Synthetic and (Un)Secure: Evaluating Generalized Membership Inference Attacks on Image Data

Pasquale Coscia[1][a], Stefano Ferrari[1][b], Vincenzo Piuri[1][c] and Ayse Salman[2][d]
[1]*Department of Computer Science, Università degli Studi di Milano, via Celoria 18, Milano, Italy*
[2]*Department of Computer Engineering, Maltepe University, 34857 Maltepe, Istanbul, Turkey*

Keywords: Membership Inference Attack, Generative Models, Fréchet Coefficient.

Abstract: Synthetic data are widely employed across diverse fields, including computer vision, robotics, and cybersecurity. However, generative models are prone to unintentionally revealing sensitive information from their training datasets, primarily due to overfitting phenomena. In this context, membership inference attacks (MIAs) have emerged as a significant privacy threat. These attacks employ binary classifiers to verify whether a specific data sample was part of the model's training set, thereby discriminating between member and non-member samples. Despite their growing relevance, the interpretation of MIA outcomes can be misleading without a detailed understanding of the data domains involved during both model development and evaluation. To bridge this gap, we performed an analysis focused on a particular category (i.e., vehicles) to assess the effectiveness of MIA under scenarios with limited overlap in data distribution. First, we introduce a data selection strategy, based on the Fréchet Coefficient, to filter and curate the evaluation datasets, followed by the execution of membership inference attacks under varying degrees of distributional overlap. Our findings indicate that MIAs are highly effective when the training and evaluation data distributions are well aligned, but their accuracy drops significantly under distribution shifts or when domain knowledge is limited. These results highlight the limitations of current MIA methodologies in reliably assessing privacy risks in generative modeling contexts.

1 INTRODUCTION

Over the past decade, Artificial Intelligence (AI) has advanced rapidly, driven by algorithms that extract patterns from large datasets to enable generalization. Progress in data processing and the emergence of novel Machine Learning (ML) paradigms have further accelerated its development and commercial adoption. Currently, concerns have arisen about the security, ethics and legality of data collection and processing. Since personal data often contains sensitive information, increasing attention has been paid to its use in training ML models and the potential risk of information leakage. This concern is part of a larger area of study focused on protecting machine learning services and understanding how attackers might misuse or disrupt a model's intended purpose. Threats that can affect an ML model development pipeline include, among others, adversarial attacks (Rosenberg et al., 2021) and data leakage (Niu et al., 2024).

Membership Inference Attacks (MIAs) (Shokri et al., 2017; Bai et al., 2024) aim to determine whether a specific sample was part of a model's training set. These attacks are studied for various reasons, including malicious data exfiltration, security auditing, and enhancing model robustness (Liu et al., 2022). As a result, MIAs have attracted significant research interest in recent years (Truong et al., 2025). MIAs exploit the tendency of overparameterized models to memorize training data, leading to better performance on seen (member) samples than on unseen (non-member) ones. They are generally categorized into white-box attacks, where the adversary has access to the model's parameters and training details, and black-box attacks, where only input-output interactions with the model are possible. MIAs can be conducted using two main approaches. The first involves training shadow models to mimic the target model, followed by training a binary classifier to discrimi-

[a] https://orcid.org/0000-0003-4726-3409
[b] https://orcid.org/0000-0002-4982-6212
[c] https://orcid.org/0000-0003-3178-8198
[d] https://orcid.org/0000-0003-2649-3061

Coscia, P., Ferrari, S., Piuri, V. and Salman, A.
Synthetic and (Un)Secure: Evaluating Generalized Membership Inference Attacks on Image Data.
DOI: 10.5220/0013657700003979
In *Proceedings of the 22nd International Conference on Security and Cryptography (SECRYPT 2025)*, pages 287-297
ISBN: 978-989-758-760-3; ISSN: 2184-7711

nate between member and non-member data based on model behavior. The second approach uses metrics derived from the target model's outputs, classifying inputs by comparing these values against predefined thresholds. An important class of MIAs targets image-based generative models, a family of machine learning models that generate samples from a distribution of images, typically conditioned on some input information. A notable example within this category is text-to-image models, which generate images based on textual descriptions (Rombach et al., 2022). Several MIA schemes exploit the differences between the generated sample (which carry information from the training dataset) and similar data that have not been used for training the model (Shokri et al., 2017; Zhang et al., 2024). The availability of well-defined member and non-member datasets critically influences the feasibility and evaluation of MIAs, as the attacker's binary classifier must reliably discriminate between the two classes while maintaining robustness to data variability.

Existing studies (Zhang et al., 2024) often overlook a systematic evaluation of the semantic and quality alignment between member and non-member datasets, collecting data without assessing class-level similarity or content quality, which may compromise the validity of MIA performance evaluations. To address this limitation, we introduce novel strategies for dataset construction that leverage image content and paired textual descriptions. Specifically, we introduce a Top-k selection approach to systematically control the degree of overlap between datasets. We then analyze the relationship between dataset similarity and MIA effectiveness by adopting established techniques from the literature to measure similarity at both the image and dataset levels using the Fréchet Coefficient (FC). Finally, we replicate a standard MIA framework on generative models to assess how the quality and selection of data influence attack success. Our results indicate that MIAs are most effective when datasets show high distributional similarity, whereas their reliability decreases significantly under distribution shifts or limited knowledge scenarios.

