IMPERIAL

Tracing the Roots: Leveraging Temporal Dynamics in Diffusion Trajectories for Origin Attribution

Andreas Floros, Seyed-Mohsen Moosavi-Dezfooli, Pier Luigi Dragotti

NeurIPS 2025, San Diego

Motivation

Toward Responsible Generative Modeling

Membership Inference
Given a trained model, the goal is to determine whether a given sample was in the training set.

- “Is my data used without permission?”
- “If so, is it secure?”

Representative MIAs threshold the training loss:

Intuitively, training samples should achieve smaller loss than non-training samples.

Model Attribution
Given a data sample, the aim is to determine, which, if any, generative model produced it.

- “Is this fake?”
- “If yes, who is responsible for it?”

Representative MA relies on reconstruction:

If a model generated the sample, there exists a latent that perfectly reconstructs it.

General Origin Attribution: Given models and data, what relationships, if any, exist between them?

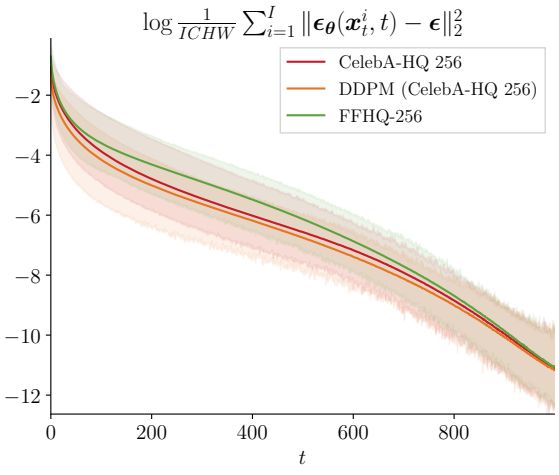
Challenges

Representative MIAs Cannot Audit Synthetic Data

Model Generations Minimize the NELBO

DDPMs show that, for a particular weighting, the loss corresponds to the NELBO of data.

- By construction, model generations should achieve the smallest diffusion loss.
- Thresholding MIAs cannot be used to audit synthetic samples or filter for memorization.
- Therefore, they are, at best, strong non-membership inference attacks.



Challenges

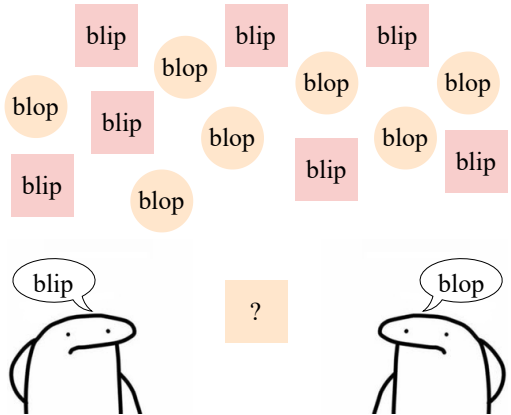
Blind-Baselines May Beat Representative MIAs

Quantifying Distribution Shifts for MIAs on CIFAR-10 & CelebA-HQ DDPMs

CIFAR-10 / CIFAR-10.1	AUC	TPR @ 1% FPR	ASR
Naive (model-blind)	52.2	0.0	52.0
Matsumoto et al. (2023)	63.2	3.3	59.7
Kong et al. (2024)	**66.9**	**5.1**	**62.4**

CelebA-HQ / FFHQ	AUC	TPR @ 1% FPR	ASR
Naive (model-blind)	**94.4**	**60.1**	**86.6**
Matsumoto et al. (2023)	85.2	26.4	76.2
Kong et al. (2024)	62.5	0.1	58.1

Model-blind baselines, with no real predictive power, may outperform engineered MIAs.



Leveraging Temporal Dynamics in Diffusion Trajectories

An Alternative to the Goldilocks Zone Conjecture

If t is large, and the noisy image is similar to noise, then predicting the added noise is easy regardless if the input was in the training set; if t is small, and so the noisy image is similar to the original, then the task is too difficult. It is hypothesized that there exists a "Goldilocks zone" for membership inference (Carlini et al., 2023).

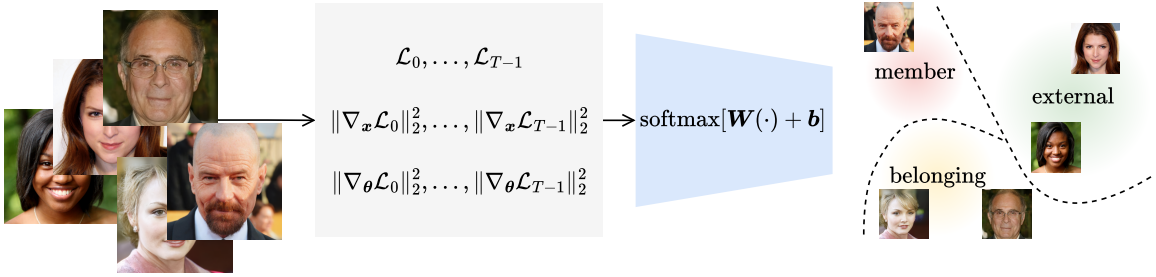
$t = 0, \ldots, 249$	$t = 250, \ldots 499$	$t = 500, \ldots, 749$	$t = 750, \ldots, 999$	$t = 0, 4, \ldots, 996$
68.5	68.1	54.7	50.7	**71.2**

Global Temporal Context Is Important

- Fix the number of queries to the diffusion model and consider classifiers operating on features $\{\mathcal{L}_t\}_{t=0}^{T-1}$.
- The best strategy (AUC) is to allocate the time-steps such that they cover the entire diffusion process.

