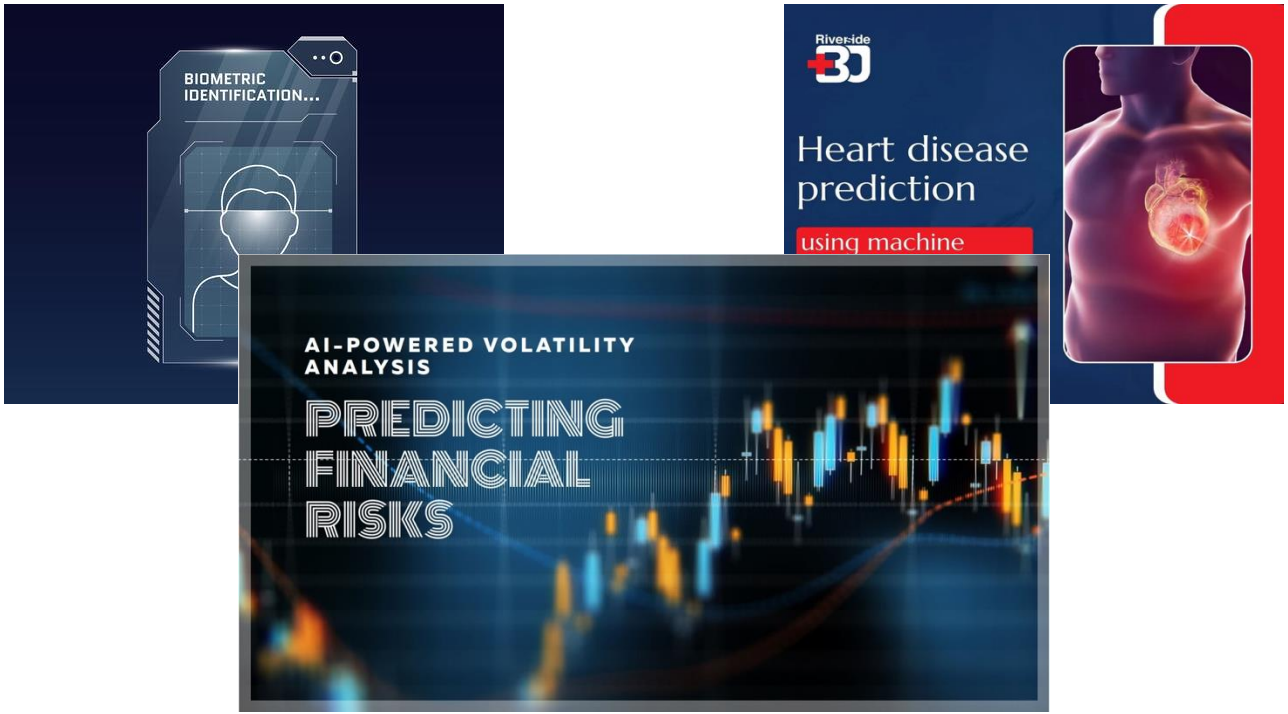
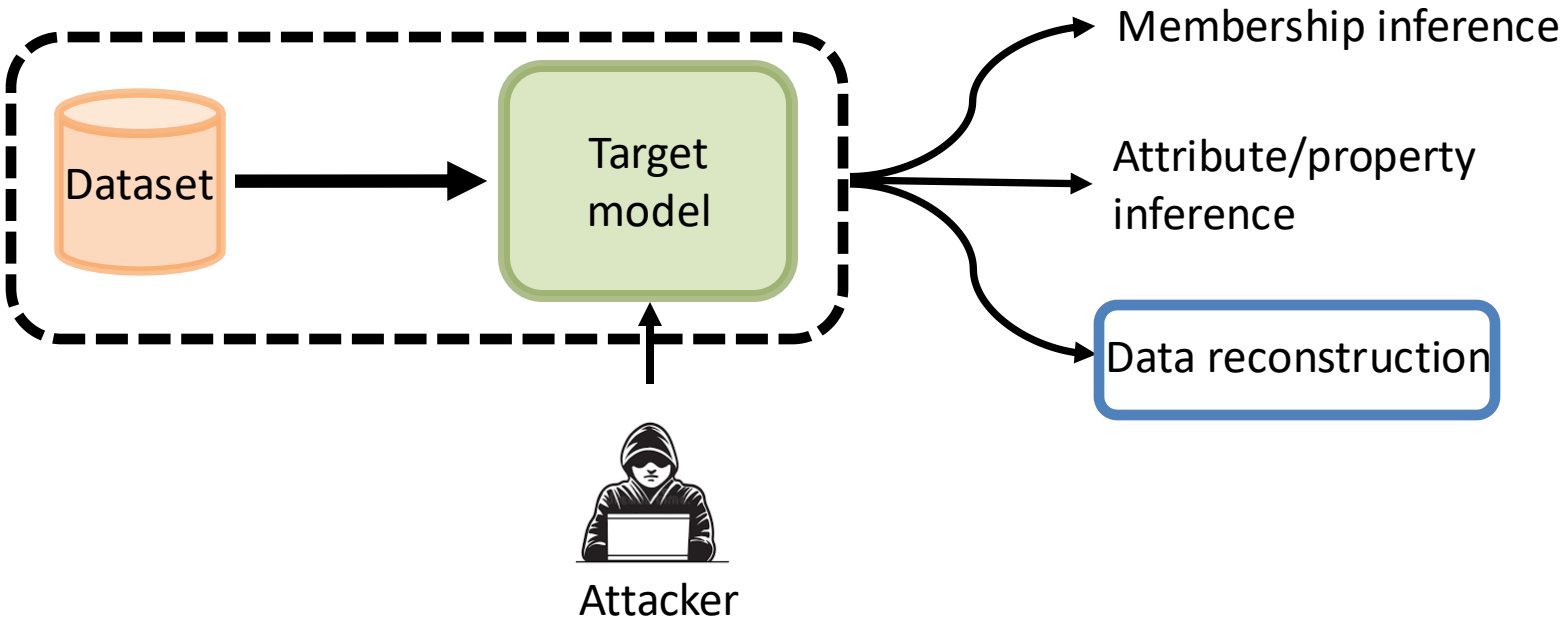
SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark

