Membership Inference Attacks for Face Images Against Fine-Tuned Latent Diffusion Models

Lauritz Christian Holme[a], Anton Mosquera Storgaard[b] and Siavash Arjomand Bigdeli[c]
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark

Keywords: Membership Inference Attack, Latent Diffusion Model.

Abstract: The rise of generative image models leads to privacy concerns when it comes to the huge datasets used to train such models. This paper investigates the possibility of inferring if a set of face images was used for fine-tuning a Latent Diffusion Model (LDM). A Membership Inference Attack (MIA) method is presented for this task. Using generated auxiliary data for the training of the attack model leads to significantly better performance, and so does the use of watermarks. The guidance scale used for inference was found to have a significant influence. If a LDM is fine-tuned for long enough, the text prompt used for inference has no significant influence. The proposed MIA is found to be viable in a realistic black-box setup against LDMs fine-tuned on face-images.

1 INTRODUCTION

Generative image models such as OpenAI's DALL·E or Stability AI's Stable Diffusion have advanced rapidly and seen a great rise in popularity over the last few years. To train models like these, millions of images are needed leading to the requirement of huge image datasets.

This need has led to image generation models being trained on images without the needed consent or necessary permissions. As a consequence - besides the violation of the ownership of images - image generation models have been able to copy the style of artists and generate images in the likeness of others without the permission to do so. Such infringements affecting both individuals and organisations can be difficult to prove, as it requires knowledge of the images used to train the generative model.

The aim of this paper and research is to investigate Membership Inference Attacks (MIA) on Latent Diffusion Models (LDM) (Rombach et al., 2022). To scope the project only Stable Diffusion v1.5 (Rombach et al., 2022) fine-tuned on face images is considered. Being able to infer if an image was part of a dataset used for training a generative model would considerably help individuals and organisations who suspect that their images have been unrightfully used. As mentioned in (Dubi'nski et al., 2023), evaluating MIA against a fine-tuned model is a potential pitfall[1]. This paper intentionally focuses on fine-tuned models as they are commonly used in real-life applications. It should be kept in mind that the results presented do not apply to non-fine-tuned models.

1.1 Definition of the Target Model

In this paper, 'Target Model' $\mathbf{M_T}$ will be used to denote the model which the MIA is performed against. The model $\mathbf{M_T}$ is in this paper characterised by its ability to turn text, $\mathcal{T}$, into some $H \times W$-dimensional image with 3 RGB colour channels, i.e. $\mathbb{R}^{(H,W,3)}$.

$$\mathbf{M_T} : \mathcal{T} \rightarrow \mathbb{R}^{(H,W,3)} \tag{1}$$

This is the model for which it is desired to infer whether an image belongs to its training set $\mathbf{D_{Target}}$. Throughout the paper, only a black-box setup will be considered. This means that the target model can only be used as intended, i.e. providing textual input to generate output. No additional knowledge of training data, model weights, and etc. is available.

The Target model $\mathbf{M_T}$ could be produced as shown on fig. 1 where a LDM (such as Stable Diffusion) is fine-tuned on a dataset to produce very domain-specific images which imitate the dataset. This could be images that are in the likeness of a specific artist's style or a group of people.

[a] https://orcid.org/0009-0001-3043-5561
[b] https://orcid.org/0009-0001-1437-6004
[c] https://orcid.org/0000-0003-2569-6473

[1] The reason it is seen as a pitfall is that fine-tuning a model easily leads to over-fitting to an image dataset resulting in higher accuracy on predicting member images. This is also shown in (Carlini et al., 2023).

Holme, L. C., Storgaard, A. M. and Bigdeli, S. A.
Membership Inference Attacks for Face Images Against Fine-Tuned Latent Diffusion Models.
DOI: 10.5220/0013182600003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 439-446
ISBN: 978-989-758-728-3; ISSN: 2184-4321

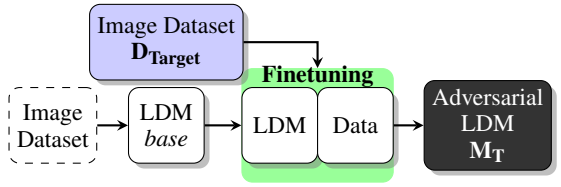


Figure 1: An approach a malicious actor could use to obtain a LDM which is trained on unrightfully obtained data.