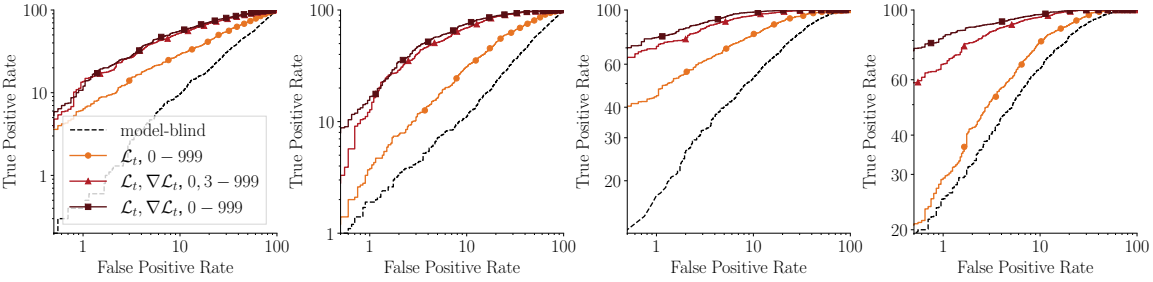
Leveraging Temporal Dynamics in Diffusion Trajectories

Our Assumptions & Overall Pipeline



Toward Origin Attribution

- We relax and quantify our assumptions, only requiring minimal data access for development (<3.4%).
- We consider a simple pipeline for modeling the diffusion trajectories based on the loss and its gradients.



Member (CIFAR-10) Belonging (CIFAR-10) Member (CelebA-HQ) Belonging (CelebA-HQ)

Model Attribution

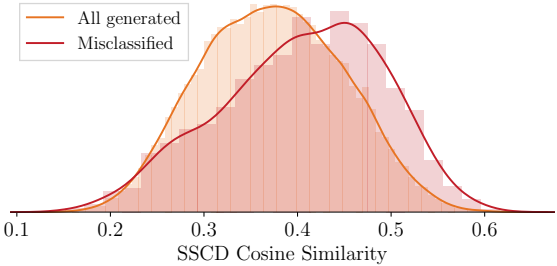
Classifiers trained only on DDPM and real data.

CIFAR-10			DDPM	DDIM	DDGAN	WDiff	Avg
Naive (model-blind)			60.8	39.1	**41.9**	44.0	51.2
Our method							
time-steps	$\mathcal{L}_t$	$\nabla\mathcal{L}_t$					
$0, 1, \ldots, 999$	✓		75.5	57.6	36.2	64.4	64.1
$0, 3, \ldots, 999$	✓	✓	86.0	72.4	18.1	76.1	70.8
$0, 1, \ldots, 999$	✓	✓	**87.2**	**86.8**	13.1	**91.7**	**75.5**

CelebA-HQ			DDPM	DDIM	DDGAN	WDiff	Avg
Naive (model-blind)			88.3	**20.2**	22.6	27.0	55.8
Our method							
time-steps	$\mathcal{L}_t$	$\nabla\mathcal{L}_t$					
$0, 1, \ldots, 999$	✓		98.9	6.3	59.6	57.8	70.0
$0, 3, \ldots, 999$	✓	✓	**100.0**	7.9	74.2	**76.9**	**76.5**
$0, 1, \ldots, 999$	✓	✓	**100.0**	3.5	**86.8**	68.9	**76.5**

Data Extraction Filter

30k generations filtered to 1.7k.



When our system misclassifies generated samples (left) as training data, they tend to be similar to samples from the training set (right).

Discussion
Our Recommendations

Revisit the threat models

- Reliance on surrogate and foundation models leads to opaque assumptions.
- We argue for methodological purity: a step toward practical origin attribution methods.

Embrace distribution shifts

- Benchmarking on idealized datasets and sanitization does not reflect reality.
- Methodological effectiveness is relative to appropriate baselines, not absolute.

Focus on data extraction

- Standard metrics do not necessarily reflect privacy and security vulnerabilities.
- Data extraction is an undeniable proof of real risks: focus on this instead.

IMPERIAL

Thank you

Tracing the Roots: Leveraging Temporal Dynamics in Diffusion Trajectories for Origin Attribution
NeurIPS 2025, San Diego