SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark

Rui Wen[1]* Yiyong Liu[2]* Michael Backes[2] Yang Zhang[2]†
[1]*Institute of Science Tokyo* [2]*CISPA Helmholtz Center for Information Security*

Abstract

Data reconstruction attacks, which aim to recover the training dataset of a target model with limited access, have gained increasing attention in recent years. However, there is currently no consensus on a formal definition of data reconstruction attacks or appropriate evaluation metrics for measuring their quality. This lack of rigorous definitions and universal metrics has hindered further advancement in this field. In this paper, we address this issue in the vision domain by proposing a unified attack taxonomy and formal definitions of data reconstruction attacks. We first propose a set of quantitative evaluation metrics that consider important criteria such as quantifiability, consistency, precision, and diversity. Additionally, we leverage large language models (LLMs) as a substitute for human judgment, enabling visual evaluation with an emphasis on high-quality reconstructions. Using our proposed taxonomy and metrics, we present a unified framework for systematically evaluating the strengths and limitations of existing attacks and establishing a benchmark for future research. Empirical results, primarily from a memorization perspective, not only validate the effectiveness of our metrics but also offer valuable insights for designing new attacks.

1 Introduction

The prosperous development of machine learning techniques has been witnessed in the past decade, which fertilizes its application in real-world scenarios. However, training a machine learning model for privacy-crucial tasks, such as person identification [6, 67], disease prediction [30], and financial risk prediction [11], demands a large volume of data which is not only valuable but also sensitive. As a result, model owners tend to release the model only, taking it for granted that the training data will not be leaked.

However, a series of works demonstrate that with limited access to a target model, the adversary is capable of inferring partial/complete information about the model's training samples. As a representative, membership inference attacks (MIA) [25, 26, 31–34, 46, 48, 61, 65, 71] denote a line of works that aim to infer whether a specific data sample is in the target model's training dataset. The disclosure of the membership status is a severe privacy breach as this information could indicate certain sensitive properties of the target sample.

A more challenging privacy attack is data reconstruction, which aims to recover the entire training dataset of a target model. This attack is considered the ultimate privacy breach, as it exposes all information about every sample. While membership inference (MIA) shares some similarities with data reconstruction, several key differences make data reconstruction a stronger attack. MIA is a sample-level attack, determining the membership status of individual samples, while data reconstruction is a dataset-level attack aimed at extracting the entire training dataset. From another angle, MIA is a decision problem, whereas data reconstruction is a search problem. In theory, MIA could aid in data reconstruction if the adversary has a large candidate dataset containing all the target samples and can perfectly predict membership. However, these assumptions are too strong, and to our knowledge, no one has attempted to use MIA for data reconstruction.

Owing to the crucial role of data reconstruction in the privacy domain, numerous attacks have been proposed in recent years. For example, Fredrikson et al. [19] propose the first data reconstruction attack, namely model inversion, which requires white-box access to the target model. Later, Yang et al. [66] relax this assumption by leveraging a training-based approach. Zhang et al. [73] adopt a Generative Adversarial Network (GAN) [21] to enhance the reconstruction quality. However, some researchers [41, 48] have claimed that model inversion can only recover a representative sample for each class of the target model, thus not an ideal data reconstruction attack. On the other hand, for certain training paradigms like federated learning [74] and online learning [44], some researchers have shown that they are able to reconstruct individual training samples. Moreover, all these works have used different types of evaluation metrics (see Section 4.1 for more

*The first two authors made equal contributions.
†Corresponding author

details).

Despite all the efforts, one of the major problems in the field of data reconstruction is that there does not exist a rigorous and unified definition for the attack. Moreover, the community has no consensus on what are the proper metrics for attack evaluation. This predicament roots in the diversity of the attack scenarios, e.g., various threat models and different training paradigms. At the same time, the lack of universal metrics exacerbates this problem since no existing metric is able to reflect all aspects of the reconstructions. We argue that without a clear-stated definition and metrics, it is hard to conduct further investigation in this direction.

In this paper, we take the first step in tackling this problem by 1) providing a definition of data reconstruction, 2) proposing a set of evaluation metrics, and 3) performing a large scale of experiments to establish a benchmark for further study. Specifically, our contributions can be summarized as follows:

- Definition: We recapitulate the reconstruction goal and attack information in existing work and formally provide an attack taxonomy. Based on this, we quantitatively and rigorously define the data reconstruction attack.
- Metrics: We formalize the desiderata for evaluation metrics and, based on this framework, propose two sets of metrics that address both macro and micro aspects of reconstruction, while emphasizing the importance of diversity in reconstruction quality. Additionally, we mitigate the limitations of using visualization as an evaluation tool by incorporating large language models (LLMs), with this metric specifically designed to assess high-quality reconstructions.
- Evaluation: We present a unified framework for data reconstruction attacks. Especially from the perspective of memorization, we conduct a thorough evaluation of ten reconstruction attacks under various attack scenarios. Our experiments demonstrate the effectiveness of the proposed metrics and provide a comprehensive analysis of existing attacks.

Implications: In this work, we propose the first rigorous definition of data reconstruction and develop a set of metrics. Our evaluation can serve as a benchmark for the current state-of-the-art approaches. We believe our results pave the way for further investigation into data reconstruction. We will share our code to facilitate research in the field in the future.

2 Background

2.1 Machine Learning Models

Machine learning algorithms aim to construct models that can accurately predict given inputs. These models are typically represented by a parameterized function $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ denotes the output space containing all possible predictions. To determine parameters θ that lead to optimal performance, a common approach is to minimize the following objective function using backpropagation:

$$\min_\theta \mathcal{L}(f_\theta(x), y)$$

where $\mathcal{L}$ denotes the classification loss, $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples used to train the target model, constituting the training dataset.

Recent research shows that given access to a trained target model f_θ, the adversary might exploit it through techniques such as membership inference [8, 26, 33, 34, 36, 37, 46, 48] and data reconstruction attacks [1, 19, 22, 44, 66, 68, 73, 74]. These attacks focus on inferring micro-level information about individual training samples, potentially leading to a direct privacy breach.

2.2 Data Reconstruction Attacks

The data reconstruction attack aims to recover the target dataset with limited access to the target model, with the aid of additional knowledge possessed by the adversary.

Existing attacks broadly fall into three categories: optimization-based, training-based, and analysis-based. In this paper, we thoroughly investigate ten representative reconstruction attacks, analyzing their performance and limitations. These attacks are briefly introduced below:

2.2.1 Optimization-based Attack

Most existing reconstruction attacks can be classified into this particular attack type. Such attacks aim to reconstruct the training dataset by iteratively optimizing the input until the desired class achieves a high likelihood score.

Recent attacks have incorporated generative models to enhance the quality of reconstruction, employing different architectural choices [13, 59] and loss functions [53]. We select seven representative attacks in our paper.

MI-Face [19]: Fredrikson et al. take the first step in reconstructing training samples from a trained model. Given a class y, the method first initializes a sample, and then updates x to maximize the likelihood/probability of belonging to that class, i.e.,

$$\min_x \mathcal{L}(f_\theta(x), y)$$

where $\mathcal{L}$ is the classification loss. Similar to training a machine learning model, MI-Face also uses backpropagating. But the difference is MI-Face focuses on optimizing x rather than the parameters θ in normal training. The backpropagation process demands white-box access to the target model, and the reconstructed sample always converges to

the most confident x near the initial point. Because x is high-dimensional, the reconstruction quality highly depends on the initialization.

DeepDream [1]: DeepDream was originally proposed to interpret machine learning models. However, this approach can also be utilized to improve MI-Face to acquire better results. The key idea is introducing regularization terms that force reconstructed samples to share similar statistics to natural images by penalizing each sample's total variance and ℓ_2 norm:

$$\min_x \mathcal{L}(f_\theta(x), y) + \alpha_{\text{tv}} \mathcal{R}_{\text{tv}}(x) + \alpha_{\ell_2} \mathcal{R}_{\ell_2}(x)$$

where $\mathcal{R}_{\text{tv}}$ and $\mathcal{R}_{\ell_2}$ denote total variance and ℓ_2 norm, respectively.

DeepInversion [68]: DeepInversion further improves DeepDream by adding another loss. Its intuition is that the batch normalization layers encoded statistical information about the training samples. Thus, minimizing the distance between reconstructed statistics and those stored in the target model helps.

Revealer [73]: Contrary to exploiting information encoded in the target model, Revealer leverages an auxiliary dataset to train a Generative Adversarial Network (GAN) G that generates samples x. Now instead of optimizing x, Revealer optimizes the input random seed z to G (and the reconstructed sample would be $x = G(z)$):

$$\min_z \mathcal{L}(f_\theta(G(z)), y)$$

The intuition is to utilize GAN to force the output $G(z)$ to always look 'real' on any optimized z, at the same time, guarantee high confidence in the prediction.

KEDMI [13]: KEDMI enhances the GAN-based approach in two primary ways. First, it optimizes the process of extracting knowledge from the auxiliary dataset by modifying the GAN's training objective. Specifically, it leverages labels assigned by the target model to the auxiliary dataset. The discriminator is trained to not only distinguish between real and fake samples but also to differentiate among the labels, enabling more nuanced learning. Second, instead of focusing on reconstructing single data points, KEDMI targets the reconstruction of the entire data distribution. To achieve this, it explicitly parameterizes the training data distribution and approximates the reconstructed distribution using its distributional parameters, such as the mean (μ) and standard deviation (σ).

PLGMI [69]: PLGMI decouples the search space by training a conditional GAN (cGAN). It utilizes pseudo-labels generated by a top-n selection strategy to steer the training process, combined with a max-margin loss, which collectively improves the effectiveness of the attack.

Deep-Leakage [74]: Deep-Leakage allows the adversary to access gradients (originally designed for federated learning systems) and aims to reconstruct training samples corresponding to the gradients. Specifically, the adversary randomly creates "dummy input" and "dummy label" and computes gradients based on this input-label pair. By optimizing this input-label pair to approximate true gradients, the "dummy input" and "dummy label" converges to target samples.