The Adversarial LDM will have obtained its high quality generation abilities from the pre-training on some image dataset (such as LAION-5B) making it able to generalise and mix styles with the images from $\mathbf{D_{Target}}$. It could be difficult to infer membership when a large image generation model contains information from hundreds of millions of images. However, as the model fine-tunes on a smaller set of images, $\mathbf{D_{Target}}$, it would enhance the information leakage from $\mathbf{D_{Target}}$ and it could be possible to infer whether the images have been used in the fine-tuning or not.

1.2 Definition of the Attack Model

The 'Attack Model' $\mathbf{M_A}$ denotes the model trained to infer the membership of a query image in relation to the target model's training set. It is characterised by taking some image $\mathbb{R}^{(H,W,3)}$ and translating it to a value $\mathcal{P}$ between 0 and 1 expressing the predicted probability of the image being part of $\mathbf{D_{Target}}$.

$$\mathbf{M_A} : \mathbb{R}^{(H,W,3)} \rightarrow \mathcal{P} \tag{2}$$

The attack model will be trained using supervised learning. To create the training set for the attack model, the training positives will be obtained by using the target model to generate images. This is done under the assumption that the target model's output leaks information of the training data. This assumption is required as the goal is to make the attack model learn to recognise the data distribution of the target model's training set and not its generated output.

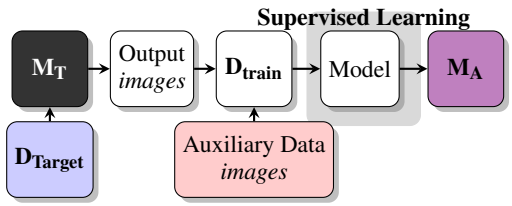


Figure 2: The approach taken to train the Attack Model $\mathbf{M_A}$. Images ($\mathbf{D_{Target}}$) are used to train $\mathbf{M_T}$ which then outputs some images which are used as positives in $\mathbf{D_{Train}}$. The negatives in $\mathbf{D_{Train}}$ come from an auxiliary dataset. A model is then trained on $\mathbf{D_{Train}}$ using supervised learning to produce the final attack model $\mathbf{M_A}$.

The training negatives are obtained from an 'auxiliary dataset'. The purpose of the images in this dataset is to represent the same or a similar domain as the one the target model has been trained on. The auxiliary dataset should only contain images that the target model has not trained on, so it can be used as training negatives for the attack model. The training process of the attack model can be seen in Figure 2.

2 RELATED WORKS

In 2017 the paper *Membership Inference Attacks against Machine Learning Models* (Shokri et al., 2017) proposed how to conduct a Membership Inference Attack against classification models in a black-box setup. The paper investigates the possibility of creating "shadow models" which mimic a target model and uses them to train an attack model for the task of classifying the membership of data samples, i.e. whether they belong to the training set of a target model or not. A shadow model as described in the paper is a model that aims to be similar to the target model, i.e. have a similar architecture, be trained in a similar way, use similar training data etc. The motivation for using shadow models is that it allows for training an attack model using supervised learning.

This is possible by using the training and test data for the shadow models along with their outputted classification vectors when training the attack models (training set are members, test set are non-members). The paper shows that MIAs can be feasible in a black-box setup.

Besides targeting classification models, multiple papers have also investigated the possibility of performing MIAs on generative models. One such paper is *GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models* (Chen et al., 2020). The paper has its main focus on MIAs against Generative Adversarial Networks and describes multiple attack scenarios including a full black-box attack. Common for all attack scenarios presented in the paper is the assumption that the probability a generator can produce a sample is proportional to that sample being a member of the generator's training set. This assumption is made as the generator model is trained to approximate the distribution of the training data. The full black-box attack presented works by sampling from a generator and then finding the sample closest to the query sample (ie. the sample to determine the membership of) using a distance metric. The most similar generated sample is then used to approximate the probability that the generator can create the sample which is then used to predict the membership.

A generalised approach to MIA was proposed in the paper *Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models* (Zhang et al., 2023). The proposed approach does not require shadow models, works in a black-box setup, and can be used against multiple generative models. The core idea for the technique presented is to take advantage of the similar data distribution between the training data of a generative model and its output as done by (Chen et al., 2020). The similarity between the data distribution relies on the model overfitting to its training data. This similarity functions as an information leakage, making the generated output describe traits of the model's training data.

In the paper, supervised learning is used for training an attack model for the membership inference task. This is done by querying the target model to generate output used as training positives. This is assumed to be representative as training positives due to the assumption of a similar data distribution between the target models training data and its output. The training negatives are obtained from an auxiliary dataset. The paper uses the Resnet-18 architecture for the attack model.