Rui Wen*, Yiyong Liu*, Michael Backes, Yang Zhang

Privacy-crucial machine learning tasks



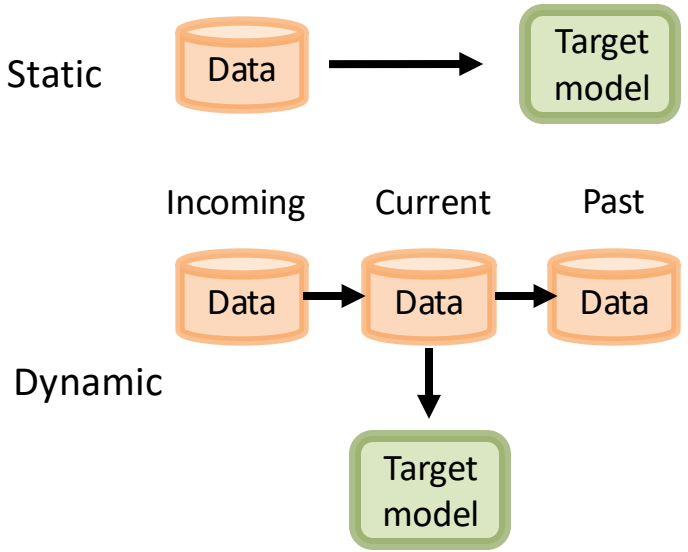
The ultimate privacy breach



The limitation of data reconstruction

No rigorous definition

- Training type
  - Static[1] or dynamic[2]



[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Salem et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. USENIX Security 2020.

The limitation of data reconstruction

No rigorous definition

- Training type
  - Static[1] or dynamic[2]
- Model access
  - Black-box[3] or white-box[1]

[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Salem et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. USENIX Security 2020.
[3] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.

The limitation of data reconstruction

No rigorous definition

- Training type
  - Static[1] or dynamic[2]
- Model access
  - Black-box[3] or white-box[1]
- **Dataset access**
  - **No data[1] or similar[4] or same distribution[4]**

Target dataset

Auxiliary dataset

Similar

Same

[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Salem et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. USENIX Security 2020.
[3] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.
[4] Yuan et al. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. AAAI 2023.