2.2.2 Training-based Attack

Inv-Alignment [66]: To overcome the limitation that data reconstruction requires white-box access to the target model, Yang et al. opt for a training-based approach that works with black-box access. Briefly, they construct an autoencoder with the target model as the encoder part. Once the autoencoder is well-trained, the decoder part can be leveraged to reconstruct inputs given corresponding posteriors.

Updates-Leak [44]: Updates-Leak considers the online-learning scenario where the adversary has access to different versions of the target model and tries to reconstruct samples used to update the model. The adversary trains numerous shadow models to mimic the updating procedure and leverage the posterior difference to reconstruct target samples.

2.2.3 Analysis-based Attack

Bias-Rec [22]: Haim et al. theoretically prove that the training data can be fully recovered given certain assumptions. In detail, if the target model is a homogeneous ReLU network and trained on a binary dataset using gradient flow, then *the parameters are linear combinations of the derivatives of the network at the training data points* [22]. According to this assertion, the adversary can derive training samples via optimization.

It is worth mentioning that the recent attack proposed by Balle et al. [5] has shown promising results in reconstructing the missing sample in a dataset. However, their attack assumes the adversary has the whole dataset except for the reconstructed one (in order to verify differential privacy properties), whose attack scenario significantly differs from common settings. Additionally, certain attacks [7, 29, 52] rely on invertible network architectures to execute their malicious actions. However, as these attack scenarios do not align with the scope of our investigation, we do not consider their attacks in this paper.

2.3 Data Reconstruction vs. Membership Inference

Membership inference attack (MIA) is another representative attack that aims to expose information about a training dataset, specifically by determining whether a target sample is part of it. There are two primary differences between data reconstruction and MIA, which necessitate the use of distinct attack methods. First, MIA takes a sample as input to

decide its membership status, whereas data reconstruction only has a target model. This difference impels data reconstruction to employ an incompatible attack approach, as the state-of-the-art MIA regularly involves training samples in the attack process. For example, Carlini et al. [8] train two sets of shadow models with datasets with/without the candidate sample. Such a training discrepancy is a crucial factor in enabling successful attacks.

From a different view, MIA can be framed as a decision problem while data reconstruction is a search problem. This distinction implies that data reconstruction is a more challenging and sophisticated attack. In an ideal scenario, perfect data reconstruction would reveal all membership statuses, as the entire training dataset would be exposed. MIA could serve as the foundation for data reconstruction, given the assumption that the adversary has a dataset that includes all training samples. However, in reality, this assumption is not feasible, and the adversary must try all possible pixel combinations iteratively, which is impractical due to the enormity of the search space. Additionally, no existing MIA model can provide perfect accuracy in determining membership status, necessitating the development of alternative approaches for conducting data reconstruction attacks.

2.4 Memorization

In the realm of attacks targeting the disclosure of information from training datasets, memorization serves as a key concept closely linked to attack performance. Within machine learning, memorization refers to a model's unintentional retention of intricate details from its training data, which is particularly evident in high-capacity models. Feldman [18] provides a succinct definition of memorization for a target sample (x_i, y_i) with index i, describing it as the impact of removing a data point on the model's prediction for that particular point:

$$mem(\mathcal{A}, \mathcal{D}, i) = \Pr_{f_\theta \sim \mathcal{A}(\mathcal{D})}[f_\theta(x_i) = y_i] - \Pr_{f_\theta \sim \mathcal{A}(\mathcal{D}^{\setminus i})}[f_\theta(x_i) = y_i] \tag{1}$$

where $\mathcal{D}^{\setminus i}$ denotes the dataset with the sample (x_i, y_i) removed.

While memorization can be beneficial, and even indispensable [18], for achieving high model performance, it simultaneously poses risks that adversaries can exploit to expose sensitive information.

Tramèr et al. [56] demonstrates that enhancing a model's memorization through data poisoning attacks significantly increases the effectiveness of various privacy attacks, including membership inference, attribute inference, and data extraction. Further investigations by Carlini et al. [9] illustrate the "privacy onion effect", highlighting that vulnerabilities arising from memorization cannot be easily mitigated by merely removing outliers.

Conversely, reducing a model's memorization of the training data can mitigate its vulnerability to privacy attacks, as seen in approaches like differential privacy [3, 17, 43]. However, this often comes at the expense of model performance. Despite its importance, the relationship between memorization and data reconstruction remains surprisingly underexplored. We aim to address this gap by benchmarking existing data reconstruction attacks from the perspective of memorization.

3 Defining Data Reconstruction

3.1 Reconstruction Taxonomy

To establish a rigorous definition of data reconstruction attacks, it is necessary to take into account various aspects such as training type, model access, and dataset access. To this end, we present a taxonomy that captures these aspects and use it to formulate a formal definition of data reconstruction. We are motivated by two key questions that arise in this context.
1) What data does the adversary aim to reconstruct? The diversity of attack scenarios poses challenges to developing a unified definition of data reconstruction attacks. Some attacks focus on the standard setting where the model is trained on one fixed dataset, while others focus on settings that entail dataset change during the training process, e.g., online learning. Additionally, some attacks only consider data that contributes to certain updates. This diversity makes it hard to incorporate all possibilities into a unified definition.
2) What information does the adversary have? This problem also stems from the diverse attack scenarios where the adversary has varying levels of information. For example, when users release their models, the adversary has white-box access to the model, while if models only provide a query interface, the adversary may lack detailed information about the model's architecture and parameters. Furthermore, the adversary may or may not have access to the same distribution of the target dataset. Thus, it is crucial to consider all possible situations to provide a comprehensive definition of data reconstruction attacks.

In the following, we categorize reconstruction attacks from three dimensions: training type, model access, and dataset access:

Training Type: We categorize training into two types: *static* and *dynamic*, based on the target dataset's role. In static, the target dataset remains fixed during training, and the attack aims to recover it. In dynamic, the target dataset is introduced during training, causing model changes, as the updating dataset in online learning.

Model Access: The attacker may have one of the following access to the target model: *Black-box Access*, which means the attacker could only query the model in an API manner. *White-box Access*, which means the attacker could get full information about the target model, including the model architecture, parameters, and even gradients of the target sample

Table 1: We group ten reconstruction attacks from three dimensions, which indicate the necessary information for the attack, including training type, model access, and dataset access. Note that attacks requiring more information about the model or dataset can be extended from attacks that require less information.

Training Type	Model Access	Dataset Access		
		No Data	Similar Distribution	Same Distribution
Static	Black-Box		Inv-Alignment	Inv-Alignment
	White-Box	MI-Face DeepDream DeepInversion Bias-Rec	Revealer KEDMI PLGMI	Revealer KEDMI PLGMI
Dynamic	Black-Box			Updates-Leak
	White-Box	Deep-Leakage		

calculated on the target model. We also acknowledge that real-world scenarios may involve intermediate access, representing a hybrid of black-box and white-box approaches. For example, an encoder might only be queryable via an API (black-box), while subsequent classification layers are directly accessible (white-box). Our framework's analysis of these two extremes, pure black-box and full white-box, effectively bounds the attack performance for all such intermediate cases.

Dataset Access: The attacker may have one of the following types of knowledge about the target dataset: *No Data*, meaning no access to any dataset information; *Same Distribution*, meaning the attacker can sample data from the same distribution as the target dataset; *Similar Distribution*, meaning the attacker can sample from distributions similar to the target. The key difference between "same" and "similar" distributions lies in the level of specificity of the distribution information that the adversary possesses: "same" refers to detailed information, e.g., images of a specific person, while "similar" refers to more general knowledge, e.g., knowing the dataset contains human faces.

In the following, we denote such information as extra knowledge ($\mathcal{K}$), defined as below:

Definition 3.1 (Extra Knowledge $\mathcal{K}$)**.** Extra knowledge $\mathcal{K}$ provides the necessary information required for the reconstruction attack. A standard extra knowledge should contain information from three dimensions:

1. **Training Type:** static or dynamic
2. **Model Access:** black-box access or white-box access
3. **Dataset Access:** no data or similar distribution or same distribution

It should be emphasized that while our present investigation concerns the realm of vision, the attack taxonomy we have formulated has broader applicability to reconstruction attacks in diverse domains.

We categorize the ten attacks in Table 1. Current reconstruction attacks most focus on the scenario where the adversary has white-box access to the target model, as this access provides crucial information that aids in dataset reconstruction. For attacks without model information, they tend to rely on information from a dataset that shares similar characteristics with the target dataset. For example, Inv-Alignment and Updates-Leak do not require white-box access but require data from the same distribution as a substitute. Furthermore, attacks with model information can be further improved with the help of dataset information. For instance, PLGMI leverages information from both the model and the dataset, resulting in improved performance compared to attacks that rely solely on model or dataset information, as demonstrated in the evaluation part.

Theoretically, attacks with black-box access can be easily extended to white-box attacks, while the inverse transition is not straightforward. One potential solution to transfer white-box attacks to black-box attacks is through the use of model stealing to extract their internal parameters. Additionally, for attacks that leverage information about the same distribution, we also investigate the feasibility of using similar distribution. These aspects are examined in Section 7.

3.2 Reconstruction Definition

Given the extra information, another remaining question is the number of samples to reconstruct. Existing attacks generate a surplus of samples and designate those that best align with the target dataset as the reconstruction outcome. However, this approach is deemed unsuitable as it imparts specific information about the target dataset that is unavailable to the adversary. For the same reason, permitting the adversary to generate an infinite number of samples is inappropriate as generating every conceivable pixel combination can subsume the target dataset, yet such a reconstruction does not furnish any useful information.

Therefore, we explicitly stipulate that the reconstructed

dataset has the same size as the target model. Generating a greater number of samples than the target samples and selecting high-quality reconstructed samples is allowed. But it is crucial that the selection procedure does not involve any information about the target dataset.

It should be emphasized that the reconstruction size is a requisite for evaluation and not for the adversary. Furthermore, our definition encompasses scenarios where the focus is on reconstructing a subset of the training dataset. A detailed description of this is presented in Section 4.3.