3 DATA

The data used can be divided into two groups. The data used to fine-tune the target model, and the data used to train the attack model.

The focus of this paper is face-images. An image dataset is needed for the experiments and the images should not be contained in LAION-5B[2]. Data from two universities are chosen which have publicly available images of their employees on their websites. The universities are the Technical University of Denmark (DTU) and Aalborg University (AAU).

3.1 Data Source for the Target Model

To fine-tune the target model with face-images three different face datasets were considered:

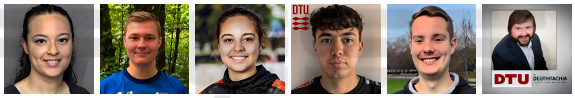
- $\mathbf{D_{DTU}}$: Images scraped from DTU orbit.
- $\mathbf{D_{AAU}}$: Images scraped from AAU vbn.
- $\mathbf{D_{LFW}}$: Labeled Faces in the Wild (Huang et al., 2007). It contains 9,452 images.

[2]The images should not be contained in LAION-5B as it was used to train the target model (Stable Diffusion v1.5) and a portion of the image dataset should be used as non-members

After collection, the two image data sets $\mathbf{D_{DTU}}$ and $\mathbf{D_{AAU}}$ were partitioned into two subsets, thus producing four different datasets: $\mathbf{D^{seen}_{DTU}}$, $\mathbf{D^{unseen}_{DTU}}$, $\mathbf{D^{seen}_{AAU}}$, and $\mathbf{D^{unseen}_{AAU}}$. The 'seen' datasets are used to fine-tune the target model while the 'unseen' datasets are only used to test/train the attack model. Two more datasets are created which are copied from $\mathbf{D^{seen}_{DTU}}$. One set of the images gets a visible DTU watermark in the top right corner, and the other set gets overlayed with an almost invisible DTU logo, this is shown in fig. 6. On table 1 there is a description of all the variations of datasets used to fine-tune the target model.

3.2 Data Source for the Attack Model

As mentioned in section 3.1 $\mathbf{D^{unseen}_{DTU}}$ and $\mathbf{D^{unseen}_{AAU}}$ are not used to fine-tune the target model, this way they can be used to test or train the attack model with the certainty that there is no data leakage into the generated images. The output of the target models are used in the training of the attack model. On table table 2 there is a description of all the different image-sets used for training / testing the attack model in different experiments. On fig. 3 examples can be seen of the generated images from the target models.



(a) $\mathbf{D^{gen}_{DTU}}$ (b) $\mathbf{D^{gen}_{AAU}}$ (c) $\mathbf{D^{gen}_{hwm}}$ (d) $\mathbf{D^{gen}_{wm}}$ $\mathbf{D^{gen}_{DTU}}$ (e) $\mathbf{D^{gen}_{DTU+LFW}}$ (f) $\mathbf{D^{gen}_{NFT}}$

Figure 3: These are examples of the data in the different generated image datasets which are used to train the attack model.

4 METHOD

The Target Dataset. $\mathbf{D_{target}}$ used for fine-tuning $\mathbf{M_T}$ is a dataset consisting of image-text pairs with the text describing the image. A BLIP model (Li et al., 2022) is used to auto-label the images. When $\mathbf{D^{seen}_{DTU}}$ is used to create $\mathbf{D_{target}}$, the labels generated with BLIP are conditioned with `"a dtu headshot of a"` and for $\mathbf{D^{seen}_{AAU}}$, `"a aau headshot of a"` is used. In other words, all labels in $\mathbf{D^{DTU}_{target}}$ begin with `"a dtu headshot of a (...)"`.

The Training Positives. $\mathbf{D^{gen}}$ are generated using the fine-tuned model $\mathbf{M_T}$. 2,500 images are generated using 100 inference steps and a guidance scale of 7.5. When using the model fine-tuned on DTU images, the prompt `"a dtu headshot"` is used and for the model fine-tuned on AAU images, the prompt `"a`

Table 1: This table describes the different datasets used for fine-tuning the target models.

