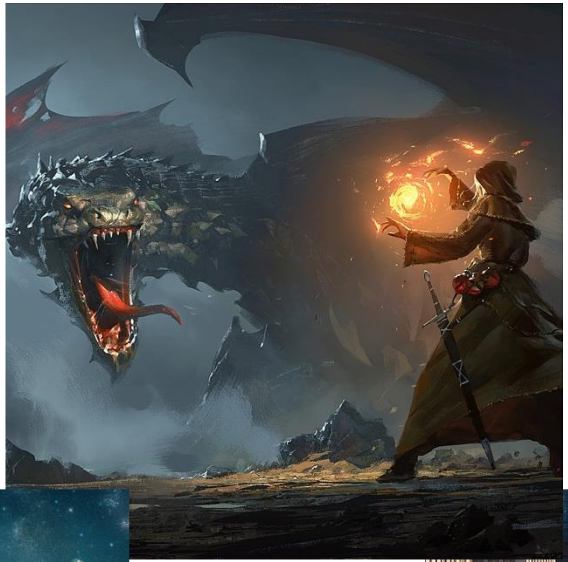
ML in PL CONFERENCE 2023

CDI: Copyrighted Data Identification in Diffusion Models

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, Adam Dziedzic

Warsaw University of Technology
CVLab@PW
IDEAS NCBR
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Diffusion Models are amazing image generators

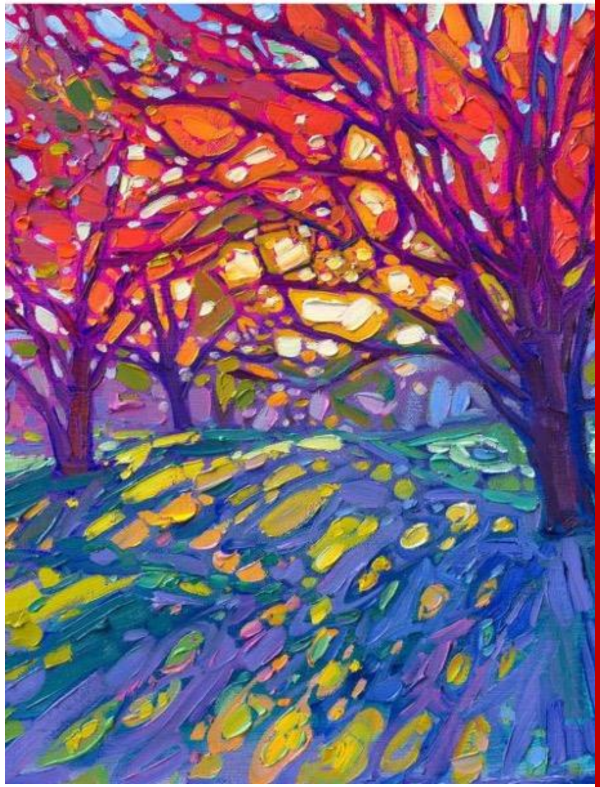


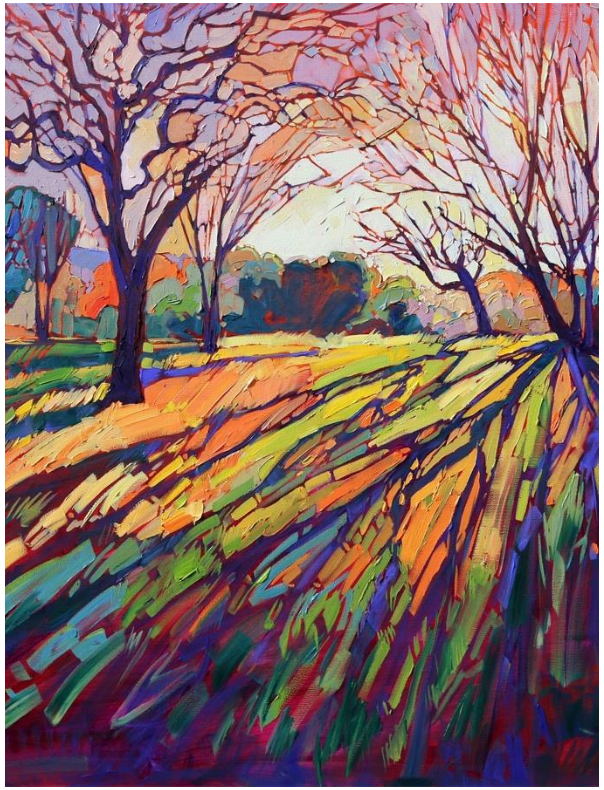
DMs are trained on billions of images

Which image is generated as

„a forest painting in the style of Erin Hanson"?

