can segmentation models be trained with fully synthetically generated data? In *SASHIMI*. Springer, 2022.

Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *The Eleventh International Conference on Learning Representations*, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017a.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPs*, 30, 2017b.

Hinz, T., Fisher, M., Wang, O., and Wermter, S. Improved techniques for training single-image gans. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. Deep models under the gan: Information leakage from collaborative deep learning.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Hu, H. and Pang, J. Membership inference of diffusion models.

Jacenkow, G., O'Neil, A. Q., and Tsaftaris, S. A. Indication as Prior Knowledge for Multimodal Disease Classification in Chest Radiographs with Transformers. *IEEE ISBI*, 2022.

Jegorova, M., Kaul, C., Mayor, C., O'Neil, A. Q., Weir, A., Murray-Smith, R., and Tsaftaris, S. A. Survey: Leakage and privacy at inference time. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–20, 2023.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., ying Deng, C., Mark, R. G., and Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

Jordon, J., Wilson, A., and van der Schaar, M. Synthetic data: Opening the data floodgates to enable faster, more directed development of machine learning methods. *arXiv preprint arXiv:2012.04580*, 2020.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. Diffusion models for medical image analysis: A comprehensive survey. *arXiv:2211.07804*, 2022.

Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O., and Toews, M. Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis. *NeuroImage*, 158:242–259, 2017. ISSN 1053-8119.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*, 2023.

Li, Z., Li, Y., Zhao, P., Song, R., Li, X., and Yang, J. Is synthetic data from diffusion models ready for knowledge distillation?, 2023.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022.

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., and Bano, A. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48: 100546, 2023.

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., and Maier, A. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Scientific Reports 2022 12:1*, 12(1):1–13, sep 2022. ISSN 2045-2322.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., and Cardoso, M. J. Brain imaging generation with latent diffusion models. In *Deep Generative Models MICCAI*. Springer, 2022a.

Pinaya, W. H. L., Graham, M. S., Gray, R., da Costa, P. F., Tudosiu, P.-D., Wright, P., Mah, Y. H., MacKinnon, A. D., Teo, J. T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *MICCAI*, 2022b.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *ICML*, 2021a.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021b.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Sanchez, P., Kascenas, A., Liu, X., O'Neil, A. Q., and Tsaftaris, S. A. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022.

Somepalli, G., Singla, V., Goldblum, M., Geiping, Wu, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, 2023.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Su, R., Liu, X., and Tsaftaris, S. A. Why patient data cannot be easily forgotten? In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer, 2022.

Yoon, J., Drumright, L. N., and van der Schaar, M. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE Journal of Biomedical and Health Informatics*, 24, 2020.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification.

Zhang, L. and Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. feb 2023. URL https://arxiv.org/abs/2302.05543v1.

Zhang, Z., Yan, C., and Malin, B. A. Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125, jan 2022. ISSN 1532-0480. doi: 10.1016/J.JBI.2021.103977. URL https://pubmed.ncbi.nlm.nih.gov/34920126/.