Provided the reconstruction size, the target model, and the necessary information indicated in extra knowledge, we formally define the reconstruction algorithm as follows:

Definition 3.2 (Reconstruction Algorithm)**.** Given a target model $m \in \mathcal{M}$ and extra knowledge $k \in \mathcal{K}$, reconstruction algorithm $\mathcal{A}$ could reconstruct a dataset $\mathcal{D}_{rec} = \mathcal{A}^k(m) \in \mathbb{D}$, i.e., $\mathcal{A}$: $\mathcal{M} \times \mathcal{K} \to \mathbb{D}$, with the same size as the target dataset $\mathcal{D}_{tar}$.

In the static setting, the target dataset $\mathcal{D}_{tar}$ is the whole training dataset of the target model; in the dynamic setting, the target dataset refers to the subset of data that directly contributes to the model change.

4 Evaluation Metric for Data Reconstruction

A robust evaluation metric is crucial for the development of data reconstruction attacks, analogous to the role of loss functions in guiding model optimization. This section reviews existing metrics, outlines the desired properties of ideal metrics, and introduces our proposed metrics in Section 4.3.

4.1 Existing Common Metrics

Visualization: Visualization is a useful tool for understanding the quality of reconstructions and has been widely used in previous work [1, 19, 44, 53, 66, 68, 70, 73, 74]. While visualization provides the most direct and intuitive impression of reconstruction quality, its non-quantitative nature and reliance on subjective human judgment limit its effectiveness as an evaluation metric. Consequently, it is crucial to also utilize quantitative metrics in order to accurately assess reconstruction quality.

MSE/PSNR/SSIM: Mean Squared Error (MSE) and other similarity measures, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), are commonly utilized to provide quantitative evaluation results [44, 66, 73, 74]. However, such metrics only measure the similarity between individual samples rather than the overall dataset, which poses two issues. First, the choice of sample pairs for comparison is not fixed, causing evaluation results to vary based on the selected pairs. Second, sample-level metrics can't capture the diversity of the reconstructed dataset. For example, if all reconstructed samples are similar to *one* target sample, the measured distance may be small, but the reconstruction may still be unsatisfactory.

Feature Distance: Feature distance measures the similarity in the feature space and has been used in previous works [13, 73]. Concretely, for each reconstructed sample, the distance to the centroid of its class in the feature space is calculated. However, as the feature space is determined by an evaluation network, the evaluation results can be inconsistent because different evaluation networks use different feature spaces with varying class centroids. Additionally, like other sample-level metrics such as MSE, this measure doesn't capture the diversity of the reconstruction.

Accuracy (Train): To incorporate macro similarity, one method uses reconstructed samples to train a model for the same task [68, 70, 73]. The model's testing accuracy reflects the reconstruction quality, with higher accuracy indicating better reconstruction. We point out that this metric overlooks the precision of the reconstruction, as demonstrated by a counterexample in the extended version [62]. Concretely, we utilize the data-free model extraction method [58] to generate a dataset, which is then used to steal the target model. Although the resulting stolen model exhibits high accuracy, the generated dataset, which serves as the training dataset of the stolen model, is vastly dissimilar from the target dataset.

Accuracy (Test): Alternatively, one can train an evaluation model on a dataset from the same distribution as the target model and assess whether the evaluation model can accurately classify each reconstructed sample [13, 53, 59, 70, 73]. The idea is that high-quality reconstructions should contain recognizable patterns that the evaluation model can capture. However, we argue that this metric is ineffective, as shown by examples that resemble random noise but are still classified with high confidence by the evaluation model, as illustrated in the extended version [62]. These counterexamples, generated using MI-Face [19], achieve prediction confidences above 0.9 for their corresponding classes.

4.2 Design Desiderata

To propose metrics that can reflect the quality of a reconstruction, we need to consider four questions: 1) can such metrics *quantitatively* measure the quality? 2) can such metrics provide a *consistent* result for a fixed reconstruction? 3) can such metrics embody the quality in a *micro* aspect? 4) can such metrics incorporate the quality in a *macro* aspect?

In response to the drawbacks of existing metrics, we propose a set of properties that suitable evaluation metrics should possess, including quantifiability, consistency, precision, and diversity:

- **Quantifiability.** The evaluation metric should provide quantitative results and eliminate the influence of subjective factors.

- **Consistency.** The evaluation metric should be consistent, that is, given a pair of the reconstructed dataset and the target dataset, the evaluated result should be determined, regardless of the testing time, place, and order.
- **Precision.** The evaluation metric should be aware of the reconstruction precision. Specifically, it should capture the sample-level similarity of reconstructed samples to target samples.
- **Diversity.** The evaluation metric should be aware of the reconstruction diversity. Concretely, the metric should reflect the percentage of data being reconstructed. An attack that can only recover a few samples accurately is not regarded as a successful reconstruction attack.

4.3 Definition of Our Quantitative Metrics

To capture both precision and diversity aspects, we introduce two metrics from different perspectives. The first is a dataset-level metric that measures the distribution similarity between the reconstructed and target datasets.

Definition 4.1 (Dataset-level Metric)**.** The dataset-level metric $\mu : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ is defined as a mapping that takes two datasets and produces a single real number D-Dis, i.e., $\text{D-Dis} = \mu(\mathcal{A}^k(m), \mathcal{D}_{tar})$.

We leverage Fréchet Inception Distance (FID) to incorporate the macro quality of the reconstructed dataset, which measures the distance between different data distributions. Specifically, we approximate the two dataset distributions by Gaussian distributions as follows:

$$\mathcal{D}_{tar} \sim \mathcal{N}(\nu_{tar}, \Sigma_{ori}), \text{ and } \mathcal{A}^k(m) \sim \mathcal{N}(\nu_{rec}, \Sigma_{rec})$$

We compute the D-Dis value as:

$$||\nu_{tar} - \nu_{rec}||_2^2 + tr\left(\Sigma_{tar} + \Sigma_{rec} - 2(\Sigma_{tar}^{\frac{1}{2}} \cdot \Sigma_{rec} \cdot \Sigma_{tar}^{\frac{1}{2}})^{\frac{1}{2}}\right)$$

where $||\cdot||_2$ denotes the Euclidean distance, and $tr(\cdot)$ denotes the trace of a matrix. ν and variance Σ denote the parameters of the corresponding Gaussian distributions.

We choose Fréchet Inception Distance (FID) to measure the dataset-level distance due to its established use as a metric for evaluating the quality of generated distributions. Furthermore, FID is model-agnostic and can be calculated using any feature extractor, but common practice involves using the pre-trained InceptionV3 regardless of the dataset[1]. Using FID to measure dataset-level distance also resolves concerns related to consistency, as feature distances are calculated using the same model across different tasks. We further introduce sample-level metrics to improve our evaluation by taking precision into account. Specifically, we concentrate on how similar the reconstructed sample is to the target sample in the original data space.

[1] https://github.com/mseitzer/pytorch-fid

Definition 4.2 (Sample-level Metric)**.** The sample-level metric $\mu : \mathbb{D} \times \mathbb{D} \to \mathbb{R}^2$ maps two datasets into two real numbers S-Dis and α, i.e., $(\text{S-Dis}, \alpha) = \mu(\mathcal{A}^k(m), \mathcal{D}_{tar})$, where the first value S-Dis calculates the averaged minimal distance between reconstructed dataset and the target dataset, and the second value α denotes the coverage of reconstructed part. Formally:

$$\text{S-Dis} = \frac{1}{|\mathcal{A}^k(m)|} \sum_{x_i \in \mathcal{A}^k(m)} d(x_i, f(x_i))$$

where $f : \mathcal{A}^k(m) \to \mathcal{D}_{tar}$, such that, $\forall x_i \in \mathcal{A}^k(m)$, f maps x_i to the sample $f(x_i) \in \mathcal{D}_{tar}$ with the minimal distance to x_i.

For coverage α, it indicates the reconstructed diversity:

$$\alpha = \frac{|f(\mathcal{A}^k(m))|}{|\mathcal{D}_{tar}|}$$

Here, $f(\mathcal{A}^k(m))$ denotes the image of f, and $\alpha \in (0, 1]$, larger coverage indicates better diversity.

Note that the choice of d may vary depending on the aspect of reconstruction quality to be assessed. Both metrics proposed above provide deterministic evaluation results, only based on the content of the reconstructed dataset, which satisfies the property of quantifiability and consistency. Furthermore, the coverage reflects the diversity of the reconstruction. The precision of the metrics is partially illustrated in Section 5.1, and we discuss it in more detail in Section 6.1.

We can see that the reconstruction of a subset is encompassed by our reconstruction definition. To clarify, we can replicate such samples, and the outcome will remain unaffected as the corresponding image in the target dataset is fixed, resulting in an unchanged coverage and distance. In the event that accurately reconstructing a subset, it is conceivable that it will be characterized by a small distance, yet it may have limited coverage. A small distance indicates an accurate reconstruction, while the small coverage indicates that the reconstructed samples constitute only a minor fraction.

Provided the evaluation metric, we can parameterize the attack to characterize the ability of data reconstruction attack. We first formulate the ideal case where the attack exactly reconstructs the target dataset under the measurement of μ.

Definition 4.3 (μ-Exact Reconstruction)**.** We say a reconstruction algorithm $\mathcal{A}$ achieves μ-exact reconstruction if $\mu(\mathcal{A}^k(m), \mathcal{D}_{tar}) = 0$.

μ denotes the distance measure that specifies the quality of the reconstruction, where the output could be a tuple if the measure has more than one evaluation metric. Note that μ-Exact Reconstruction is a necessary and insufficient condition for the perfect reconstruction, depending on the choice of measure μ, which further highlights the importance of selecting appropriate metrics to evaluate the reconstruction performance.