Defining data reconstruction

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		

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., $\mathcal{A}$: $\mathcal{M} \times \mathcal{K} \rightarrow \mathbb{D}$, with the same size as the target dataset $\mathcal{D}_{tar}$.

Information the adversary have

Data the adversary aim to reconstruct

The limitation of data reconstruction

No proper evaluation metrics

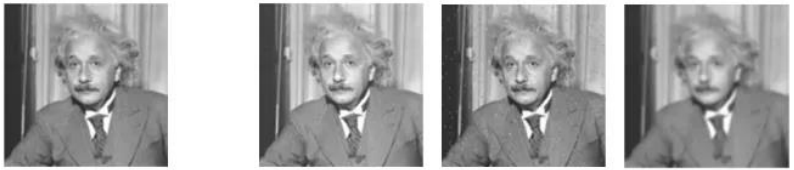
- Visualization[1]

[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.

The limitation of data reconstruction

No proper evaluation metrics

- Visualization[1]
- MSE/PSNR/SSIM[2]

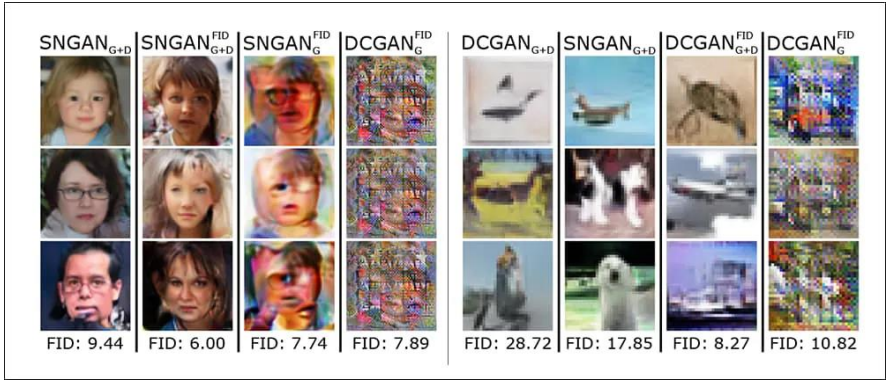


[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.

The limitation of data reconstruction

No proper evaluation metrics

- Visualization[1]
- MSE/PSNR/SSIM[2]
- Feature distance[3]

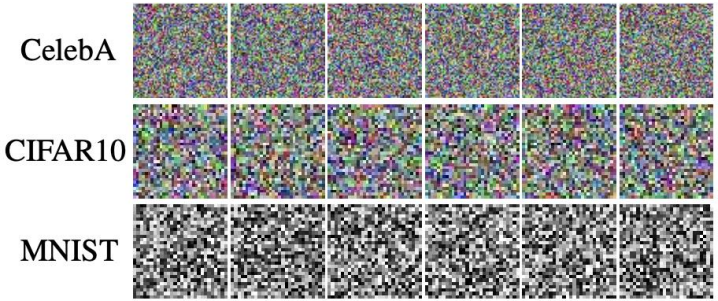


[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.
[3] Chen et al. Knowledge-Enriched Distributional Model Inversion Attacks. ICCV 2021,

The limitation of data reconstruction

No proper evaluation metrics

- Visualization[1]
- MSE/PSNR/SSIM[2]
- Feature distance[3]
- Accuracy (train)[4]



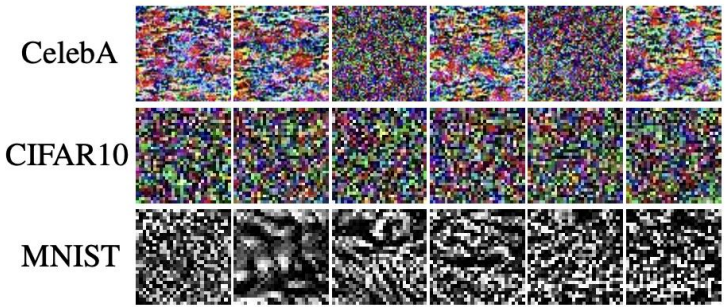
(a) Accuracy (Train)

[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.
[3] Chen et al. Knowledge-Enriched Distributional Model Inversion Attacks. ICCV 2021,
[4] Zhang et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. CVPR 2020.

