Towards More Realistic Membership Inference Attacks on Large Diffusion Models

Jan Dubiński[1], Antoni Kowalczuk[1,2], Stanisław Pawlak[1], Przemysław Rokita[1], Tomasz Trzciński[1,3,4], Paweł Morawiecki[5]

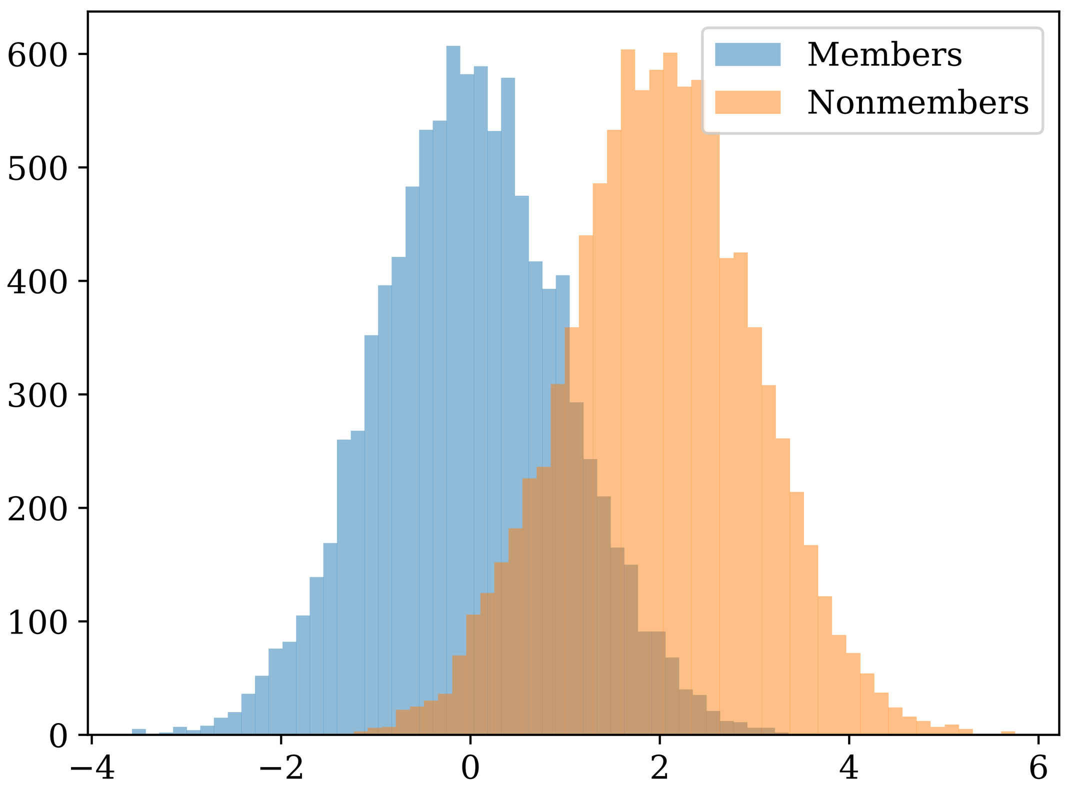
[1]Warsaw University of Technology, [2]AI Society Golem, [3]IDEAS NCBR, [4]Tooploox, [5]Polish Academy of Sciences

TL;DR

- We identify the **pitfalls** of existing approaches to *membership inference attacks* on large diffusion models.
- We provide a new **dataset** along with construction methodology.
- We propose a **fair** and **rigorous** evaluation protocol on the **SOTA Stable Diffusion model**.
- We thoroughly evaluate a set of *MIAs* using our dataset and methodology.

Membership Inference Attacks

Was this example used to train the model? **Yes** or **No**?