Symbol	Size, n	Description
$\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$	1,120	Partition of $\mathbf{D_{DTU}}$ used to fine-tune the target model $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$\mathbf{D}_{\mathbf{AAU}}^{\mathbf{seen}}$	1,120	Partition of $\mathbf{D_{AAU}}$ used to fine-tune the target model $\mathbf{M}_{\mathbf{T}}^{\mathbf{AAU}}$
$^{\mathbf{wm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$	1,120	Partition of $\mathbf{D_{DTU}}$ with watermarks used to fine-tune the target model $^{\mathbf{wm}}\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$^{\mathbf{hwm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$	1,120	Partition of $\mathbf{D_{DTU}}$ with hidden watermarks used to fine-tune $^{\mathbf{hwm}}\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$\mathbf{D_{DTU+LFW}}$	2,240	A combination of images from $\mathbf{D_{DTU}}$ and $\mathbf{D_{LFW}}$ used to fine-tune $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU+LFW}}$

Table 2: This table describes the different datasets used for fine-tuning the attack models.

Symbol	Size, n	Description
$\mathbf{D}_{\mathbf{DTU}}^{\mathbf{unseen}}$	1,103	A partition of $\mathbf{D_{DTU}}$ that is **not** used to fine-tune any image generation model.
$\mathbf{D}_{\mathbf{AAU}}^{\mathbf{unseen}}$	978	A partition of $\mathbf{D_{AAU}}$ that is **not** used to fine-tune any image generation model.
$\mathbf{D_{LFW}}$	9,452	Labeled Faces in the Wild: A Database for Studying Face Recognition (Huang et al., 2007)
$\mathbf{D}_{\mathbf{DTU}}^{\mathbf{gen}}$	2,500	Generated by a LDM which has been fine-tuned on the $\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$ image set, i.e. $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$\mathbf{D}_{\mathbf{AAU}}^{\mathbf{gen}}$	2,500	Generated by a LDM that has been fine-tuned on the $\mathbf{D}_{\mathbf{AAU}}^{\mathbf{seen}}$ image set, i.e. $\mathbf{M}_{\mathbf{T}}^{\mathbf{AAU}}$
$^{\mathbf{wm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{gen}}$	2,500	Generated by a LDM which has been fine-tuned on the $^{\mathbf{wm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$ image set, i.e. $^{\mathbf{wm}}\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$^{\mathbf{hwm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{gen}}$	2,500	Generated by a LDM which has been fine-tuned on the $^{\mathbf{hwm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$ image set, i.e. $^{\mathbf{hwm}}\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$
$\mathbf{D}_{\mathbf{DTU+LFW}}^{\mathbf{gen}}$	2,500	Generated by a LDM that has been fine-tuned on the $\mathbf{D_{DTU+LFW}}$ image set, i.e. $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU+LFW}}$
$\mathbf{D}_{\mathbf{NFT}}^{\mathbf{gen}}$	2× 2,500	Generated by a Non Fine-Tuned (NFT) target model $\mathbf{M_T}$, i.e. Stable Diffusion out-of-the-box. This dataset also comes in two versions, one which was prompted with "`a dtu headshot`" and one which was prompted with "`a aau headshot`"

`aau headshot`" is used. 25 images are generated per seed.

The Auxiliary Data. i.e. the training negatives for the supervised learning of $\mathbf{M_A}$ are supposed to represent all images from the same (or a similar) domain as $\mathbf{D_{target}}$ not seen by $\mathbf{M_T}$. The auxiliary data must not have been used for training $\mathbf{M_T}$, i.e. the data should not be part of $\mathbf{D_{target}}$ or the original training data for $\mathbf{M_T}$. In this paper the domain of interest is face images, so the auxiliary data should be face images not seen by $\mathbf{M_T}$.
In (Zhang et al., 2023) they propose using an auxiliary dataset consisting of real images from other datasets. The experiments conducted in this paper will default to constructing the auxiliary datasets using another generative model to generate images. This is done to ensure the attack model does not simply learn to classify real images and generated images. This is also mentioned as the 2nd pitfall for MIA on LDM's in (Dubi'nski et al., 2023). It should be noted that this is more computational expensive than simply using some dataset, as it will require training a generative model when performing a MIA.

4.1 Model Specifications

The Target Model. $\mathbf{M_T}$ used for this project is a fine-tuned version of Stable Diffusion v1.5 (Rombach et al., 2022). Stable Diffusion v1.5 is trained on a subset of LAION-5B. The model is then fine-tuned on each of the datasets presented in section 3.1 resulting in multiple models, e.g. $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU}}$ which is the SD v1.5 model fine-tuned on $\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$ or $\mathbf{M}_{\mathbf{T}}^{\mathbf{DTU\ WM}}$ which is the SD v1.5 model fine-tuned on $^{\mathbf{wm}}\mathbf{D}_{\mathbf{DTU}}^{\mathbf{seen}}$ instead.