The limitation of data reconstruction

No proper evaluation metrics

- Visualization[1]
- MSE/PSNR/SSIM[2]
- Feature distance[3]
- Accuracy (train)[4]
- Accuracy (test)[4]



(b) Accuracy (Test)

[1] Fredrikson et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. ACM CCS 2015.
[2] Yang et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. ACM CCS 2019.
[3] Chen et al. Knowledge-Enriched Distributional Model Inversion Attacks. ICCV 2021,
[4] Zhang et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. CVPR 2020.

Designing evaluation metrics

- Quantifiability
  - Provide quantitative results and eliminate the influence of subjective factors
- Consistency
  - The evaluation results should be consistent and determined
- Precision
  - Capture the sample-level similarity of reconstructed samples to target samples
- Diversity
  - Reflect the percentage of data being reconstructed

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

Designing evaluation metrics

- **Quantifiability**
  - Provide quantitative results and eliminate the influence of subjective factors
- **Consistency**
  - The evaluation results should be consistent and determined
- **Precision**
  - Capture the sample-level similarity of reconstructed samples to target samples
- **Diversity**
  - Reflect the percentage of data being reconstructed

Definition 4.2 (Sample-level Metric)**.** The sample-level metric $\mu : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}^2$ 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.

Designing evaluation metrics

- **Quantifiability**
  - Provide quantitative results and eliminate the influence of subjective factors
- **Consistency**
  - The evaluation results should be consistent and determined
- **Precision**
  - Capture the sample-level similarity of reconstructed samples to target samples
- **Diversity**
  - Reflect the percentage of data being reconstructed

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

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.

Designing evaluation metrics

- **Quantifiability**
  - Provide quantitative results and eliminate the influence of subjective factors
- **Consistency**
  - The evaluation results should be consistent and determined
- **Precision**
  - Capture the sample-level similarity of reconstructed samples to target samples
- **Diversity**
  - Reflect the percentage of data being reconstructed

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.

Experimental settings

- Datasets
  - CelebA, CIFAR10 and MNIST
- Attack details
  - Ten attacks
  - Balanced dataset, sizes 100->20000
  - Disjoint auxiliary dataset
- Metric details
  - Sample-level: SSIM, PSNR and MSE
  - Dataset-level: FID