The remainder of this paper is organized as follows. Section 2 presents the context of privacy and security in ML-based systems. Section 3 presents the general scheme of MIAs in generative models, discusses related issues, and describes the problem of evaluating the degree of similarity between images along with the proposed procedure for this task. Section 4 outlines our case study and the selected member/non-member datasets. Section 5 describes the experimental activity and discusses the obtained results. Finally, Section 6 summarizes our outcomes.

2 RELATED WORK

Membership inference attacks (MIAs) can be applied across diverse model types, such as classification, generative, regression, and embedding models. In the following, we survey the relevant literature exploring these different contexts.

Classification Models. Classification models are typically categorized into binary and multi-class classifiers. (Shokri et al., 2021) investigated binary classifiers that assess membership inference attacks using feature-based model explanations across datasets, showing that backpropagation through explanations can leak significant information about training data. Additionally, they empirically explore the balance between privacy and explanation quality by analyzing perturbation-based model explanations. Multi-class classifiers are explored in (Shokri et al., 2021), (Long et al., 2018), (Long et al., 2020), and (Salem et al., 2019). They explored various datasets and attack strategies on the training data of shadow models, using a shadow model trained on data from the same distribution as the target model to mimic its behavior. Similar MIA success rates are achieved even when the attacker's data comes from a different distribution.

Generative Models. (Liu et al., 2019) introduced co-membership attacks targeting deep generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). The attack assumes knowledge of whether a given bundle of samples was part of the training set. The authors find that co-membership attacks are more powerful than single attacks, and VAEs are more prone to attacks than GANs. (Duan et al., 2023) proposed a gray-box attack, where the attacker can access to the loss function of a given sample at a given time-step of the diffusion process. The attack can be mitigated by applying data augmentation techniques during the target model's training. (Carlini et al., 2023) studied black-box MIAs on generative models by issuing extensive queries to detect memorized data. (Zhou et al., 2022) proposed to extract properties of the training set, such as the statistical distribution of sample groups, by querying the target model and analyzing the distribution of the generated outputs. Although the approach is designed for GANs, it can be adapted to other generative models. In (Hu and Pang, 2021), various levels of knowledge about the target model have been explored for MIA on GANs. The proposed attack, which requires partial knowledge of the training dataset and the discriminator, is categorized as a gray-box attack. (Chen et al., 2020) studied MIAs against

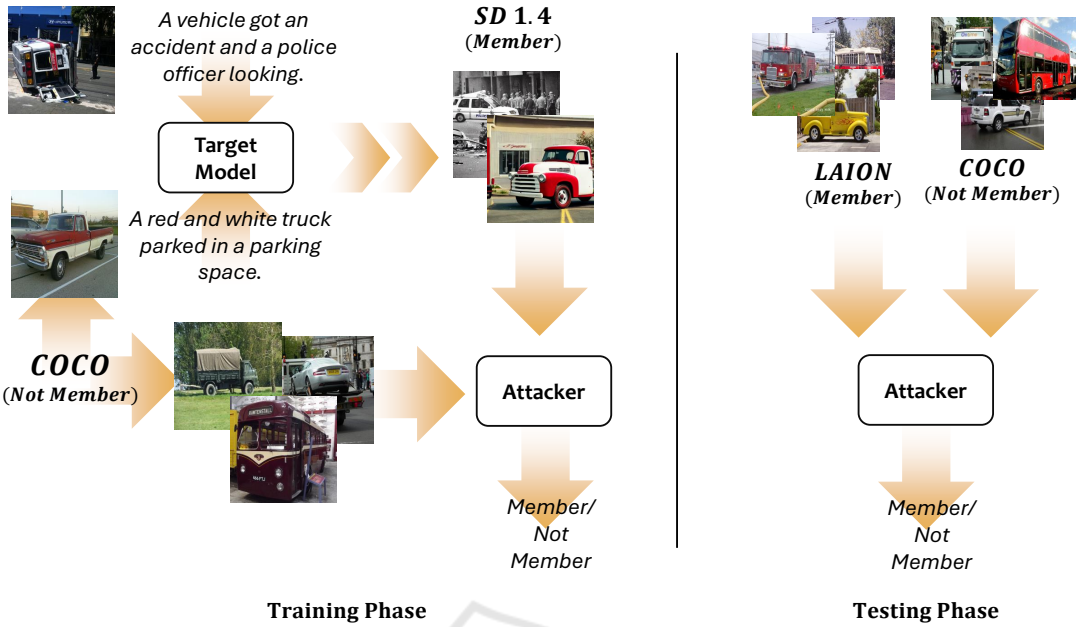


Figure 1: Attack scheme used as case study. The captions of non-member images (COCO) are used to generate the positive examples (SD-1.4) for training the attack model by querying the target model. The trained attack model is then tested using LAION images as positive examples.

GANs in several scenarios, in which the various elements of the model are exposed at different levels. For each of the proposed attack methods, empirical tests to evaluate their effectiveness are suggested. (Azadmanesh et al., 2021) focused on GANs assuming that a given number of member images are known and proposes a method to discriminate them among a larger group that includes non-member images. The attack is built on the training of an autoencoder the decoding part of which is played by the generator of the target model, while the encoder is trained to output the latent code of the given images. The discrimination of the member samples is then realized by analyzing the loss of the autoencoder. (Zhang et al., 2024) proposed a black-box MIA applicable to any generative model. This approach involves sampling the generated distribution and using a set of non-member samples that share the same distribution as the member samples of the target model. The underlying assumption is that the generated samples implicitly encode information about the training set. A classifier can then be trained to discriminate between the generated samples and the auxiliary images, ultimately enabling the discrimination between member and non-member samples. M4I (Hu et al., 2022b) proposed an MIA-based shadow model on image-to-text models. The discrimination is on the basis of the matching between the input-features and the corresponding output-features.