The relaxed version allows the reconstructed dataset to have a small distance ε with the target dataset, we refer to this version as (ε,μ)-Approximate Reconstruction:

Definition 4.4 ((ε,μ)-Approximate Reconstruction)**.** We say a reconstruction algorithm $\mathcal{A}$ achieves (ε,μ)-approximate reconstruction if $\mu(\mathcal{A}^k(m),\mathcal{D}_{tar}) \leq \varepsilon$.

Remark. (ε,μ)-Approximate Reconstruction corresponds to the specific case of μ-Exact Reconstruction when $\varepsilon = 0$.

Our framework's core metrics effectively address a broad spectrum of attack goals. For instance, reconstructing a single missing sample [45] translates within our definition to a low coverage score, indicating its narrow scope, combined with a reconstruction-distance metric to assess fidelity. In contrast, an attack targeting an entire class aims for high coverage [19], signifying the retrieval of numerous distinct dataset items. Overall, the two metrics work together: distance measures reconstruction accuracy, while coverage scales with the breadth of the attacker's objective, from single-sample inference to full-dataset recovery.

5 Evaluation

Due to space constraints, we defer the description of the experimental setup to the extended version [62], where we provide details on the dataset, attack configurations, and evaluation metrics.

5.1 Results Under Quantitative Metrics

We divide ten reconstruction attacks into two groups based on the training type. To provide an overview of the attack performance, we visualize the reconstruction results of all ten attacks in Figure 1. Overall, none of the attacks can achieve exact reconstruction, and their quality varies significantly. To better analyze and compare these attacks, we apply the proposed metrics to evaluate the reconstruction datasets, and report the results in Table 2.

Dataset-level Metric: For the CelebA dataset with the largest size, we visually confirm that three GAN-based attacks (Revealer, KEDMI, and PLGMI) outperform other methods, which is consistent with their lower FID scores. For dynamic training-type attacks, Updates-Leak shows a lower FID score than Deep-Leakage, indicating better reconstruction quality. This pattern holds for the other two datasets. We also observe that DeepInversion's performance worsens as dataset complexity[2] increases, with FID scores rising from 64.678 (MNIST) to 131.545 (CIFAR10) to 234.672 (CelebA). In contrast, GAN-based attacks, such as PLGMI, maintain clear semantic reconstruction, reflected in their low FID scores: 82.940 (MNIST), 123.018 (CIFAR10), and 85.143 (CelebA).

[2]We follow the criteria used in [37] to determine the complexity of the dataset. Briefly, gray-scale datasets are simpler than colored datasets.

In general, the FID score reflects reconstruction quality, with lower values indicating better reconstructions. However, we argue that relying solely on FID may not fully capture reconstruction quality, as shown with the CelebA dataset. As seen in Figure 1, Inv-Alignment generates reconstructions that clearly show the semantic meaning of the target dataset, i.e., human faces, whereas some DeepInversion reconstructions lack sufficient information, despite having a lower FID score (234.672) compared to Inv-Alignment (357.610). This highlights the need to incorporate sample-level metrics like SSIM, PSNR, and MSE to more comprehensively evaluate reconstruction quality.

Sample-level Metric: Sample-level metrics complement dataset-level metrics by providing a micro-scale evaluation of reconstructions, which, in certain contexts, may better align with human perception. As previously noted, Inv-Alignment surpasses DeepInversion on the CelebA dataset; however, this superiority is not reflected by the FID score. In contrast, all three sample-level metrics deliver results consistent with our expectations, as shown in Figure 1. Additionally, sample-level metrics and dataset-level metrics generally provide consistent outcomes, as demonstrated in Figure 2. Specifically, attacks achieving higher performance at the dataset level also tend to perform well on sample-level metrics. For instance, Figure 2a shows that reconstructions with higher FID scores typically exhibit lower SSIM values, with similar trends observed across other metrics.

Sample-level metrics also play a vital role in assessing the diversity of reconstructed samples (coverage), a key indicator of a reconstruction attack's success. Diversity is challenging to evaluate using dataset-level metrics, as these rely on a few representative samples to approximate the distribution. Diversity is crucial for determining reconstruction quality, as generating identical data similar to a few samples in the training dataset might achieve high performance but does not constitute an effective reconstruction attack, as it limits the exposed information to only a few data points. To measure coverage, we identify the nearest pair for each reconstructed sample in the target dataset within the same class. The proportion of such pairs in the target dataset indicates the reconstruction's diversity. Given that coverage is influenced by both the number of classes and reconstructed samples, we present the results for CIFAR10 and MNIST together in the extended version [62], and separately for CelebA. The findings reveal an intriguing observation: high-quality reconstructions do not necessarily correspond to greater diversity in the reconstructed samples. Furthermore, results in Table 2 show a decline in coverage as the size of the training dataset increases. This trend aligns with our hypothesis that current reconstruction attacks tend to focus on capturing general information about the training dataset rather than individual sample details. Consequently, the ability to reconstruct diverse, unique samples diminishes as the training dataset grows, leading to lower coverage metrics for larger datasets.

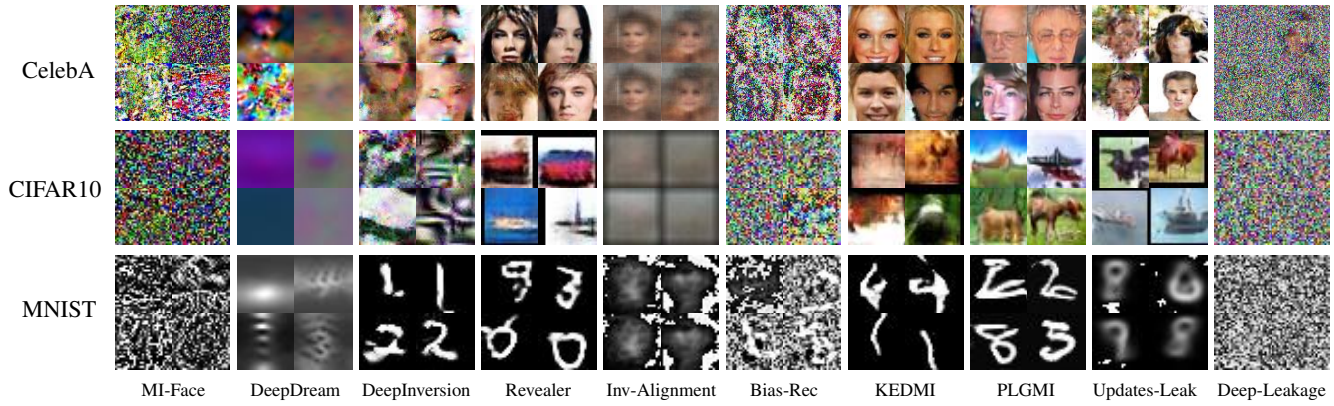
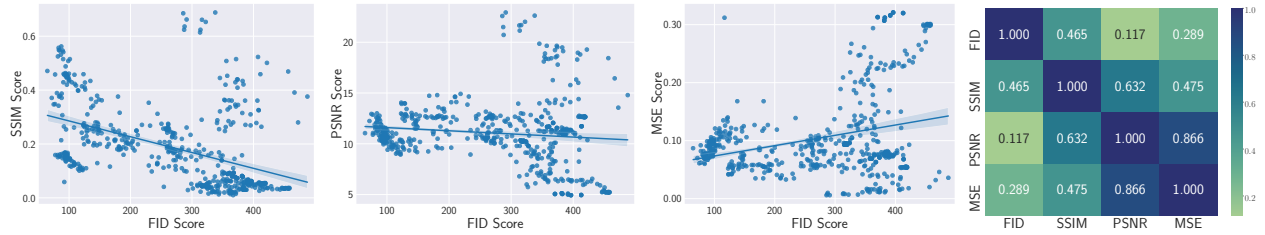


Figure 1: Visualization of existing reconstruction attacks. For each attack, the left two images are reconstructed from the target model with a smaller training size (1,000 for CelebA and 100 for CIFAR10 and MNIST), and the right two images are from the larger one (20,000).



(a) FID vs. SSIM (b) FID vs. PSNR (c) FID vs. MSE (d) Correlation

Figure 2: Relationship between sample-level metrics and the dataset-level metric, we plot the correlation heatmap using the absolute value of correlation between different metrics.

Summary: While there is a relationship between dataset-level and sample-level metrics, their correlation is relatively weak, especially when evaluating low-quality reconstructions, as illustrated in Figure 2d. For example, both MI-Face and DeepDream perform poorly on the CelebA task, which is reflected in their FID scores (336.906 vs. 370.396). However, their SSIM scores vary significantly (0.023 vs. 0.346), highlighting the inconsistency between metrics.

To examine the relationship between metrics in the context of high-quality reconstructions, we focus on three GAN-based attacks that generate higher-quality outputs. In these cases, the correlation coefficients improve: the correlation between FID and SSIM increases from 0.465 to 0.597, FID and PSNR from 0.117 to 0.721, and FID and MSE from 0.289 to 0.511.

Furthermore, low-quality reconstructions introduce discrepancies in coverage, even within the same attack. For instance, in the case of Bias-Rec on CelebA, coverage varies significantly between SSIM (24.28%) and PSNR (0.57%). When low-quality reconstructions are excluded, coverage across sample-level metrics aligns more closely, with the correlation coefficient between SSIM and PSNR, as well as between SSIM and MSE, increasing dramatically from 0.429 to 0.993.

Despite these improvements, the correlation between dataset-level and sample-level metrics remains insufficiently strong to ignore their distinctions. Dataset-level metrics provide a broad perspective on reconstruction quality, whereas sample-level metrics offer detailed insights into the fidelity of individual samples. Therefore, it is essential to utilize multiple metrics during evaluation. Although a single score can be derived using the minimum of normalized metrics to represent reconstruction quality, it is generally advisable to consider all available metrics. Ultimately, the decision between prioritizing usability and ensuring accuracy or effectiveness involves a careful trade-off.

6 Memorization in Data Reconstruction

We observe that existing attacks show varying performance, and even the same attack may exploit different levels of vulnerabilities when target models are trained with datasets of varying sizes. From the perspective of model owners, understanding which models are more vulnerable to data reconstruction attacks is essential. In this section, we examine model vulnerability to data reconstruction attacks through

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA with 6 different sizes. Attacks with a gray background belong to the dynamic training type. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance. Experimental results for other model architectures and datasets can be found in the extended version [62].