Visualization results

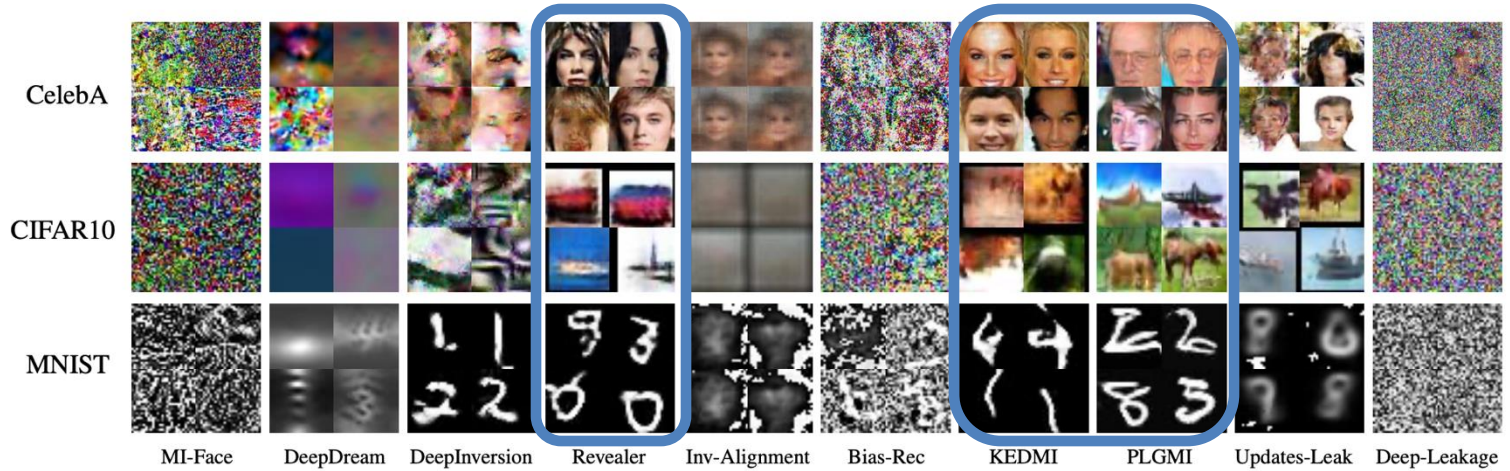


Figure 1: Visualization of existing reconstruction attacks. For each attack, the left two images are reconstructed from the target model with a smaller training size (1,000 for CelebA and 100 for CIFAR10 and MNIST), and the right two images are from the larger one (20,000).

Quantitative results

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
Memorization			**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)



Quantitative results

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
Memorization			**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Quantitative results

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

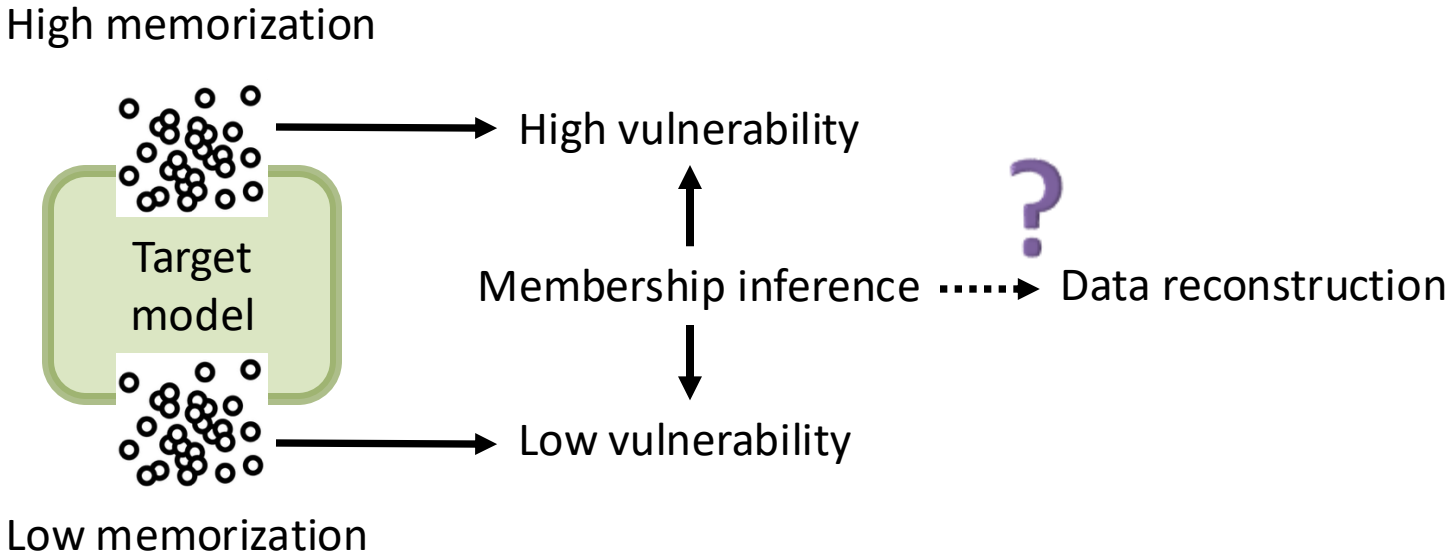
Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
	Memorization		**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Quantitative results

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
	Memorization		**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Model memorization vs. dataset leakage



[1] Carlini et al. The Privacy Onion Effect: Memorization is Relative. NeurIPS 2022.
[2] Tramèr et al. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. ACM CCS 2022.

Definition of model memorization

Leave-one-out change in the label[1]

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

[1] Feldman et al. Does Learning Require Memorization? A Short Tale about a Long Tail. ACM STOC 2020.