Neural Network Embeddings. Graph embeddings (Duddu et al., 2021) are also studied to determine whether a graph node, representing an individual user's data, is part of the model's training set. Image encoders represent another important class of models. (Liu et al., 2021), for example, proposed a membership inference method against image encoders pre-trained by contrastive learning. Their black-box approach aims to identify whether an input is part of the training set of the encoder. Using a custom encoder, EncoderMI, which exploits overfitting to its training data, the method detects when two augmented versions of an input (from within or outside the training set) yield more or less similar feature vectors. This approach can be used (i) by a data owner to audit unauthorized use of public data in pre-training an image encoder, or (ii) by an attacker to compromise the privacy of sensitive/private training data.

Several MIAs have been studied for different architectures, e.g., deep learning regression (Gupta et al., 2021) and Natural Language Processing (NLP) models (Song and Raghunathan, 2020).

3 METHODOLOGY

MIA attacks on image-based generative models are often based on the possibility of discriminating between generated and original images. Figure 1 illustrates our attack model, which builds upon the framework proposed in (Zhang et al., 2024). This black-box attack scheme involves collecting an auxiliary dataset

by generating images using the model under attack. The attack model is then a classifier trained to discriminate between images from the auxiliary dataset and images from a similar dataset that was not used during the model's training (e.g., a dataset collected after the model's release). The underlying rationale is that the generated images capture details of the membership dataset, enabling the classifier to effectively discriminate between member and non-member images. This attack can be considered an almost black-box approach, as it mainly relies on the knowledge of the datasets. Specifically, it does not assume access to the member dataset, but only to the non-member dataset. Additionally, unlike shadow model-based schemes, it does not require significant computational resources, though it does necessitate black-box access to the target model.

In this work, we focus on a text-to-image target model. The attack begins by training a classifier using a carefully selected non-member dataset (negative examples). Positive examples for the attack model are generated by querying the target model with captions corresponding to the non-member images, resulting in a set of generated images whose content closely resembles that of the non-member images. Despite their similarity, these generated images still retain traces of the model's original training data. For this reason, the classifier trained to discriminate between negative and positive examples should, to some extent, also be able to differentiate between member and non-member images, as the generated images indirectly reflect information from the training set.

For our experiments, we consider Stable Diffusion v1.4 (Rombach et al., 2022) as the target model, an open-source latent diffusion model trained on the LAION-aesthetics v2 5+ dataset (LAION, Large-scale Artificial Intelligence Open Network, 2022). This model is widely used due to its ability to run and be fine-tuned on consumer-grade PCs equipped with GPUs.

Empirical approaches such as this require careful consideration during validation experiments, as the classifier may learn to focus on features that effectively distinguish the two datasets but do not directly relate to their membership in the target model's training set. For example, if the two datasets belong to distinct domains (e.g., animals vs. vehicles), the attack model may simply learn to discriminate between these domains, rather than between training set membership. Furthermore, when the datasets share similar content, the success of the attack can be evaluated based on the degree of overlap at the content level between the datasets that are used to construct the attack model. To this end, we propose a Top-k approach to categorize the collected images into distinct sets, for which we assess both the degree of overlap and the attack's performance.

3.1 Evaluation Metrics

In the following, we detail our procedure to evaluate the similarity between images and datasets, respectively, and the attack performance.

Images Similarity Evaluation. The similarity between two images can be assessed at various levels. At the pixel level, the comparison is based on the values of selected pixels or pixel subsets in both images. In contrast, at the feature level, the comparison is made using the features extracted from the images. While pixel-level methods are effective for evaluating the (quasi-)identity of two images, feature-level methods are generally more computationally efficient and better suited for handling large datasets.

In this study, we consider the cosine similarity method, which operates at the feature level and is commonly used in computer vision tasks. Using features enables comparison of images based on their content, regardless of the actual pixel layout.

Given two images I_a and I_b, their features $\phi(a)$ and $\phi(b)$, respectively, are extracted and their distance, $d(I_a, I_b)$ is computed as:

$$d(I_a, I_b) = \frac{\phi(a) \cdot \phi(b)}{||\phi(a)|| ||\phi(b)||} \quad (1)$$

Datasets Similarity Evaluation. In addition to image similarity, the comparison can also be extended to datasets of images. In this regard, the Fréchet inception distance (FID) (Dowson and Landau, 1982) is a metric used to measure the similarity of two multivariate normal distributions and can be used to asses the quality of images created by a generative model (Heusel et al., 2017). While the comparison between image distributions can be in general very complex, FID can be used to measure the similarity between the feature distributions of two datasets of images by comparing their multivariate Gaussian distributions. A lower FID score indicates that the two sets are similar. However, one of the drawbacks of this method is the lack of an absolute reference system. In fact, FID score can be used only to measure the relative similarity of two sets. Two FIDs are comparable only if they are computed on a common dataset.

Recently, a new metric based on FID, the Fréchet Coefficient (FC) (Kucharski and Fabijańska, 2025), has been proposed. FC score improves the FID because its value can be only in the range $[0, 1]$. This makes the interpretation of the resulting score easier, since higher values accounts for a higher similarity,

being score 1 the complete overlapping of the two distributions.