Attack	Metrics		Target Data Size					
			1,000	2,000	5,000	10,000	15,000	20,000
Memorization			**1.000**	**0.981**	**0.862**	**0.539**	**0.386**	**0.301**
MI-Face	Dataset-level	FID ↓	362.361	365.570	361.517	340.959	334.520	336.906
	Sample-level	SSIM ↑	0.014(100.00%)	0.019(72.15%)	0.028(57.88%)	0.024(57.75%)	0.024(57.28%)	0.023(54.80%)
		PSNR ↑	7.978(100.00%)	8.453(53.80%)	8.872(30.74%)	8.851(17.63%)	9.005(13.61%)	9.037(10.08%)
		MSE ↓	0.163(100.00%)	0.145(53.80%)	0.132(30.74%)	0.131(17.63%)	0.127(13.61%)	0.126(10.08%)
DeepDream	Dataset-level	FID ↓	337.139	306.974	297.590	315.676	345.472	370.396
	Sample-level	SSIM ↑	0.079(100.00%)	0.140(57.65%)	0.234(26.44%)	0.290(14.38%)	0.321(8.95%)	0.346(6.81%)
		PSNR ↑	9.162(100.00%)	10.304(53.70%)	11.957(26.60%)	13.006(14.43%)	13.980(9.29%)	14.027(7.63%)
		MSE ↓	0.128(100.00%)	0.100(53.70%)	0.069(26.60%)	0.053(14.43%)	0.042(9.29%)	0.041(7.63%)
DeepInversion	Dataset-level	FID ↓	287.497	283.429	273.183	273.415	245.736	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.110(61.70%)	0.118(42.44%)	0.140(28.97%)	0.150(27.07%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.094(56.10%)	10.550(31.68%)	11.143(21.07%)	11.388(19.09%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.105(56.10%)	0.094(31.68%)	0.081(21.07%)	0.076(19.09%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	103.428	94.899	93.961	93.781	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.116(66.75%)	0.135(50.50%)	0.150(44.05%)	0.157(40.47%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	9.720(60.90%)	10.087(43.68%)	10.449(37.20%)	10.622(33.66%)	10.733(32.17%)
		MSE ↓	0.123(100.00%)	0.113(60.90%)	0.103(43.68%)	0.094(37.20%)	0.091(33.66%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	243.849	359.419	229.609	361.569	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.312(51.45%)	0.285(22.56%)	0.353(13.09%)	0.336(9.46%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	12.463(52.40%)	13.023(23.10%)	13.858(13.93%)	14.007(9.74%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.061(52.40%)	0.052(23.10%)	0.043(13.93%)	0.041(9.74%)	0.039(8.02%)
Bias-Rec	Dataset-level	FID ↓	327.883	323.892	316.452	319.774	318.895	315.227
	Sample-level	SSIM ↑	0.039(38.20%)	0.040(35.55%)	0.043(30.88%)	0.044(27.50%)	0.045(25.44%)	0.047(24.28%)
		PSNR ↑	9.783(3.20%)	10.211(0.75%)	10.249(0.80%)	10.311(0.53%)	10.222(0.60%)	10.354(0.57%)
		MSE ↓	0.105(3.20%)	0.096(0.75%)	0.095(0.80%)	0.093(0.53%)	0.095(0.60%)	0.093(0.57%)
KEDMI	Dataset-level	FID ↓	121.689	110.088	101.730	94.341	95.852	97.667
	Sample-level	SSIM ↑	0.111(100.00%)	0.124(57.30%)	0.144(32.60%)	0.157(24.29%)	0.160(21.10%)	0.167(18.93%)
		PSNR ↑	9.167(100.00%)	9.715(54.50%)	10.605(30.28%)	11.126(20.60%)	11.428(17.12%)	11.560(15.36%)
		MSE ↓	0.133(100.00%)	0.115(54.50%)	0.092(30.28%)	0.081(20.60%)	0.075(17.12%)	0.073(15.36%)
PLGMI	Dataset-level	FID ↓	127.722	107.883	104.842	97.730	84.836	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.132(60.40%)	0.136(37.02%)	0.149(26.15%)	0.154(21.83%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.164(56.10%)	10.588(31.58%)	10.921(20.55%)	11.299(16.44%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.102(56.10%)	0.092(31.58%)	0.085(20.55%)	0.078(16.44%)	0.074(13.54%)
Updates-Leak	Dataset-level	FID ↓	192.446	259.231	263.899	275.845	312.011	260.100
	Sample-level	SSIM ↑	0.203(18.00%)	0.184(11.00%)	0.194(9.00%)	0.155(22.00%)	0.170(18.00%)	0.187(5.00%)
		PSNR ↑	13.173(14.00%)	13.165(4.00%)	12.693(6.00%)	12.554(14.00%)	12.537(10.00%)	12.509(6.00%)
		MSE ↓	0.049(14.00%)	0.050(4.00%)	0.056(6.00%)	0.058(14.00%)	0.058(10.00%)	0.058(6.00%)
Deep-Leakage	Dataset-level	FID ↓	376.852	383.272	382.227	384.841	384.023	383.338
	Sample-level	SSIM ↑	0.031(49.00%)	0.030(48.00%)	0.034(43.00%)	0.033(45.00%)	0.037(43.00%)	0.040(52.00%)
		PSNR ↑	10.957(3.00%)	10.965(2.00%)	11.014(4.00%)	11.076(2.00%)	11.152(2.00%)	11.199(2.00%)
		MSE ↓	0.080(3.00%)	0.080(2.00%)	0.079(4.00%)	0.078(2.00%)	0.077(2.00%)	0.076(2.00%)

the lens of memorization. This choice is motivated by the close connection between memorization and membership inference attacks. Previous work [9, 56] suggests that strongly memorized samples are more vulnerable to membership inference. Given that data reconstruction can be framed as a search problem within the context of membership inference, we investigate whether models with higher memorization scores are likewise more prone to data reconstruction attacks.

As shown in Table 2, we trained models using six datasets of varying sizes, which resulted in different levels of memorization for individual samples. To quantify memorization at the model level, we extend the sample-based memorization definition (Equation 1) to cover the entire model. Specifically, we use the average memorization score of the first 1,000 samples in the training dataset as a proxy for the model's overall

Table 3: Evaluation with GPT-4o.

Attack	Metrics	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
	Memorization	**1.000**	**0.981**	**0.862**	**0.539**	**0.386**	**0.301**
Revealer	# of Major	349	182	188	133	94	54
	# of All	166	35	45	46	26	15
	Pred Rate	0.335	0.185	0.194	0.137	0.096	0.053
KEDMI	# of Major	303	157	227	188	90	35
	# of All	115	14	37	37	14	6
	Pred Rate	0.281	0.169	0.217	0.189	0.097	0.047
PLGMI	# of Major	325	133	162	192	120	68
	# of All	116	12	12	44	15	6
	Pred Rate	0.283	0.142	0.174	0.200	0.125	0.076
PLGMI (Pre)	# of Major	484	146	186	126	38	20
	# of All	246	15	29	21	8	0
	Pred Rate	0.446	0.160	0.188	0.125	0.054	0.026

memorization score:

$$\textit{model-mem}(\mathcal{A},\mathcal{D}) = \mathop{\mathbb{E}}_{x_i\in\mathcal{D}}\Big[\Pr_{f_\theta\sim\mathcal{A}(\mathcal{D})}[f_\theta(x_i)=y_i] - \Pr_{f_\theta\sim\mathcal{A}(\mathcal{D}^{\setminus i})}[f_\theta(x_i)=y_i]\Big] \tag{2}$$

As expected, we observe that the model memorization score decreases from 1.000 to 0.301 as the training size increases from 1,000 to 20,000. However, the performance of different attacks is inconsistent. For instance, with the DeepDream attack, the FID distance to the training dataset is 337.139 when the memorization score is 1.000, and it decreases as the memorization score drops to 0.301. In contrast, for attacks such as Revealer, reconstruction quality (as measured by FID) improves as the memorization score decreases.

We attribute this discrepancy to two potential factors. First, the attack methods may lack the necessary capacity to accurately capture underlying vulnerabilities, particularly in the case of non-generative model-based attacks. This further raises the question of whether current methods are genuinely extracting private information from the model or merely imputing plausible samples, we provide some initial discussion in Section 7. Second, the evaluation metrics used may not effectively capture the true performance of these attacks. As discussed in Section 5.1, while existing metrics provide a rough estimate of reconstruction quality, there remains a gap between the numerical results and human perception. In some cases, measurements do not align with visual evaluations. We discuss this challenge in the next section.

6.1 Linkage Between Model Memorization and Dataset Leakage

To evaluate the effectiveness of data reconstruction, we have introduced a set of metrics designed to quantitatively assess reconstruction quality. While these metrics offer a coarse-grained view of reconstruction success, previous findings suggest that high quantitative scores do not necessarily correlate with high-quality reconstructions. Although human inspection remains the most direct and intuitive method for quality assessment, it is often prohibitive in terms of cost and scale for extensive datasets. Therefore, to achieve an efficient and scalable evaluation, we propose utilizing GPT-4o. This model is recognized for its robust performance and has been further refined during the Reinforcement Learning from Human Feedback (RLHF) phase to align with human preferences.

Table 4: Evaluation with InternVL 2.5 and Claude 3.7 for PLGMI (pre) on VGG16 trained on CelebA.