The Attack Model. used is Resnet-18 which was introduced in 2015 (He et al., 2015). The pretrained weights[3] are kept, however a fully connected layer with two neurons replaces the standard 1000-neuron final layer. The loss function used is categorical cross-entropy and the Adam optimizer is used for fast convergence.

4.2 Experimental Setup

Multiple experiments are performed to determine the success and performance of MIAs against the target model, $\mathbf{M_T}$. Similar for all tests is that they aim to help investigate the possibility of a successful MIA on a $\mathbf{M_T}$ in the setup depicted in Figure 4. As illustrated in the figure, some $\mathbf{M_T}$ is being fine-tuned on an image dataset which has been unrightfully obtained, $\mathbf{D_{target}}$.

The target model $\mathbf{M_T}$ is then queried to generate output images used as training positives in $\mathbf{D_{train}}$ and some auxiliary data is added to $\mathbf{D_{train}}$ as negatives. The attack model $\mathbf{M_A}$ is then trained on $\mathbf{D_{train}}$ using supervised learning. The resulting model $\mathbf{M_A}$ is then queried on images to classify whether they were part of $\mathbf{D_{target}}$ or not. 15% of the test data is used for validation of $\mathbf{M_A}$.

[3]The weights are available here https://download.pytorch.org/models/resnet18-f37072fd.pth

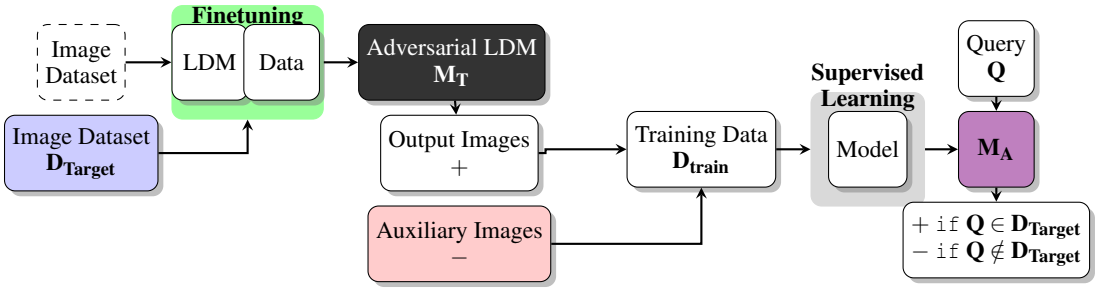


Figure 4: A graphic representation of how an attack model $\mathbf{M_A}$ could be built and later queried with an image $\mathbf{Q}$.

For the experiments $\mathbf{M_T}$ is fine-tuned using a known $\mathbf{D_{target}}$, hereby the ground-truth is known when creating $\mathbf{D_{train}}$ and querying the resulting model $\mathbf{M_A}$. This makes it possible to determine the performance of $\mathbf{M_A}$ and the MIA. On fig. 5 the approach to testing the attack model is shown.

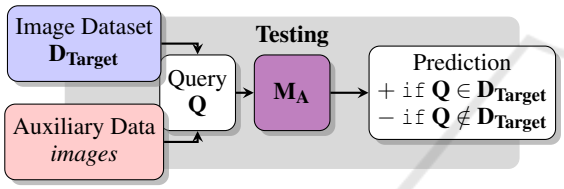


Figure 5: $\mathbf{M_A}$ is given a set $\mathbf{Q}$ of images combined from $\mathbf{D_{Target}}$ and an auxiliary image data set. This allows for evaluating the performance of $\mathbf{M_A}$.

5 RESULTS

For each test, the attack model has been trained 5 different times with 5 different seeds to calculate the 95% confidence intervals of the metrics of interest. Zero-shot classification using the CLIP model (Radford et al., 2021) is used as a baseline. On table 3, all the results are summarised. Note that the motivation/basis of several tests are different, therefore they should not all be directly compared on only the AUC score.

The Impact of the Relationship Between Training and Test Negatives. can be seen on table 3 No. 3, 4, and 7. The MIA is successful across all three tests: $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D}^{\mathbf{seen}}_{\mathbf{AAU}}$, $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$, and $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D_{LFW}}$. All three test uses $\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$ and $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ for the training of $\mathbf{M_A}$. No significance difference was found in the MIA performance for the different settings of test positives and negatives. In our case it does not seem to matter if the train negatives were generated using the test negatives (cf. table 3 No. 3 vs 4). Neither did it seem to matter if the training and test negatives are sampled from the same data distribution distinct from the training positives or not (cf. table 3 No. 3 vs 7). $\mathbf{M_A}$ clearly outperforms the baseline in test 3 and 4, but the baseline model performs better on experiment 7 with a near perfect AUC for the baseline. This indicates that using a not-generated image-set against a generated image-set artificially boosts the MIA performance.