Attack Performance Evaluation. The evaluation of the effectiveness of an MIA can be realized using several metrics. Among them, the average-case accuracy (or other aggregate metrics, such as AUC) is often used (Hu et al., 2022a). However, in (Carlini et al., 2022), its suitability is questioned and another metric, the True Positive Rate at a low False Positive Rate (TPR @ low FPR) is considered. The rationale is rooted in the nature and objectives of an MIA, as even if the majority of training set members cannot be identified, the attack can still be deemed successful if a small minority is identified with high confidence. If the success of a method can be measured in terms of TPR, its reliability can be measured in terms of FPR. Hence, the TPR when FPR is low can be assumed as the right metric to evaluate the success of an MIA technique. For the above reasons, we use the True Positive Rate at low False Positive Rate (TPR at low FPR) (Carlini et al., 2022) metric for measuring the performance of the attack.

4 CASE STUDY

Datasets Selection. To select images with content similar to a target, candidate images for our datasets are evaluated using an image classifier. The classifier assigns a probability measure, $p(c_i)$, to each class c_i in the set $C = \{c_i\}$, where C represents the set of all the considered classes by the classifier. Since the target content may correspond to several classes within C, a suitable subset of target classes, $C_t \subset C$, is defined. Images are considered relevant if at least one of their five most probable classes belongs to C_t; otherwise, they are discarded. The selected images are further categorized based on the reliability of their content classification into six subsets: from Top-0 to Top-5. An image is assigned to the Top-k set if at least k of the five most probable classes belong to C_t. For example, Top-0 images are those where at least one target class appears in the top five, but a non-target class is the most probable. In contrast, a Top-3 image has at least the first three most probable classes belonging to C_t while the other two can belong to non-target class. All of the selected images are intended to be used in the attack experiment, but finer-grained content-based classification helps in assessing the suitability of the image similarity measure. The image classifier employed in this work is ResNet34 (He et al., 2016) pretrained on ImageNet-1k, which provides probability estimates across 1000 classes.

Non-Member Dataset. The non-member dataset for the experiments is obtained from the MS-COCO (Lin et al., 2014) dataset. This dataset provides images with several additional information that make it suitable to develop and test several image-based applications (e.g., segmentation, recognition, captioning). In particular, COCO images are provided with a textual description and object segmentation of the scene. It is composed of more than 200k labeled images with 80 object categories. Images having objects belonging to terrestrial vehicles (bus, car, motorcycle, train, and truck) are selected as candidates and further filtered using the procedure described in Section 4. Hereafter, the selected images set will be referred as COCO.

Member Dataset Since the target model of the case study is Stable-Diffusion 1.4, the LAION-aesthetic v2 5+ dataset (LAION, Large-scale Artificial Intelligence Open Network, 2022) will be used as the candidate member set. It is not feasible to fully download the LAION dataset using standard processing resources, since it comprises nearly 1 billion web images. In fact, the dataset is provided as URLs paired with textual descriptions. However, for the purposes of this paper, ground truth member images are only required for testing the attack model. Therefore, the following sampling and selection procedure is implemented to collect the member dataset. Images are downloaded from their URLs (if still available), and only those with textual descriptions sufficiently similar to the descriptions of the selected non-member images are retained. The similarity between textual descriptions is quantified using cosine similarity (1) applied to the CLIP (Radford et al., 2021) embeddings (512-dimensional features). We collect two versions of this dataset, corresponding to cosine similarity thresholds of 0.50 and 0.85.

To generate synthetic images, we use Stable Diffusion (SD) v1.4 as our target generative framework. This model generates images from textual descriptions, provided in our case by the COCO annotations. Specifically, we employ a version of SD pretrained on the LAION-2B-en dataset and fine-tuned on the LAION-aesthetics v2 5+ dataset. For the inference process, the number of steps is set to 150, while the guidance scale factor to 7.5, with no negative prompt used. The generated images have a resolution of 512×512 pixels. To ensure a sufficient quantity of images within a reasonable time frame, we attempt to generate three synthetic images per COCO annotation. Then, the ResNet34-based filtering mechanism is applied to each image; images that meet the required "terrestrial vehicular"-content criterion are assigned to one of the Top-k categories, while those not meeting the criterion are rejected. Furthermore,