LLMs	Metrics	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
InternVL 2.5	# of Major	531	134	92	149	65	29
	# of All	69	0	0	2	0	0
	Pred Rate	0.410	0.157	0.130	0.165	0.096	0.042
Claude 3.7	# of Major	479	188	41	125	130	37
	# of All	143	31	0	23	17	0
	Pred Rate	0.434	0.179	0.056	0.134	0.158	0.039

We evaluate the attack performance using 1,000 randomly chosen CelebA images as targets. Each target appears in six training sets of increasing size. For every set, we run the reconstruction algorithm and keep the 1,000 results with the highest PSNR score relative to their targets. Thus, each target image has six candidate reconstructions—one from each training size. To decide which reconstruction looks closest to the ground truth, we prompt GPT-4o: "*From the second to the seventh image, which image is more similar to the first one? Please make sure your response must be the index of that image and don't say any other words.*" We repeat this query five times and tally how often each training size is chosen. We report three summary measures: "# of major" (how often a setting wins the majority vote), "# of all" (how often the vote is unanimous), and "pred rate" (its overall selection frequency across all queries).

Our analysis concentrates on three GAN-based approaches, which consistently yield high-quality reconstructions. As indicated in Table 3, despite some variability, all three metrics across the 1,000 identities generally exhibit a trend where reconstruction quality diminishes as the memorization score decreases. These results can also be generalized to other leading LLMs, such as InternVL 2.5 and Claude 3.7, as demonstrated in Table 4. Visual evidence in Figure 3 further supports the observation that reconstructions from the smallest target dataset size (1,000 samples) resemble the target images more closely than those from the largest size (20,000 samples). This observation aligns with our expectations but contrasts with results from previous metrics.

We interpret these findings in two ways. From the perspective of GAN-based attacks, these methods reconstruct data at a class level and aim to closely approximate the distribution of the targets. Consequently, when a target dataset includes multiple samples within a class, achieving a close match for any specific target sample becomes unlikely. Conversely, smaller

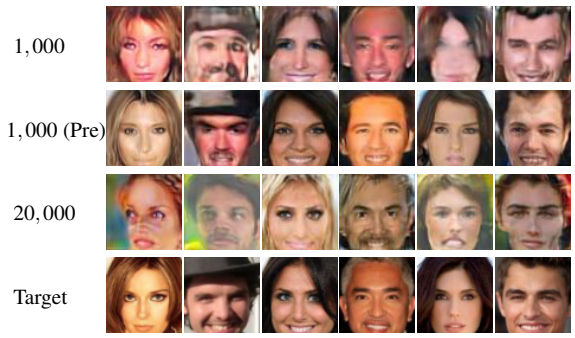


Figure 3: Visualization of PLGMI on different target models.

target datasets result in less generalized feature learning, manifesting as blurred areas in reconstructions, like the eyes and cheeks. While this blurriness negatively impacts performance as measured by both dataset-level and sample-level metrics, its effect on visual assessments is relatively modest, provided the blurring is not extensive.

This second interpretation brings forth another noteworthy observation. For each attack method, the optimal performance is achieved at a dataset size of 1000. Interestingly, the second-best performance does not occur at a slightly larger size, such as 2000, but frequently occurs at medium dataset sizes, such as 5,000 or 10,000 samples. This suggests that the reconstruction process benefits from the model's ability to learn additional features. However, this advantage is counterbalanced by the challenges posed by larger datasets, as a more extensive training dataset complicates the accurate recovery of individual samples, consistent with our earlier findings.

To further explore the benefits of feature learning, we trained a model pre-trained on disjoint datasets and fine-tuned it with the target data. This approach enables the model to learn additional features, as evidenced by the improved testing accuracy. We evaluate the attack on this fine-tuned model in Table 3. The results indicate that pre-training enhances the advantage of using a smaller dataset size, as the pre-trained model has already developed a degree of generalizability and is able to learn sample-specific features more efficiently.

To better understand the effect of pre-training, we compare attack performance between models trained from scratch and those fine-tuned from pre-trained weights, as presented in Table 5. Fine-tuning a pre-trained model enables it to learn additional features beyond those required for generalization, leading to improved reconstruction performance.

Together, these findings demonstrate that learning additional features enhances reconstruction quality. Notably, reconstructions from pre-trained models exhibit finer structural details and less blurring, as shown in Figure 3.

We further validate the role of learning additional features from the opposite perspective by reducing the bottleneck size (i.e., the width of the final layer after convolutions), thereby limiting the model's capacity to retain features [55]. As shown

Table 5: PLGMI vs. PLGMI (Pre).

Metrics	Target Data Size					
	1,000	2,000	5,000	10,000	15,000	20,000
# of Major	11 : 989	17 : 983	6 : 994	6 : 994	3 : 997	6 : 994
# of All	1 : 956	2 : 939	1 : 964	0 : 961	0 : 979	0 : 952
Pred Rate	0.015 / 0.985	0.023 / 0.977	0.012 / 0.988	0.013 / 0.987	0.006 / 0.994	0.014 / 0.986

Table 6: GPT-4o evaluation with different sizes of feature embeddings on CelebA with data size of 1,000 and 20,000.

Attack	Metrics	Feature size (1000)			Feature size (20000)		
		2,048 (Ori)	512	128	2,048 (Ori)	512	128
PLGMI	# of Major	917	75	8	742	251	7
	# of All	719	12	0	465	39	0
	Pred Rate	0.884	0.105	0.011	0.725	0.263	0.012

in Table 6, this reduction in learned feature information leads to a degradation in reconstruction performance.

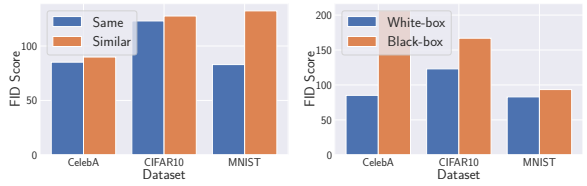
From this study, we derive two key insights. First, from an attack perspective, for GAN-based methods that produce high-quality reconstructions, we observe a trade-off between the generalizability of the target model and the quality of the reconstructions. This insight could inspire further advancements in data reconstruction attacks. Second, from an evaluation perspective, the quantitative metrics employed reliably assess attack performance when the differences between methods are pronounced (e.g., comparing GAN-based methods to techniques producing less semantically meaningful reconstructions, such as MI-Face). However, since current reconstruction attacks are far from perfectly recovering target datasets, existing dataset-level and sample-level metrics often fail to align precisely with human perception, particularly when differences are subtle.

With the advancements in large language models (LLMs), some limitations of visualization as an evaluation metric can be mitigated. We encourage future research to explore integrating LLMs with quantitative metrics for a more comprehensive evaluation of data reconstruction attacks.

7 Utilization of Attack Knowledge

This section investigates whether data reconstruction reveals sensitive information or just replicates attack knowledge, and how different levels of attack information affect performance.

Influence of Data Access: We first assess how current methods use auxiliary datasets to enhance attacks, focusing on whether reconstruction merely involves imputing data from a similarly distributed dataset. By replacing auxiliary datasets with similar but distinct ones, that is, Kuzushiji-MNIST [15] for MNIST, CIFAR100 [2] for CIFAR10, and FFHQ [28] for CelebA, we observe moderate performance drops in GAN-based attacks (see Figure 4a and more in the extended ver-



(a) Influence of data access (b) Influence of model access

Figure 4: Influence of auxiliary information for PLGMI. More results can be found in the extended version [62].

Clean Backdoor

Figure 5: Effect of additional data to the attack performance of Inv-Alignment.

Table 7: Effect of batch normalization for DeepInversion on VGG16 trained on CelebA.

BatchNorm	Target Data Size					
	1,000	2,000	5,000	10,000	15,000	20,000
updated	287.497	283.429	273.183	273.415	245.736	234.672
fixed	372.803	348.387	346.099	346.149	325.639	311.791

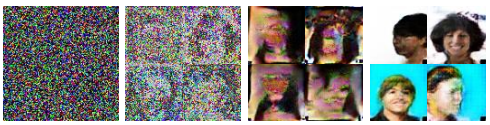
sion [62]). This suggests reconstruction extends beyond mere imputation from auxiliary data.

The picture changes for low-fidelity methods such as Inv-Alignment. Supplying auxiliary data drawn from a different distribution can even yield better results than providing data from the original distribution. For instance, providing CIFAR100 as the auxiliary set yields a FID of 330.31, whereas supplying data from the same distribution gives 357.91.

To further investigate this limitation, we conducted a backdoor experiment. A model was trained on a mixed dataset that included both clean and backdoored images, where the backdoored images contained a 16×16 black square in the bottom-right corner. We then reconstructed the training set while giving Inv-Alignment both clean and backdoored samples. Although the trigger should be a strong cue, it never appears in the reconstructions (see Figure 5). These findings suggest that low-quality reconstruction methods fail to exploit critical attack information.

Influence of Model Access: We next investigate whether access to model parameters increases information leakage by comparing attack performance under black-box and white-box settings.

For black-box evaluations, we first apply state-of-the-art model stealing techniques [58] to create a white-box surrogate model from the black-box target model. We then execute standard attacks on this surrogate model.



MI-Face Bias-Rec DeepInversion Revealer

Figure 6: Visualization of attacks on VGG16 trained on backdoored CelebA with size 20,000.

The results presented in Figure 4b demonstrate that several attack methods—particularly those with stronger baseline performance—benefit from white-box access, as expected. The performance gap widens for more complex reconstruction tasks, indicating that these methods effectively exploit model parameters to improve their performance. To illustrate this point, we examine two specific attacks: DeepDream and DeepInversion. The key distinction between them is that DeepInversion explicitly leverages the statistical information encoded in BatchNorm layers. As shown in Table 2, DeepInversion consistently outperforms DeepDream, highlighting the benefit of incorporating internal model statistics.

To further investigate the role of statistical information stored in model parameters, we conduct an experiment where a target model is trained with BatchNorm parameters fixed, allowing only the remaining parameters to be updated. We then apply DeepInversion to both the standard and modified models. As shown in Table 7, the reconstruction performance significantly degrades in the modified setting, despite using the same attack method. This performance drop underscores the importance of access to BatchNorm statistics, which appear critical for high-quality reconstruction.