The Effect of Using Real Images for the Auxiliary Dataset. can be seen on test (table 3 No. 2). It uses 2,000 images from $\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$ and 2,000 images from $\mathbf{D_{AAU}}$ to train $\mathbf{M_A}$. The attack model is then tested on 952 $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ images (positives) and 8,034 $\mathbf{D_{LFW}}$ images (negatives). This was done to test the effect of using real images for the auxiliary set (training negatives). As can be seen on table 3 experiment No. 2, $\mathbf{M_A}$ is successful in the test and achieves an AUC of $\sim 0.71 \pm 0.02$.
When comparing this with table 3 No. 7, which shows the same test except $\mathbf{M_A}$ is trained using the generated auxiliary set $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ instead, it becomes apparent that it is beneficial for the attack model to use a generated auxiliary set as it's training negatives. Zhang et al. finds that using a generated auxiliary dataset is comparable to using real images (Zhang et al., 2023). In our case, we find that using a generated auxiliary dataset is significantly better than using real images.

It can also be seen on table 3 No. 2, that the baseline model excels in classifying $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ and $\mathbf{D_{LFW}}$ as members and non-members and clearly outperforms $\mathbf{M_A}$ for this test with a near perfect AUC[4]. Note that the perfect AUC is likely due to the non-generated auxiliary data being too dissimilar to the positives, which artificially boosts the MIA performance.

Results of MIA Without Fine-Tuning. can be seen on Test no. 10. It is carried out where $\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$ is used as both the training positives and negatives for $\mathbf{M_A}$. As

[4]This test is identical to the test depicted on table 3 No. 7, as only the training negatives differ in the two tests, and the training negatives are not considered by the baseline.

Table 3: This table contains a summary of all results from all tests. It also includes training and test datasets. The interval on the Resnet-18 AUC Score is the 95%-confidence interval calculated over 5 repetitions.

No.	Experiment	Train Pos.	Train Neg.	Test Pos.	Test Neg.	R18 AUC	CLIP AUC
1	Generated DTU vs generated AAU	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	1.00 ± 0.00	0.94
2	Generated DTU vs non-generated AAU	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}_{\mathbf{LFW}}$	0.71 ± 0.02	0.99
3	DTU vs AAU Seen	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{AAU}}$	0.86 ± 0.01	0.63
4	DTU vs AAU Unseen Trained 50, 100, and 400 Epochs	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	0.86 ± 0.02, 0.82 ± 0.01, 0.86 ± 0.02	0.63, 0.62, 0.63
5	Generalised Prompt					0.82 ± 0.05	0.56
6	Guidance Scale s = 0, 4, 8, 12, 16					0.70, 0.87, 0.89, 0.85, 0.58	-
7	DTU vs LFW	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}_{\mathbf{LFW}}$	0.89 ± 0.04	0.99
8	DTU vs DTU	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{DTU}}$	0.53 ± 0.00	0.52
9	NFT vs AAU	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	0.66 ± 0.03	0.55
10	NFT vs NFT	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	0.54 ± 0.03	0.55
11	Watermark	$^{\mathbf{wm}}\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$^{\mathbf{wm}}\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	1.00 ± 0.00	0.95
12	Hidden Watermark	$^{\mathbf{hwm}}\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$^{\mathbf{hwm}}\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	0.83 ± 0.03	0.57
13	DTU+LFW vs AAU	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU+LFW}}$	$\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$	$\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$	$\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$	0.83 ± 0.02	0.64

to stay consistent with the other tests, $\mathbf{M_T}$ is prompted with `"a dtu headshot"` and `"a aau headshot"` to generate two different $\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$. As no information leakage can exist between neither the positives or negatives (unless the pretraining of SD 1.5 includes images from AAU or DTU), the MIA should not perform better than random guessing. However it did, which could be explained by a coincidence in data distribution similarity.

For the Test no. 9, $\mathbf{M_A}$ was first trained on $\mathbf{D}^{\mathbf{gen}}_{\mathbf{NFT}}$ as positives and $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ as negatives. Then it was tested using $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ as positives and $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$ as negatives. The MIA was still successful although the AUC score achieved by $\mathbf{M_A}$, 0.66 ± 0.03, is significantly worse than in all other tests on $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$.