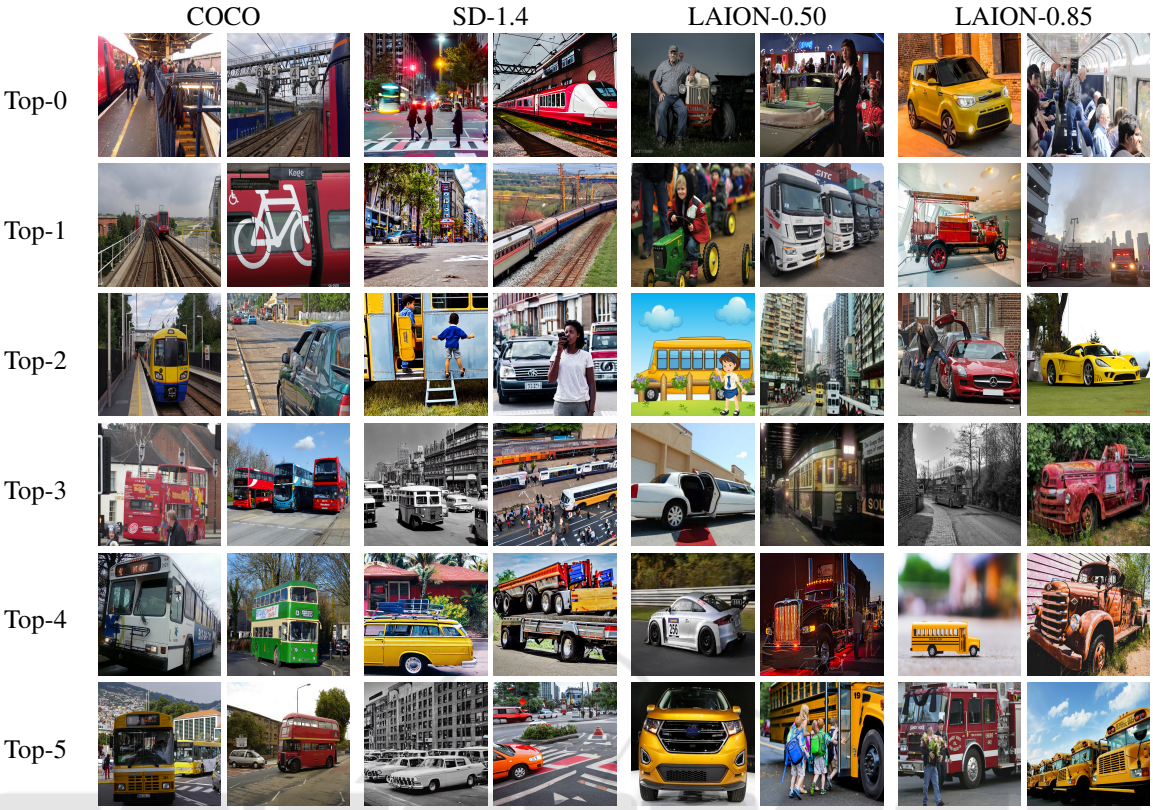


Figure 2: Examples of the selected images. The columns correspond to four datasets (COCO, SD-1.4, LAION-0.50, LAION-0.85), while the rows represent images categorized according to their reliability level in content classification (Top-0 to Top-5, as detailed in Section 4).

Table 1: Terrestrial vehicles categories from ImageNet-1k used in our data selection procedure as target classes, C_t.

Class ID	Description	Class ID	Description	Class ID	Description
407	ambulance	654	minibus	817	sports car
436	station wagon	656	minivan	829	tram
468	taxicab	675	moving van	864	tow truck, tow car, wrecker
511	convertible	705	passenger car	866	tractor
555	fire engine, fire truck	717	pickup, pickup truck	867	trailer truck, tractor trailer
569	garbage truck	734	police van	874	trolleybus
586	half-track	751	race car		
627	limousine	779	school bus		

for each annotation, the filtering process is limited to a maximum of 25 attempts. If none of these attempts yields an acceptable image, the annotation is skipped.

We refer to the previous datasets as LAION-0.50, LAION-0.85, and SD-1.4. Figure 2 shows some random samples of these datasets. They are organized column-wise by the belonging dataset and row-wise by the Top-k subset. Although all the images show a content that is clearly related to the target class (terrestrial vehicles), a slight improvement in the coherence of the scene appears moving from Top-0 to Top-5 subsets.

5 EXPERIMENTS

In the following, we present our findings on dataset similarity and the effectiveness of the attacks.

Datasets Similarity. According to the procedure outlined in Section 4, we select the non-membership images from COCO. Table 1 lists the ImageNet categories corresponding to the macro-category of terrestrial vehicles, which has been selected for the purpose of this study. Then, we partitioned the images into six subsets (Top-0 to Top-5), ranking them in increasing

Table 2: Size of the subsets of the selected datasets.

	COCO	SD-1.4	L-0.50	L-0.85
Top-0	6323	11090	11143	2377
Top-1	5743	9983	10559	2247
Top-2	4022	5733	7146	1504
Top-3	3005	3985	3973	947
Top-4	2253	2953	1653	513
Top-5	1608	2203	800	296

order of coherence with the target content ("terrestrial vehicles"), denoted as C_t. Table 2 reports the sizes of the datasets and their six subsets, where LAION is shortened as L for compactness.

The reliability of the attack model should be tested on a dataset that is similar to the training images of the target model. For this reason, we select image datasets that depict scenes belonging to the macro-category of terrestrial vehicles. Although the images may be qualitatively considered similar, assessing quantitatively their similarity is important both methodologically and experimentally. The similarity between sets can be evaluated by comparing the distribution of their features. In particular, the FC (see Section 3.1) compares mean and covariance matrices of two distributions and outputs a score in $[0, 1]$, where 0 is completely dissimilar and 1 means that the distributions completely overlap. The features of the datasets have been computed using the InceptionV3 embedding (2048 features). Since the cardinality of the Top-5 subsets is smaller for most of the considered datasets, the computation of the covariance matrix of 2048 features can be unreliable. To make the estimate of the similarity more robust, the PCA has been applied on the features and only the first 20 principal components have been used in the computation of the FC score.