However, it is important to note that not all attack methods exhibit improved performance under white-box conditions. In some cases, the lack of performance gain suggests that these methods fail to effectively utilize the information memorized by the target model. To further validate this, we evaluate attacks on a backdoored model, where a predefined trigger, known to be strongly memorized, is embedded during training. As shown in Figure 6, among the selected methods, DeepInversion and Revealer are able to clearly reconstruct the square trigger, and they also produce the highest-quality reconstructions overall. This finding aligns with earlier observations and further emphasizes the limitations of certain attack methods in fully leveraging model-internal information.

8 Discussion and Conclusion

Our findings in Table 3 reveal a clear trend: model memorization is strongly correlated with reconstruction performance. Specifically, models that exhibit higher levels of memorization toward individual training samples tend to be more vulnerable

Table 8: Evaluation on Swin with GPT-4o.

Attack	Metrics	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
PLGMI	# of Major	101	435	229	185	46	4
	# of All	10	43	11	10	3	0
	Pred Rate	0.120	0.369	0.251	0.197	0.056	0.007

Table 9: Evaluation of the Vec2Text [42] attack performance across varying target data sizes on different datasets, measured by BLEU, Sim. (Similarity), and R-L (ROUGE-L).

Dataset	Metrics	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
SST2	BLEU	0.169	0.172	0.152	0.083	0.066	0.058
	Sim.	0.707	0.706	0.663	0.514	0.460	0.440
	R-L	0.457	0.463	0.432	0.298	0.251	0.231
AGNews	BLEU	0.039	0.042	0.036	0.028	0.026	0.025
	Sim.	0.658	0.657	0.585	0.503	0.470	0.453
	R-L	0.203	0.210	0.195	0.175	0.166	0.163
IMDB	BLEU	0.010	0.010	0.009	0.009	0.008	0.008
	Sim.	0.531	0.501	0.477	0.468	0.463	0.465
	R-L	0.136	0.135	0.135	0.132	0.132	0.132

to reconstruction attacks.

We extend this analysis to larger, contemporary model architectures—particularly transformer-based models, which are widely adopted in current practice. In our experiments with Swin [38] and MAE [24] architectures, results presented in Table 8 support similar observations: models trained on smaller datasets exhibit greater vulnerability compared to those trained on larger datasets.

We further broaden our analysis to encompass additional data modalities, with a focus on the text domain, motivated by the rapid rise of Large Language Models. In this setting, we investigate two representative reconstruction attack strategies: Vec2Text [42], which reconstructs input text from its embeddings, and Complete [10, 61], which attempts to recover the original prompt provided to an LLM. Detailed descriptions of these methods and the experimental setup are provided in Appendix B. To quantitatively measure attack efficacy, we utilized three metrics. Semantic similarity served as a macro-level metric, assessing the overall likeness between the reconstructed and original texts. For micro-level evaluation, analogous to pixel-level comparisons in image reconstruction, we employed BLEU and ROUGE-L scores to capture finer-grained textual similarities.

As shown in Table 9, despite the difference in modality, we observe a consistent pattern: training on larger datasets reduces the model's memorization of individual samples, which in turn leads to poorer reconstruction quality. This suggests that memorization plays a critical role in determining vulnerability to reconstruction.

Building on these observations, a natural defense strategy emerges: reducing memorization may decrease susceptibility to reconstruction attacks. This insight helps explain why Differential Privacy (DP) is effective—by limiting the model's exposure to individual samples, DP inherently reduces memorization. Beyond DP, we also explore model pruning as an alternative. As shown in Table 10, pruning a substantial number of neuron connections discards stored information, which can mitigate unnecessary memorization and thus reduce reconstruction attack success. Notably, pruning results in minimal performance degradation—typically under 1%—whereas DP often incurs a much larger accuracy drop [23]. This suggests that pruning may be a more practical defense method in some cases and highlights the potential of developing defenses that target sample-specific memorization.

Table 10: Effect of the features learned by the target model to the attack performance for VGG16 trained on CelebA. Lower FID indicates better attack performance.

Attack	Target Data Size					
	1,000	2,000	5,000	10,000	15,000	20,000
PLGMI (DP)	173.735	156.141	140.480	127.090	122.218	119.856
PLGMI (Prune)	154.435	153.764	138.074	126.206	122.993	121.437
PLGMI	127.722	107.833	104.842	97.730	84.836	85.143
PLGMI (Pre)	94.603	82.571	86.860	78.429	82.768	80.814

This intuition is further supported by the observation that datasets used for fine-tuning pre-trained models are often more susceptible to reconstruction. During fine-tuning, the model, having already learned general data distributions from pre-training, tends to develop stronger memorization of the specific features unique to the fine-tuning samples. This underscores the need for caution when fine-tuning pre-trained models, as this process can heighten data exposure risks. This finding also resonates with existing research on attacks that manipulate pre-trained models to render fine-tuned versions more vulnerable to membership inference [35, 64]. Such research indicates that adversaries might modify pre-trained models to make the fine-tuning dataset easier to reconstruct, potentially by encouraging the model to memorize more sample-specific features.

Furthermore, caution should be exercised when releasing model parameters, particularly those that encapsulate statistical information about the training data, as demonstrated in Table 7. Such parameters can be exploited in reconstruction efforts.

Finally, for the advancement of reconstruction attack methodologies, researchers should aim to optimally utilize all available information. This includes strategically leveraging knowledge of the dataset distribution or statistical details embedded within model parameters. Concurrently, a complementary and important research avenue involves developing robust reconstruction techniques that can succeed even without access to such auxiliary information.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and constructive suggestions. We are especially grateful to Tianhao Wang for the in-depth discussions and valuable feedback throughout the development of this work. This work is partially funded by the European Health and Digital Executive Agency (HADEA) within the project "Understanding the individual host response against Hepatitis D Virus to develop a personalized approach for the management of hepatitis D" (DSolve, grant agreement number 101057917) and the BMBF with the project "Repräsentative, synthetische Gesundheitsdaten mit starken Privatsph ärengarantien" (PriSyn, 16KISAO29K).

Ethics Considerations

All experiments in this study were conducted using publicly available open-source datasets, ensuring that no private or sensitive data was involved. The target models were trained exclusively on open-source benchmark datasets, which were utilized solely for the purposes of this research. Furthermore, the reconstruction attacks were designed to reconstruct data from these open-source datasets only, and no attempt was made to access or infer any sensitive or personal information. This approach aligns with ethical research practices and ensures that the work complies with privacy standards and community guidelines.

Open Science

In compliance with the open science policy, we are committed to promoting transparency and reproducibility in our research. To this end, we share the artifacts associated with our work, including the data reconstruction attack framework and its evaluation code, available at https://doi.org/10.5281/zenodo.15603060. These resources are provided to facilitate further research and the development of more efficient data reconstruction attack methods, enabling researchers to better evaluate privacy leakage in machine learning models. By making these tools publicly available, we aim to contribute meaningfully to the broader scientific community and uphold the principles of open science.

References

[1] https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

[2] https://www.cs.toronto.edu/~kriz/cifar.html.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318. ACM, 2016.

[4] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Networks*, 2015.

[5] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing Training Data with Informed Adversaries. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1138–1156. IEEE, 2022.

[6] Martin Bäuml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3602–3609. IEEE, 2013.

[7] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible Residual Networks. In *International Conference on Machine Learning (ICML)*, pages 573–582. PMLR, 2019.

[8] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1897–1914. IEEE, 2022.

[9] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. The Privacy Onion Effect: Memorization is Relative. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[10] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *USENIX Security Symposium (USENIX Security)*, pages 2633–2650. USENIX, 2021.

[11] Paola Cerchiello and Paolo Giudici. Big Data Analysis for Financial Risk Management. *Journal of Big Data*, 2016.

[12] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan R. Ullman. SNAP: Efficient Extraction of Private Properties with Poisoning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1935–1952. IEEE, 2023.

[13] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-Enriched Distributional Model Inversion Attacks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16158–16167. IEEE, 2021.

[14] Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pages 1964–1974. PMLR, 2021.

[15] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep Learning for Classical Japanese Literature. *CoRR abs/1812.01718*, 2018.

[16] Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the Privacy Risk of In-context Learning. In *Workshop on Trustworthy Natural Language Processing (TrustNLP)*, 2023.

[17] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc., 2014.

[18] Vitaly Feldman. Does Learning Require Memorization? A Short Tale about a Long Tail. In *Annual ACM Symposium on Theory of Computing (STOC)*, pages 954–959. ACM, 2020.

[19] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1322–1333. ACM, 2015.

[20] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 619–633. ACM, 2018.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680. NIPS, 2014.

[22] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing Training Data from Trained Neural Networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[23] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both Weights and Connections for Efficient Neural Networks. *CoRR abs/1506.02626*, 2015.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988. IEEE, 2022.

[25] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-Doctor: Comprehensive Assessment of Membership Inference Against Machine Learning Models. *CoRR abs/2208.10445*, 2022.

[26] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-Level Membership Inference Attacks Against Graph Neural Networks. *CoRR abs/2102.05429*, 2021.

[27] Sanjay Kariyappa, Atul Prakash, and Moinuddin K. Qureshi. MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13814–13823. IEEE, 2021.

[28] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410. IEEE, 2019.

[29] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 10236–10245. NeurIPS, 2018.

[30] Pahulpreet Singh Kohli and Shriya Arora. Application of Machine Learning in Disease Prediction. In *International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4. IEEE, 2018.

[31] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. SeqMIA: Sequential-Metric Based Membership Inference Attack. *CoRR abs/2407.15098*, 2024.

[32] Zheng Li, Xinlei He, Ning Yu, and Yang Zhang. Membership Inference Attack Against Masked Image Modeling. *CoRR abs/2408.06825*, 2024.

[33] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. Auditing Membership Leakages of Multi-Exit Networks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1917–1931. ACM, 2022.

[34] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021.

[35] Ruixuan Liu, Tianhao Wang, Yang Cao, and Li Xiong. PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3511–3524. ACM, 2024.

[36] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership Inference Attacks by Exploiting Loss Trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2085–2098. ACM, 2022.

[37] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 4525–4542. USENIX, 2022.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9992–10002. IEEE, 2021.