DMs are trained on billions of images



„a forest painting in the style of Erin Hanson” by Stable Diffusion

DMs are trained on billions of images

Real paintings by Erin Hanson

Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and ‘Collaging’ Their Work Without Consent

The plaintiffs claim the A.I. tools have unlawfully scraped and used their artwork in training datasets.

Artists and Illustrators Are Suing

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model 'without permission ... or compensation.'

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Feb 6, 2023, 5:56 PM GMT+1

16 Comments (16 New)

Artists and Illustrators Are Suing

ARTIFICIAL INTELLIGENCE / TECH /

Getty Images Stable Diffus infringement

An illustration from Getty Images' l
similar image (complete with Getty I
Diffusion. Image: Getty Images

Generative AI Lawsuits Timeline: Legal Cases vs. OpenAI, Microsoft, Anthropic, Nvidia and More

March 13, 2024 by Joe Panettieri

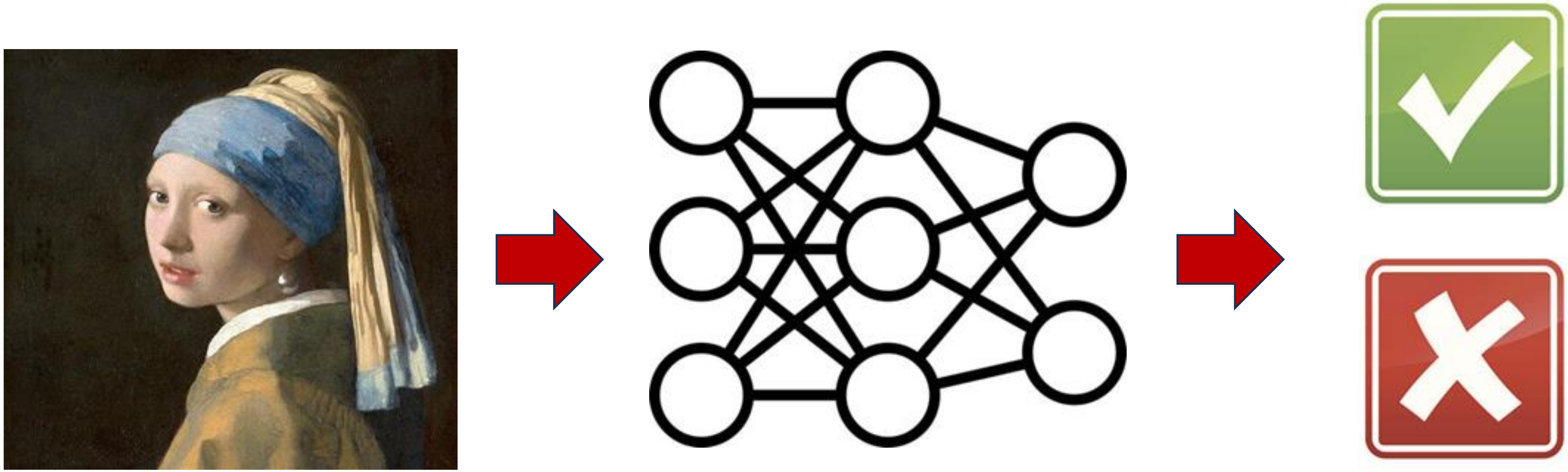
Was the sample used to train the Diffusion Model?