To assess the reliability of this method, we computed the FC score between random splits of the same dataset. The results computed on ten splits are reported in Figures 3(a)–(d), for COCO, SD-1.4, LAION-0.50, and LAION-0.85, respectively. For all the datasets, the values are very close to 1, meaning that the sets are coherent and their images span a narrow neighborhood in the features space. The FC is computed on the Top-k ($k \in \{0, ..., 5\}$) subsets, without substantial digression from the unity. This assesses the validity of the selection procedure. Only for the LAION-0.85 dataset, the similarity index appears to be affected by the finer classification. For this dataset, a noticeable decrease in the FC for Top-4 and Top-5 is observed, likely due to the small number of elements in these subsets. However, the FC remains close to 1. In fact, for reliably selected images, the position and number of target classes within the five most probable positions do not substantially affect the measured FC score.

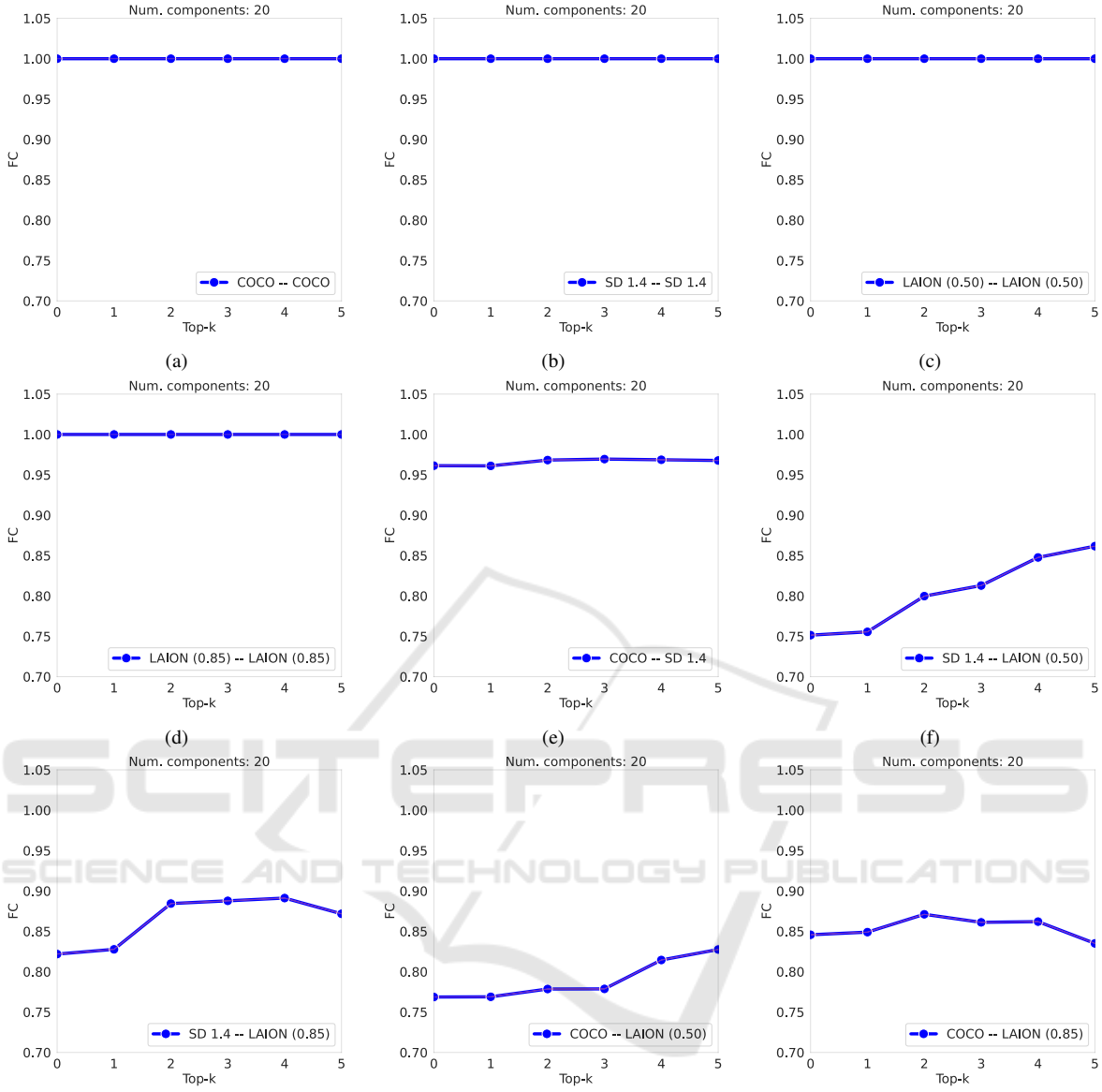
For all the comparisons (panels (a)–(i)) the FC score is quite constant, which can be interpreted that the subsets share the same distance and dispersion in the feature space, with the exception of the SD-1.4 vs. LAION-0.50 case, where the FC score ranges from 0.72 and 0.85. A similar trend can also be noticed in COCO vs. LAION-0.50 (panel (h)), although with a smaller span ($[0.75, 0.82]$). However, the scores reflect a large overlap in the features space and the trend may be due to the smaller threshold used in the selection of the images, which could have allowed scenes in which the vehicles are not the main element, especially for the lowest values of k.

In panel (e), the set COCO and SD-1.4 (that is the non-member images and the images generated from their captions) are compared. The FC is smaller than 1, but larger than 0.95, confirming that their contents are related. The panels (f)–(i) report the comparison between COCO or SD-1.4 and the member sets, LAION-0.50 and LAION-0.85. Although the FC values are large (never below 0.7), they are noticeably smaller than the FC values of COCO vs. SD (in panel (e)), which can be a sign of a slight difference between the member and not member images.