Results of model memorization

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
Memorization			**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Results of model memorization

Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
	Memorization		**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Not effective

Results of model memorization

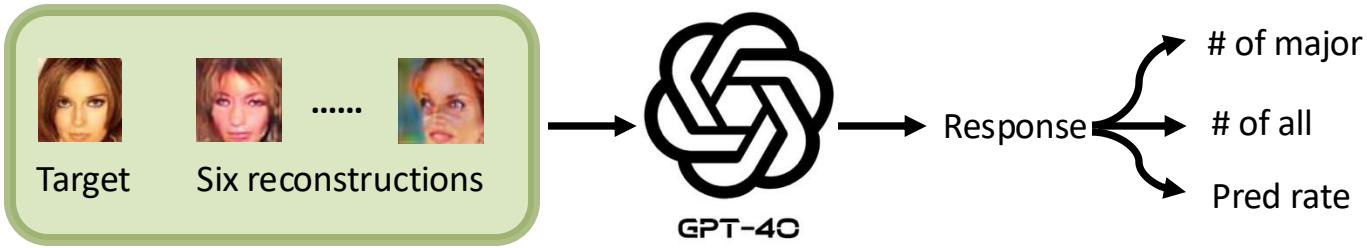
Table 2: Evaluation results of existing reconstruction attacks. The target model is VGG16 trained on CelebA. For FID and MSE, a lower score indicates better reconstruction quality; while for SSIM and PSNR, a higher score indicates better performance.

Attack	Metrics		Target Data Size			
			1,000	5,000	10,000	20,000
Memorization			**1.000**	**0.862**	**0.539**	**0.301**
DeepInversion	Dataset-level	FID ↓	287.497	273.183	273.415	234.672
	Sample-level	SSIM ↑	0.100(100.00%)	0.118(42.44%)	0.140(28.97%)	0.153(22.47%)
		PSNR ↑	9.676(100.00%)	10.550(31.68%)	11.143(21.07%)	11.343(15.73%)
		MSE ↓	0.119(100.00%)	0.094(31.68%)	0.081(21.07%)	0.077(15.73%)
Revealer	Dataset-level	FID ↓	116.712	94.899	93.961	92.982
	Sample-level	SSIM ↑	0.101(100.00%)	0.135(50.50%)	0.150(44.05%)	0.162(38.33%)
		PSNR ↑	9.144(100.00%)	10.087(43.68%)	10.449(37.20%)	10.733(32.17%)
		MSE ↓	0.312(100.00%)	0.103(43.68%)	0.094(37.20%)	0.088(32.17%)
Inv-Alignment	Dataset-level	FID ↓	344.049	359.419	229.609	357.910
	Sample-level	SSIM ↑	0.255(100.00%)	0.285(22.56%)	0.353(13.09%)	0.328(7.89%)
		PSNR ↑	11.292(100.00%)	13.023(23.10%)	13.858(13.93%)	14.253(8.02%)
		MSE ↓	0.081(100.00%)	0.052(23.10%)	0.043(13.93%)	0.039(8.02%)
PLGMI	Dataset-level	FID ↓	127.722	104.842	97.730	85.143
	Sample-level	SSIM ↑	0.110(100.00%)	0.136(37.02%)	0.149(26.15%)	0.161(18.54%)
		PSNR ↑	9.448(100.00%)	10.588(31.58%)	10.921(20.55%)	11.513(13.54%)
		MSE ↓	0.122(100.00%)	0.092(31.58%)	0.085(20.55%)	0.074(13.54%)

Not effective

Evaluation with LLMs

- Reason
  - Human inspection is expensive
  - GPT-4o aligns with human preferences
- Evaluation details
  - Input seven images
  - Run five trials with the prompt
  - Three metrics: “# of major”, “# of all” and “pred rate”



Evaluation with LLMs

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

Figure 3: Visualization of PLGMI on different target models.