As there is no relation between training and test positives, this result shows that there is information leakage between training negatives $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ and test negatives $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$. The implication of this discovery is that the tests using $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ as training negatives and $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$ (or $\mathbf{D}^{\mathbf{seen}}_{\mathbf{AAU}}$) as test negatives have a artificial boost in their performance and thus inflated metrics. This could be explained by the fact that they originate from the same data distribution.

Relationship Between Training Time of Target Model and Success of MIA. The MIA performance against target models fine-tuned on $\mathbf{D}^{\mathbf{DTU}}_{\mathbf{target}}$ for an increasing number of epochs can be seen on table 3 experiment No. 4. For all tests, $\mathbf{M_A}$ is trained on $\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$ and $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ and tested on $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ and $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$. For 400 epochs, the AUC score is significantly better than the one for 100 epochs (the same can not be concluded for 400 vs 50 epochs, as the intervals barely overlap). As seen on the 50 epoch experiment, the MIA is still successful when $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ has trained for 50 epochs and results in comparable performance to the tests against target models trained for 400 epochs. The baseline appears to perform equally well across the different training times and thus does not appear to gain extended knowledge of the underlying training data distribution with increasing training time.

A Mix of DTU and LFW in the Target Dataset. Here $\mathbf{M_A}$ is trained using $\mathbf{D}^{\mathbf{gen}}_{\mathbf{AAU}}$ as negatives and a mix of DTU and LFW as positives: $\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU+LFW}}$ (balanced training with 2,500 images from each set). Note that only $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ are test positives. The result shown on table 3 experiment No. 13 shows that $\mathbf{M_A}$ still performs a successful MIA even if only half of the data used to fine-tune $\mathbf{M_T}$ is of interest. It is not significantly worse than when the target model is only fine-tuned on $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$, which was the case in experiment No. 4 (0.83 ± 0.02 vs 0.86 ± 0.02). While much worse than $\mathbf{M_A}$, the baseline model is also able to perform a successful MIA in this test with close to the same performance when compared to test No. 4.

A Different Prompt for Target Model Inference. has influence on the generated output. The result of using the prompt `"a profile picture"` instead of `"a dtu headshot"` for inference to generate $\mathbf{D}^{\mathbf{gen}}_{\mathbf{DTU}}$ used as training positives for $\mathbf{M_A}$ is shown on table 3 test No. 5. The performance of the attack model is not significantly different when conditioning the target model with `"a profile picture"` for infer-

ence compared to table 3 test No. 4 (0.82 ± 0.05 vs 0.86 ± 0.02). This could indicate that the generated distribution still contains the features which allow the Resnet-18 to make good predictions. An explanation could be that the 400 epochs of fine-tuning on the "dtu" label and images has made the latent space more uniform. This is supported by the fact that on fig. 8 the image with guidance scale $s = 0$ still produces a headshot, even though according to eq. (3) it is unconditioned.

$$\varepsilon_t = \varepsilon_{t,uncond} + s \cdot (\varepsilon_{t,\tau(y)} - \varepsilon_{t,uncond}) \quad (3)$$

Recognising Seen vs Unseen Samples from the Same Data Distribution. The MIA on $\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{DTU}}$ performed poorly as seen on table 3 test No. 8. This shows that the MIA presented in this paper does not seem successful in the task of identifying membership on an individual basis (for a single data-point), but instead is viable for the task of inferring the membership of a dataset as a whole. These results indicate some sort of shared characteristic among the collection of images. The nature of this shared characteristic is unknown. It is a reasonable assumption that images taken at the same locations/universities/organisations share some features.

Using Watermarks for MIA Enhancement. can be seen on table 3 test No. 11 and 12. The visible watermark tested in $^{\mathbf{wm}}\mathbf{D}^{\mathbf{seen}}_{\mathbf{DTU}}$ vs $\mathbf{D}^{\mathbf{unseen}}_{\mathbf{AAU}}$ were very effective and lead to a near perfect classification of the test set by the attack model. The use of a hidden watermark however did not show any improvement compared to using no watermarks (cf. table 3 test No. 12 vs No. 4) which shows that either the target model did not learn to mimic the hidden watermark, or that the attack model was not able to pick up on the nearly invisible watermark. An example of the watermarks is shown on fig. 6.