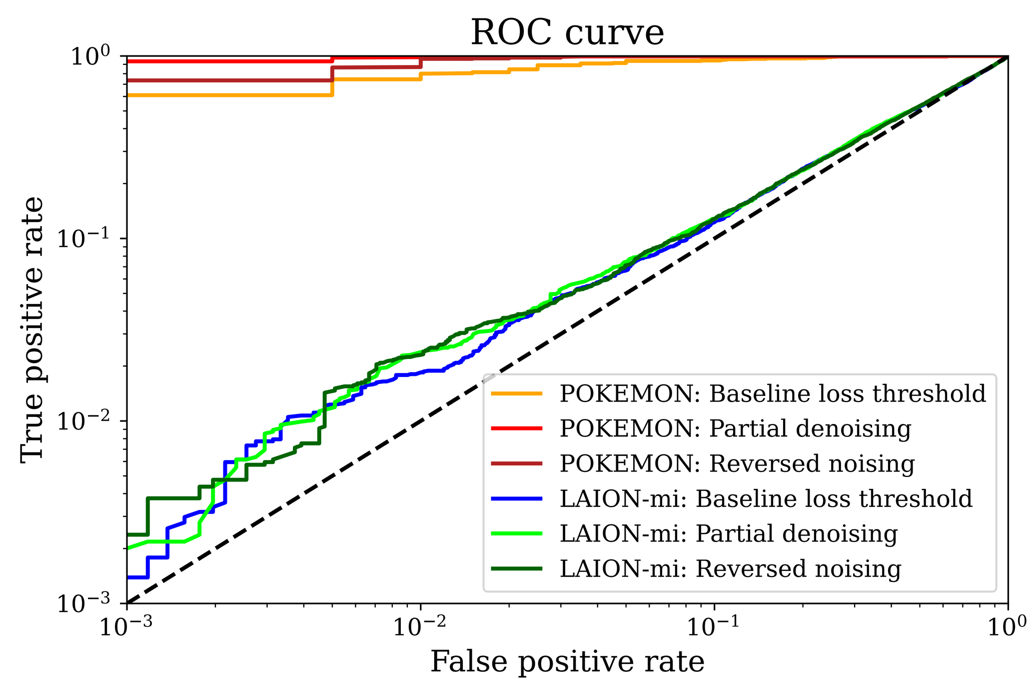
Loss Threshold Attack: $IF\ loss(sample) < threshold\ THEN\ member\ ELSE\ nonmember.$

Problem: Lack of Nonmembers Set

We cannot run MIA evaluation without nonmembers. A few approaches has been proposed:

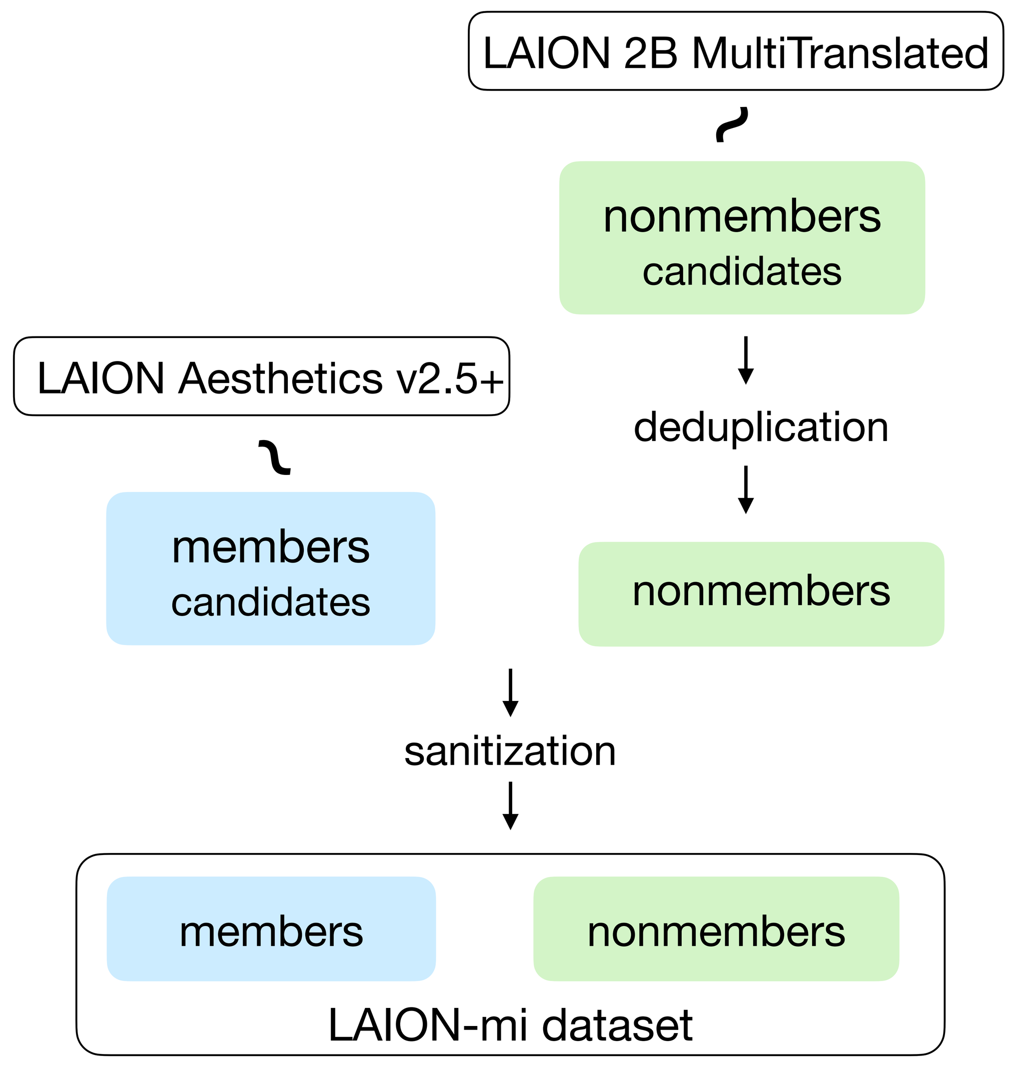
1. Fine-tune Stable Diffusion on a new dataset[1]. **Pitfall**: too trivial problem due to overfitting.
2. Train a new model on a new dataset. **Problem**: too expensive.
3. Create a dataset with similar properties to the original one. **Challenge**: distribution mismatch.