[39] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–150. ACL, 2011.

[40] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property Inference from Poisoning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1120–1137. IEEE, 2022.

[41] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting Unintended Feature Leakage in Collaborative Learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 497–512. IEEE, 2019.

[42] John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text Embeddings Reveal (Almost) As Much As Text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12448–12460. Association for Computational Linguistics, 2023.

[43] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable Private Learning with PATE. In *International Conference on Learning Representations (ICLR)*, 2018.

[44] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In *USENIX Security Symposium (USENIX Security)*, pages 1291–1308. USENIX, 2020.

[45] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella Béguelin. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 327–345. IEEE, 2023.

[46] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[47] Sunandini Sanyal, Sravanti Addepalli, and R. Venkatesh Babu. Towards Data-Free Model Stealing in a Hard Label Setting. *CoRR abs/2204.11022*, 2022.

[48] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 3–18. IEEE, 2017.

[49] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. ACL, 2013.

[50] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine Learning Models that Remember Too Much. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 587–601. ACM, 2017.

[51] Congzheng Song and Vitaly Shmatikov. Overlearning Reveals Sensitive Attributes. In *International Conference on Learning Representations (ICLR)*, 2020.

[52] Yang Song, Chenlin Meng, and Stefano Ermon. MintNet: Building Invertible Neural Networks with Masked Convolutions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 11002–11012. NeurIPS, 2019.

[53] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks. In *International Conference on Machine Learning (ICML)*, pages 20522–20545. PMLR, 2022.

[54] Anshuman Suri and David Evans. Formalizing and Estimating Distribution Inference Risks. *CoRR abs/2109.06024*, 2021.

[55] Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. *CoRR abs/1503.02406*, 2015.

[56] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2022.

[57] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium (USENIX Security)*, pages 601–618. USENIX, 2016.

[58] Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-Free Model Extraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4771–4780. IEEE, 2021.

[59] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard S. Zemel, and Alireza Makhzani. Variational Model Inversion Attacks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9706–9719. NeurIPS, 2021.

[60] Rui Wen, Michael Backes, and Yang Zhang. Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm? In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

[61] Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership Inference Attacks Against In-Context Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.

[62] Rui Wen, Yiyong Liu, Michael Backes, and Yang Zhang. SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark. *CoRR abs/2506.07888*, 2025.

[63] Rui Wen, Tianhao Wang, Michael Backes, Yang Zhang, and Ahmed Salem. Last One Standing: A Comparative Analysis of Security and Privacy of Soft Prompt Tuning, LoRA, and In-Context Learning. *CoRR abs/2310.11397*, 2023.

[64] Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. Privacy Backdoors: Enhancing Membership Inference through Poisoning Pre-trained Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2024.

[65] Yixin Wu, Rui Wen, Michael Backes, Pascal Berrang, Mathias Humbert, Yun Shen, and Yang Zhang. Quantifying Privacy Risks of Prompts in Visual Prompt Learning. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[66] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, page 225–240. ACM, 2019.

[67] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[68] Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8712–8721. IEEE, 2020.

[69] Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 2023.

[70] Zhuowen Yuan, Fan Wu, Yunhui Long, Chaowei Xiao, and Bo Li. SecretGen: Privacy Recovery on Pre-trained Models via Distribution Discrimination. In *European Conference on Computer Vision (ECCV)*, pages 139–155. Springer, 2022.

[71] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 4827–4837. IEEE, 2024.

[72] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 649–657. NIPS, 2015.

[73] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 250–258. IEEE, 2020.

[74] Ligeng Zhu, Zhijian Liu, and Song Han. Deep Leakage from Gradients. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 14747–14756. NeurIPS, 2019.

A Transferability to Larger Architectures

We extend our analysis of the relationship between memorization and reconstruction performance to larger, contemporary model architectures—specifically, transformer-based models that are widely used in modern machine learning practice. Our experiments with Swin [38] and MAE [24] architectures, as presented in Table 8 and Table 11, reveal consistent trends: models trained on smaller datasets tend to be more vulnerable, in line with our previous findings.

We also observe that the worst-case performance does not occur at the smallest dataset size. This can be attributed to the nature of transformer-based models, which typically require substantially larger datasets to train effectively. As the dataset grows, the model begins to capture sample-specific features, leading to increased unnecessary memorization. This nuanced behavior aligns with our analysis in Section 6.1.

B Experimental Setup for the Text Modality

Our experimental evaluation is performed on three established text datasets: SST2 [49], AGNews [72], and IMDB [39]. For fine-tuning the target models, we utilize subsets of these datasets ranging from 1,000 to 20,000 samples. This focus on fine-tuning, rather than training language models from scratch, is adopted due to the substantial data requirements of the latter and aligns with common practices in recent literature [16, 63].

We investigate two distinct attack methodologies:

1. **Vec2Text Attack [42]:** In this configuration, the `sentence-transformers/gtr-t5-base` model is fine-tuned. The corresponding decoder is a `T5-base` model, which is further trained using an auxiliary dataset drawn from the same data distribution as the fine-tuning set.

2. **Complete Attack [10, 61]:** For this attack, the `GPT2` model is fine-tuned for 20 epochs. To reconstruct the original input, the fine-tuned model is queried using the first three words of the said input as a prompt.

Detailed results of these attack evaluations are presented in Table 9 and Table 12. To assess the quality of the reconstructed text, we employ several metrics. Semantic similarity is quantified by the cosine similarity between sentence embeddings; these embeddings are generated using the `sentence-transformers/all-MiniLM-L6-v2` model. Additionally, we report BLEU scores to measure n-gram precision and ROUGE-L scores to evaluate the longest common subsequence, thereby providing insights into the lexical overlap and fluency of the reconstructed outputs.

Table 11: Evaluation on MAE with GPT-4o.

Attack	Metrics	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
PLGMI	# of Major	79	233	312	275	76	25
	# of All	7	12	24	24	0	1
	Pred Rate	0.097	0.222	0.299	0.267	0.090	0.025

Table 12: Evaluation of the Complete [10, 61] attack performance across varying target data sizes on different datasets, measured by BLEU, Sim. (Similarity), and R-L (ROUGE-L). Higher scores on these metrics correspond to better attack performance.

Dataset	Metric	Target Data Size					
		1,000	2,000	5,000	10,000	15,000	20,000
SST2	BLEU	0.606	0.527	0.450	0.115	0.103	0.097
	Sim.	0.875	0.774	0.721	0.496	0.449	0.417
	R-L	0.777	0.651	0.575	0.229	0.205	0.203
AGNews	BLEU	0.412	0.397	0.370	0.328	0.291	0.258
	Sim.	0.865	0.859	0.848	0.825	0.806	0.784
	R-L	0.576	0.563	0.536	0.493	0.452	0.415
IMDB	BLEU	0.344	0.282	0.123	0.051	0.035	0.029
	Sim.	0.747	0.697	0.552	0.463	0.442	0.431
	R-L	0.557	0.481	0.272	0.182	0.164	0.157

C Related Work

C.1 Membership Inference Attack

Membership inference attack reveals the membership status of a target sample, i.e., whether the target sample is in the training dataset or not, which leads to a direct privacy breach.

Shokri et al. [48] proposed the seminal work on membership inference attack against machine learning models, wherein several shadow models were trained to imitate the behavior of the target model. This attack requires access to data from the same distribution as the training dataset. Later, Salem et al. [46] relax the assumption of the same distribution and demonstrate the effectiveness of using only one shadow model, largely reducing the computational cost required. Subsequent research [14, 34] explores a more challenging setting where the adversary only has hard-label access to the target model. Specifically, Li and Zhang [34] utilize adversary examples to approximate the distance between the target sample to its decision boundary in order to make decisions based on this distance. Recently, more work [8, 36, 56] aims at enhancing the performance of membership inference attacks. For example, Carlini et al. [8] take advantage of the discrepancy of models trained with and without the target sample. Liu et al. [36] demonstrate the effectiveness of loss trajectory.

C.2 Other Privacy Attacks

Property inference attack differs from data reconstruction and membership inference attack as it aims to infer macro-level information about the target dataset, such as the gender proportion. Ateniese et al. [4] presented the first property inference attacks against Hidden Markov Models (HMMs) and Support Vector Machine (SVM), which was later extended to Fully Connected Neural Networks (FCNNs) by Ganju et al. [20]. Both attacks rely on a meta-classifier to infer the property of the training dataset using white-box access to the target model. Training the meta-classifier requires multiple shadow models, making it computationally expensive.

Suri and Evans [54] first formalized the property inference attack and provided a method for conducting the attack with black-box access to the target model. Subsequent research [12, 40], has focused on improving the performance of property inference attacks by adding a small amount of "poisoned" data to the training dataset. For example, Chaudhari et al. [12] select a limited number of samples in one class, and flip their labels to increase the discrepancy of posterior distributions for different properties.

Additionally, model stealing attacks aim to construct a local surrogate model from the target model. Tramèr et al. [57] propose the first attack against neural networks by querying the target model to construct the training dataset. However, their attack demands high-quality data to be effective, which motivates recent research to relax this assumption through the development of data-free attack paradigms [27, 47, 58].

C.3 Memorization and Privacy Leakage

Song et al. [50] demonstrate that malicious trainers can easily encode training samples into the model parameters, which can later be extracted. This highlights the potential threats for model leakage and emphasizes the need for researchers to consider the security implications of memorization in their models. Song and Shmatikov [51] further develop this point by uncovering the intrinsic behavior of models, specifically their tendency to retain information that is not pertinent to the designed classification task. Despite attempts to eliminate this extraneous information, the unintentional disclosure of sensitive data remains a persistent issue.

Several studies have leveraged memorization to enhance the attack performance. Tramèr et al. [56] show that with access to a tiny fraction of the training dataset, the adversary can boost the performance of membership inference by poisoning data samples as the memorization of these poisoned samples increases. Wen et al. [60] investigate the relationship between data importance (used as a proxy for memorization) and privacy attacks, observing that data samples with higher importance exhibit increased vulnerability to certain attacks.