(a) visible watermark (b) 25% visible (c) 1% visible

Figure 6: (a) Shows an example of the visible watermark. (b) illustrates the shape of the hidden watermark. (c) shows the actual hidden watermark used for testing.

MIA Performance for Different Guidance Scales. can be seen on fig. 7, the performance of a MIA in our case is sensitive to the guidance scale used when generating training positives. The AUC score achieved by $\mathbf{M_A}$ is highest when the guidance scale is between 4 and 12. Looking at fig. 8, we see that even with a guidance scale of $s = 0$, $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ still generates a headshot, despite having no guidance of what to generate. This indicates that fine-tuning $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ for 400 epochs has introduced enough bias that it assumes noise to stem from images of headshots. For $s = 16$ visible artifacts appear, distorting the face (fig. 8e) which might also explain the drop in AUC score. It is also notable that for $s = 0$, $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ achieves an AUC of ~ 0.7 even though it is trained on positives generated without any guidance, which demonstrates that the target model $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ trained on 400 epochs leaks information of its training data even when not conditioned on a prompt.

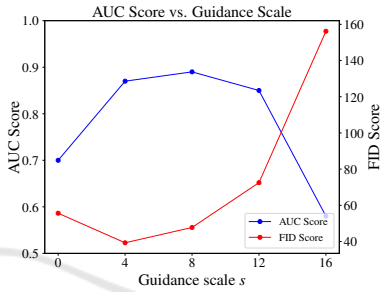


Figure 7: AUC scores for the test of $\mathbf{M_A}$ where the training positives have been generated with different guidance scales. The FID score compared to the original DTU data is also shown, a higher AUC score correlates with a lower FID score. This is consistent with (Carlini et al., 2023).

(a) $s = 0$ (b) $s = 4$ (c) $s = 8$ (d) $s = 12$ (e) $s = 16$

Figure 8: Examples of images generated by $\mathbf{M}^{\mathbf{DTU}}_{\mathbf{T}}$ using different guidance scales.

6 CONCLUSION

It is concluded that it is possible to perform Membership Inference Attacks on Latent Diffusion Models fine-tuned on face images. The attack proposed is restricted to the task of inferring membership on a dataset level, and is not successful in the task of inferring which specific images of a dataset are used for fine-tuning a target model.

The circumstances or prerequisites for the cases where the MIA is successful are investigated and multiple factors are considered in relation to its performance. It is found that using a generated auxiliary dataset leads to a significantly better performance for the MIA than using real images. Diluting the members of interest by mixing two datasets in the fine-

tuning of the target model did not lead to significantly worse performance of the MIA. It should however be noted that this was only tested by reducing the share of target members to 1:2.

It is found that by introducing visible watermarks to the target dataset, our MIA sees a significant boost in performance. Using hidden watermarks was not found to have a positive impact on the performance of the MIA. No significant effect was found when investigating the influence of the relationship between the labels used for fine-tuning the target model and the prompt used for inference, i.e. whether they match or not. Upon investigating the importance of the guidance scale used by the target model, it is found to have a significant influence on the performance of our MIA, with best performance at $s \sim 8$.

Overall the proposed MIA is a realistic and feasible attack in a real-life application. However, it is computationally expensive to fine-tune a generative "shadow model" for the task of producing an auxiliary dataset related to the domain of interest as well as training the Resnet-18 attack model. The nature of the tests performed restricts our conclusion to the case of LDMs fine-tuned on face images. The smallest amount of fine-tuning that was still found to be effective was 50 epochs on the member images (however it could be lower - as it was not tested). The only LDM used for testing was Stable-Diffusion-v1.5 (Rombach et al., 2022), which limits the generalizability of the conclusions drawn. The approach using a Resnet-18 as an attack model is found to be generally stable on several different hyperparameters in the target LDM. In conclusion, the method for Membership Inference Attack shown in this paper is realistic and could be used as a tool to infer if one's face images have been used to fine-tune a Latent Diffusion Model in a black-box setup.

REFERENCES

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. *ArXiv*, abs/2301.13188.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. (2020). Ganleaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20. ACM.

Dubi'nski, J., Kowalczuk, A. D., Pawlak, S., Rokita, P., Trzci'nski, T., and Morawiecki, P. (2023). Towards more realistic membership inference attacks on large diffusion models. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4848–4857.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models.

Zhang, M., Yu, N., Wen, R., Backes, M., and Zhang, Y. (2023). Generated distributions are all you need for membership inference attacks against generative models.