Pitfall: Fine-tuning



Pitfalls in the evaluation setting can lead to incorrect conclusions on the effectiveness of *membership inference attacks* against large diffusion models such as Stable Diffusion.

Solution: LAION-mi Dataset

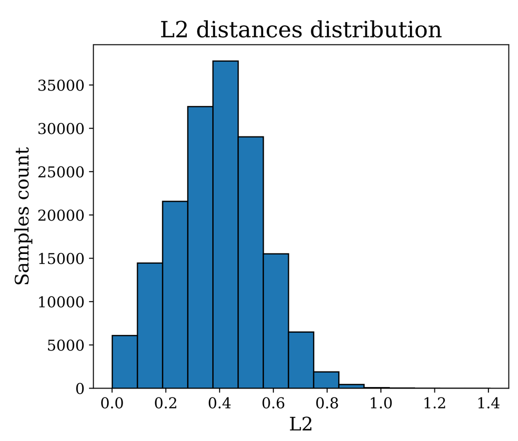


A general scheme of constructing LAION-mi dataset.

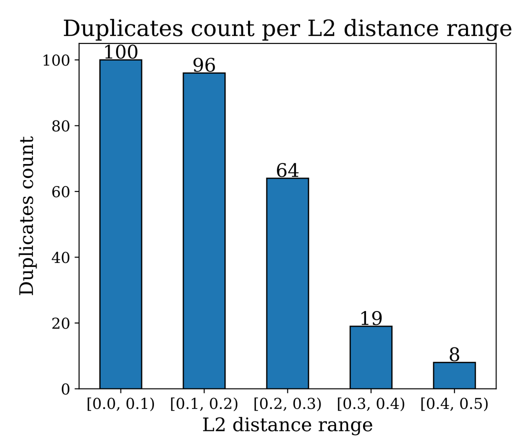
Challenges

- Duplicates: LAION-2B EN constains 30% duplicates[2].
- Distribution mismatch: LAION-2B EN and LAION-2B Multi Translated may have different distributions.

Deduplication



Distribution of L2 distances.



Duplicates count per 100 samples for L2 distances buckets.

Evaluation

Scenario	Loss	Method	TPR@FPR=1%. ↑ LAION-mi	POKEMON
White-box	Model Loss	Baseline loss thr.	1.92%±0.59	80.9%±2.27
		Reversed noising	2.51%±0.73	97.3%±0.93
		Partial denoising	2.31%±0.61	94.5%±1.34
		Reversed denoising	2.25%±0.64	91.5%±1.63
	Latent Error	Reversed noising	1.26%±0.62	11.5%±1.84
		Partial denoising	2.42%±0.62	99.5%±0.4
		Reversed denoising	2.17%±0.64	61.1%±2.74
	Pixel Error	Reversed noising	1.90%±0.51	8.36%±1.66
		Reversed denoising	2.03%±0.55	12.0%±1.97
		Partial denoising	1.75%±0.68	25.38%±2.55
Grey-box	Latent Error	Generation from prompt	0.93%±0.41	7.15%±1.5
Black-box	Pixel Error	Generation from prompt	0.35%±0.19	12.0%±1.9

References

[1] Jinhao Duan et al. *Are Diffusion Models Vulnerable to Membership Inference Attacks?* 2023. arXiv: 2302.01316 [cs.CV].