Besides, we also tested the similarity of COCO vs. the union of the others (SD-1.4, LAION-0.5, and LAION-0.85), which resulted in an FC score slightly larger than 0.9, coherently with the similarity indices for the pairs of sets (panels (e), (h), and (i) of Figure 3).

Attack Performance. We follow the MIA scheme proposed in (Zhang et al., 2024) which serves as a case study to evaluate the influence of dataset similarity on the attack model's performance. To this end, a classifier based on the ResNet50 model (He et al., 2016) is fine-tuned to discriminate between member and non-member images. For each training epoch, the model is tested on the test dataset, and the best performance is selected. Although this procedure may be methodologically questionable for obtaining an unbiased evaluation of the final classifier, it is chosen to provide an optimal case for assessing the attack scheme's performance.

Several settings test different datasets for training and testing. In all cases, we use COCO as the non-member dataset, and split all sets into 80-20% proportions for training and testing. To increase confidence in the image content, we use the Top-1 subsets. The attacks have been evaluated in terms of accuracy, AUC, and TPR@0.2%FPR (Section 3.1)



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Similarity of the datasets used in the attack. The similarity is computed as the feature correlation (FC) on the first 20 PCA components of the image embeddings (features) for all subsets with increasing content scores. In all comparisons (panels (a)–(i)), the FC score remains relatively constant, indicating that the subsets are quite similar to each other in the feature space. Panels (a)–(d) show the similarity of random partitions of the same datasets, averaged over 10 random splits. Since the FC values are close to 1, the sets are coherent, and their images are highly similar. Panel (e) compares the COCO and SD sets (the non-member images and the images generated from their captions). The FC is slightly smaller than 1 but greater than 0.95, confirming that their contents are related. Panels (f)–(i) compare COCO or SD with the member sets, LAION-0.50 and LAION-0.85. Although the FC values remain large (never below 0.7), they are noticeably smaller than the FC values between COCO and SD (in panel (e)), suggesting a slight difference between the member and non-member images.

Table 3: Attack performances

Training images		Testing images		Evaluation metrics		
Non-Member	**Member**	**Non-Member**	**Member**	**Accuracy**	**AUC**	**TPR@0.2%FPR**
COCO	SD-1.4	COCO	LAION-0.50	52.7%	0.619	0.02
COCO	SD-1.4	COCO	LAION-0.85	75.6%	0.650	0.06
COCO	SD-1.4	COCO	SD-1.4	99.3%	1.00	0.97
COCO	LAION-0.50 LAION-0.85	COCO	LAION-0.50 LAION-0.85	92.7%	0.972	0.20

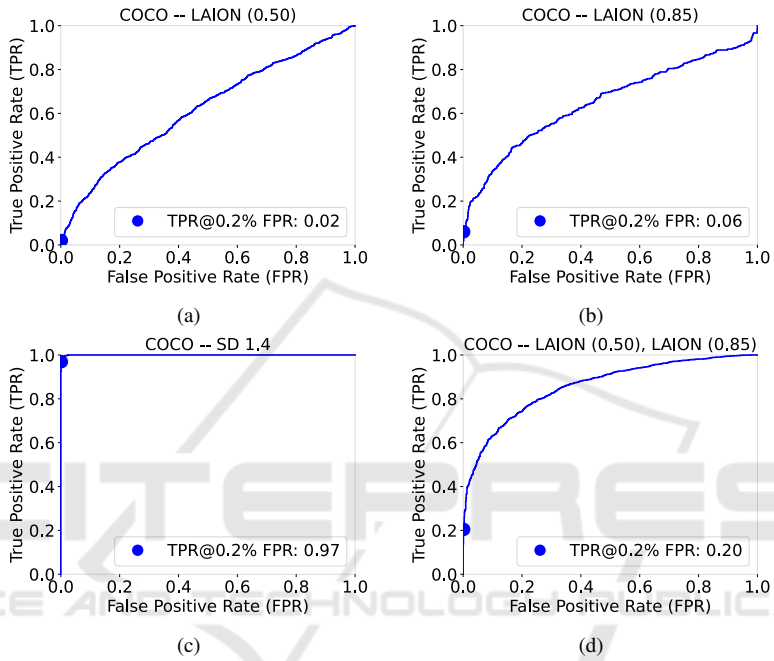


Figure 4: ROC of the attack models.

and the performances are reported in Table 3 and the ROC of the models are reported in Figure 4. In detail, we test the setup for the real attack scheme using SD-1.4 as positive examples and COCO as negative examples in the training, and test the models on LAION-0.50 or LAION-0.85 as positive examples, with COCO as negative examples. For both models, the performance is below expectations: in particular, the model tested on LAION-0.50 performs close to random choice. The performance of the model tested on LAION-0.85 is much better, achieving an average accuracy of 75.6%. However, the TPR@0.2% FPR is low for both models: 0.02 and 0.06. This suggests that the classifier is not effectively identifying the majority of member images, as only 2% and 6% are detected. Hence, MIAs do not provide a reliable prediction of membership.

We carried out two additional experiments to evaluate the goodness of the model and the effectiveness of the separability of the two datasets. In the first, the classifier is trained and tested on SD-1.4 and COCO. It achieves a good level of reliability, scoring 99.3% of accuracy and 0.97 of TPR@0.2%FPR. The ROC in Figure 4(c) shows a very sharp trend. The model shows that it is possible to reliably discriminate between the generated and genuine images, although their content is very similar.