Evaluation with LLMs

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

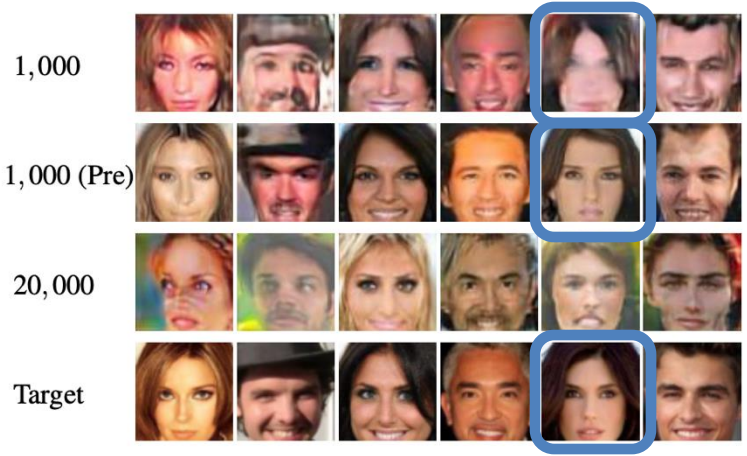


Figure 3: Visualization of PLGMI on different target models.

Evaluation with LLMs

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

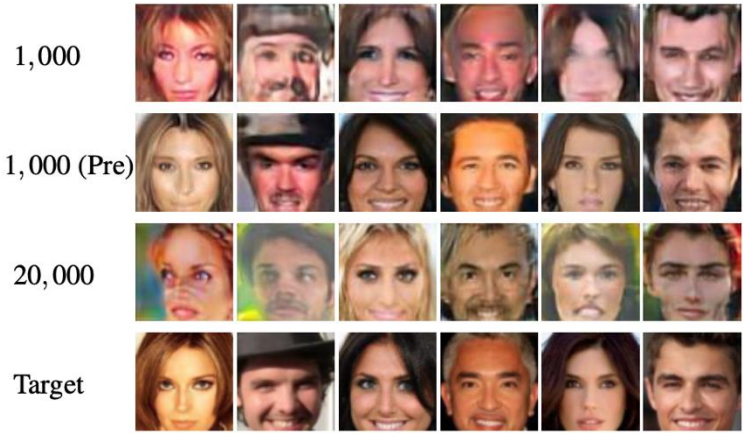


Figure 3: Visualization of PLGMI on different target models.

Evaluation with LLMs

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

Figure 3: Visualization of PLGMI on different target models.

More discussions

- Influence of data access
  - Low-quality reconstruction methods fail to exploit critical attack information

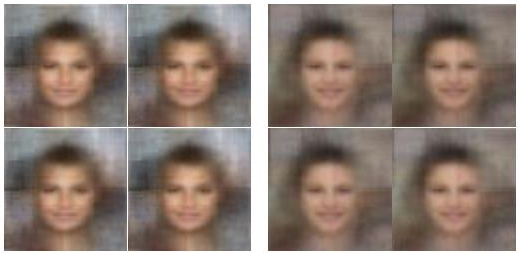


Figure 5: Effect of additional data to the attack performance of Inv-Alignment.

More discussions

- Influence of data access
  - Low-quality reconstruction methods fail to exploit critical attack information
- Influence of model access
  - Certain attack methods can't fully leverage model-internal information

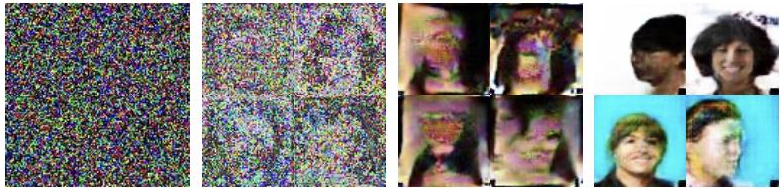


Figure 6: Visualization of attacks on VGG16 trained on backdoored CelebA with size 20,000.

More discussions

- Influence of data access
  - Low-quality reconstruction methods fail to exploit critical attack information
- Influence of model access
  - Certain attack methods can't fully leverage model-internal information
- Extend the analysis to text domain
  - Memorization plays a critical role in determining vulnerability to reconstruction

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

More discussions

- Influence of data access
  - Low-quali… …nation
- Influence of…
  - Certain at…
- Extend the…
  - Memorization plays a critical role in determining vulnerability to reconstruction

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

- Attack and defense strategies based on model memorization
  - Reduce or encourage the memorization can be the direction for designing the attack and defense methods in the future

Thanks!

Welcome to use our framework: https://doi. org/10.5281/zenodo.15603060