$\in$

Sample $\boldsymbol{s}$

Training Data $\boldsymbol{D}$

Membership Inference Attack



1. Choose sample ***s***

2. Query the model ***M*** with sample ***s***

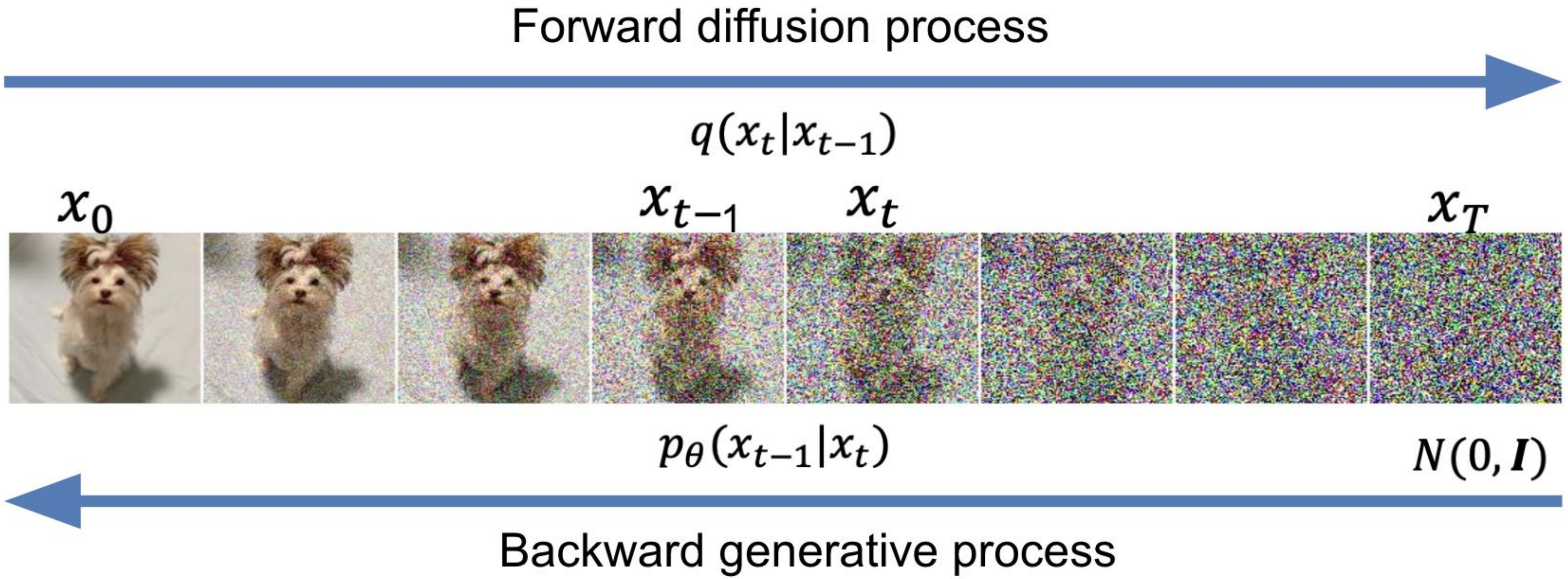
3. Decision: was ***s*** in the train set?

Loss threshold membership inference attack

Machine learning models minimize loss
on the training (*members*) set

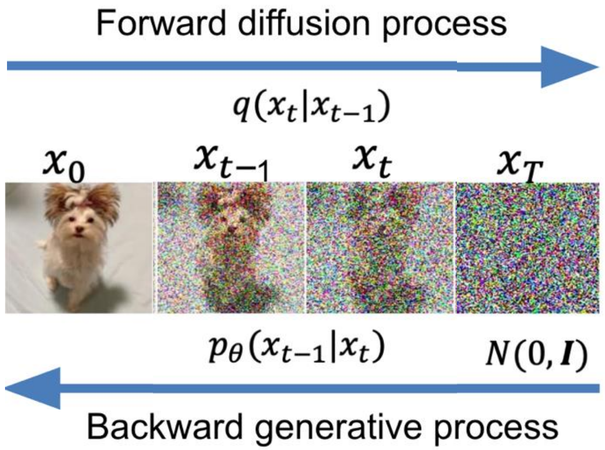
$$\begin{aligned} &IF\ Loss(sample) < Threshold \\ &\quad THEN\ Member \\ &\quad ELSE\ Nonmember \end{aligned}$$