In the second experiment, the classifier is trained and tested on the union of the LAION datasets as positive examples and COCO as negative examples. Although the accuracy reaches 92.7%, the TPR@0.2%FPR remains quite low, scoring 0.20. Also the ROC (Figure 4(d)) does not reach the sharpness of the previous model. This experiment corresponds to an MIA in a gray-box set-up, since there is the knowledge of a small subset of the training model. However, the TPR@0.2%FPR achieved does not allow to consider this attack as successful, since its low reliability. This partial failure of the MIA scheme is likely due to the large size of the target model's

training dataset. The small size of the selected subset may inadequately represent the richness of the entire dataset.

6 CONCLUSIONS

Membership Inference Attacks (MIAs) are a critical tool for evaluating the privacy protection properties of generative models. This paper investigates the influence of dataset similarity, i.e., the degree to which the attack dataset resembles the target model's training set, on the effectiveness of MIAs. A methodology for selecting images from the same content domain for both member and non-member datasets is proposed. Our results suggest that greater dissimilarity between the datasets may hinder the model's ability to attribute membership accurately.

Future work could involve selecting images from different macro-categories, enabling a broader sampling of the feature space. Additionally, further investigations may explore the use of diverse image classifiers for different tasks. Furthermore, all classifiers in the current study are pre-trained on the same dataset (ImageNet-1k), which could introduce bias into the evaluation process.

ACKNOWLEDGEMENTS

This work was supported in part by the EC under projects Chips JU EdgeAI (101097300) and GLACIATION (101070141), and by project SERICS (PE00000014) under the MUR NRRP funded by the EU - NGEU. Project EdgeAI is supported by the Chips Joint Undertaking and its members including top-up funding by Austria, Belgium, France, Greece, Italy, Latvia, Netherlands, and Norway. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, the Chips Joint Undertaking, or the Italian MUR. Neither the European Union, nor the granting authority, nor Italian MUR can be held responsible for them. We thank Federico Florio, for the preliminary study he performed during his Master thesis at Università degli Studi di Milano.

REFERENCES

Azadmanesh, M., Ghahfarokhi, B. S., and Talouki, M. A. (2021). A white-box generator membership inference attack against generative models. In *2021 18th International ISC Conference on Information Security and Cryptology (ISCISC)*, pages 13–17.

Bai, L., Hu, H., Ye, Q., Li, H., Wang, L., and Xu, J. (2024). Membership inference attacks and defenses in federated learning: A survey. *ACM Comput. Surv.*, 57(4).

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. (2022). Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

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. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA. USENIX Association.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. (2020). Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pages 343–362, New York, NY, USA. Association for Computing Machinery.

Dowson, D. and Landau, B. (1982). The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. (2023). Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, pages 8717–8730.

Duddu, V., Boutet, A., and Shejwalkar, V. (2021). Quantifying privacy leakage in graph embedding. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous '20, page 76–85, New York, NY, USA. Association for Computing Machinery.

Gupta, U., Stripelis, D., Lam, P. K., Thompson, P., Ambite, J. L., and Steeg, G. V. (2021). Membership inference attacks on deep regression models for neuroimaging. In Heinrich, M., Dou, Q., de Bruijne, M., Lellmann, J., Schläfer, A., and Ernst, F., editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 228–251. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.

Hu, H. and Pang, J. (2021). Stealing machine learning models: Attacks and countermeasures for generative adversarial networks. In *Proceedings of the 37th Annual Computer Security Applications Conference*, ACSAC '21, page 1–16, New York, NY, USA. Association for Computing Machinery.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang,

X. (2022a). Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s).

Hu, P., Wang, Z., Sun, R., Wang, H., and Xue, M. (2022b). M^4I: Multi-modal models membership inference. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1867–1882. Curran Associates, Inc.

Kucharski, A. and Fabijańska, A. (2025). Towards improved evaluation of generative neural networks: The Fréchet coefficient. *Neurocomputing*, 623:129422.

LAION, Large-scale Artificial Intelligence Open Network (2022). https://laion.ai/blog/laion-aesthetics/.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Liu, H., Jia, J., Qu, W., and Gong, N. Z. (2021). Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 2081–2095, New York, NY, USA. Association for Computing Machinery.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., and Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Trans. Intell. Syst. Technol.*, 14(1).

Liu, K. S., Xiao, C., Li, B., and Gao, J. (2019). Performing co-membership attacks against deep generative models. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 459–467.

Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C. A., and Chen, K. (2018). Understanding membership inferences on well-generalized learning models.

Long, Y., Wang, L., Bu, D., Bindschaedler, V., Wang, X., Tang, H., Gunter, C. A., and Chen, K. (2020). A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 521–534.

Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., Shen, Y., Jiang, X., Ma, J., and Zhang, Y. (2024). A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2(5):404–454.

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. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Comput. Surv.*, 54(5).

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. (2019). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*.

Shokri, R., Strobel, M., and Zick, Y. (2021). On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 231–241, New York, NY, USA. Association for Computing Machinery.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Song, C. and Raghunathan, A. (2020). Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA. Association for Computing Machinery.

Truong, V. T., Dang, L. B., and Le, L. B. (2025). Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Comput. Surv.* Just Accepted.

Zhang, M., Yu, N., Wen, R., Backes, M., and Zhang, Y. (2024). Generated distributions are all you need for membership inference attacks against generative models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4827–4837.

Zhou, J., Chen, Y., Shen, C., and Zhang, Y. (2022). Property inference attacks against gans. In *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society.