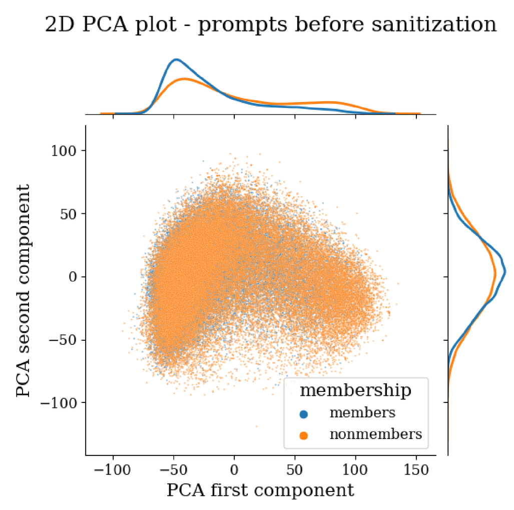
[2] Ryan Webster et al. *On the De-duplication of LAION-2B.* 2023. arXiv: 2303.12733 [cs.CV].

Distribution Mismatch

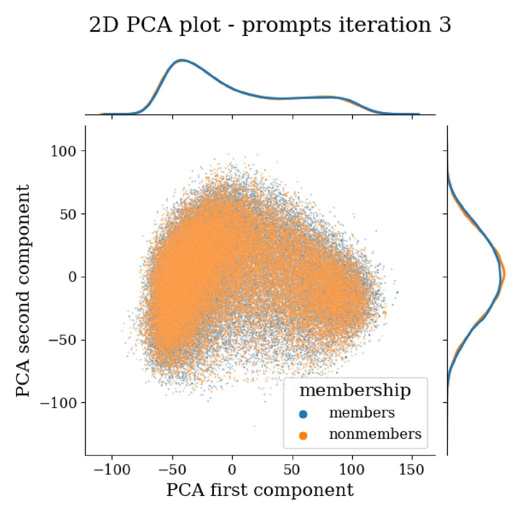
We assess the mismatch using the following metrics:

- Visual inspection of the PCA projection of the dataset.
- FID score between the subsets.
- Classifier-based evaluation.

Sanitization: PCA Projection



Prompts before sanitization.

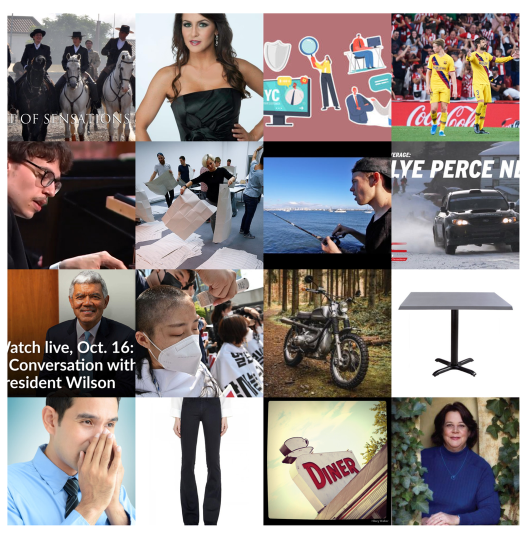


Prompts after sanitization.

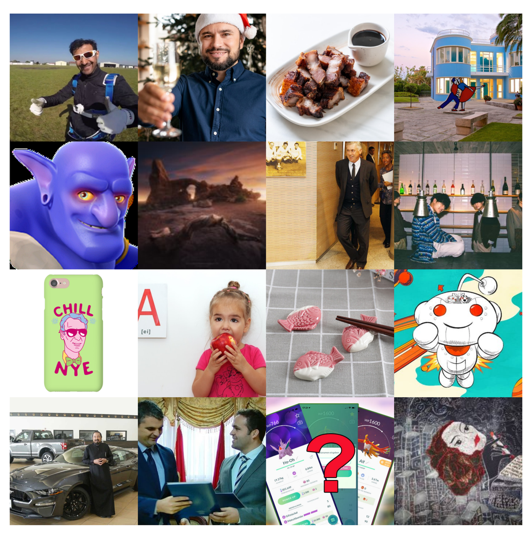
Sanitization: FID

Data subset	FID text	FID images
Members internal - random	9.84	7.00
Members internal - sanitized	9.77	7.06
Nonmembers internal	9.73	7.01
Comparative - random	66.43	13.90
Comparative - sanitized	**13.54**	**8.87**

LAION-mi Samples



Sample 16 members.



Sample 16 nonmembers.

Paper

Contact: antonikowalcz@gmail.com