Diffusion Models



Membership Inference Attack on Diffusion Models

1. Choose sample ***s***



2. Query the model ***M*** with sample ***s***

3. Decision: was ***s*** in the train set?

$$\boxed{IF\ loss(sample) < threshold\ THEN\ member\ ELSE\ nonmember}$$

Membership Inference for Large Diffusion Models Fails

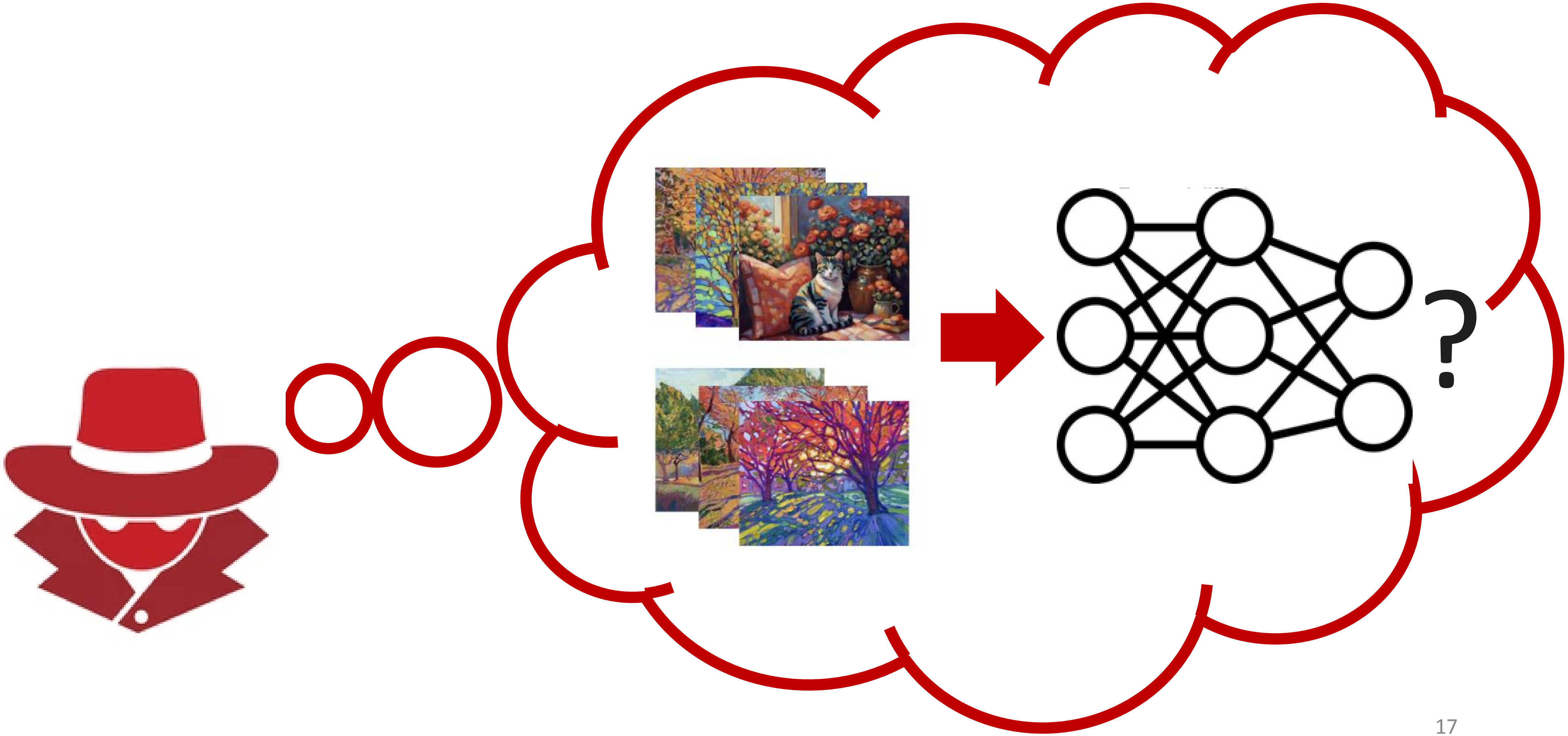
MIA	Max. TPR@FPR=1%
Denoising Loss	2.24%
Secmi	2.44%
PIA	5.57%
PIAN	1.53%

Membership Inference for Large Diffusion Models Fails

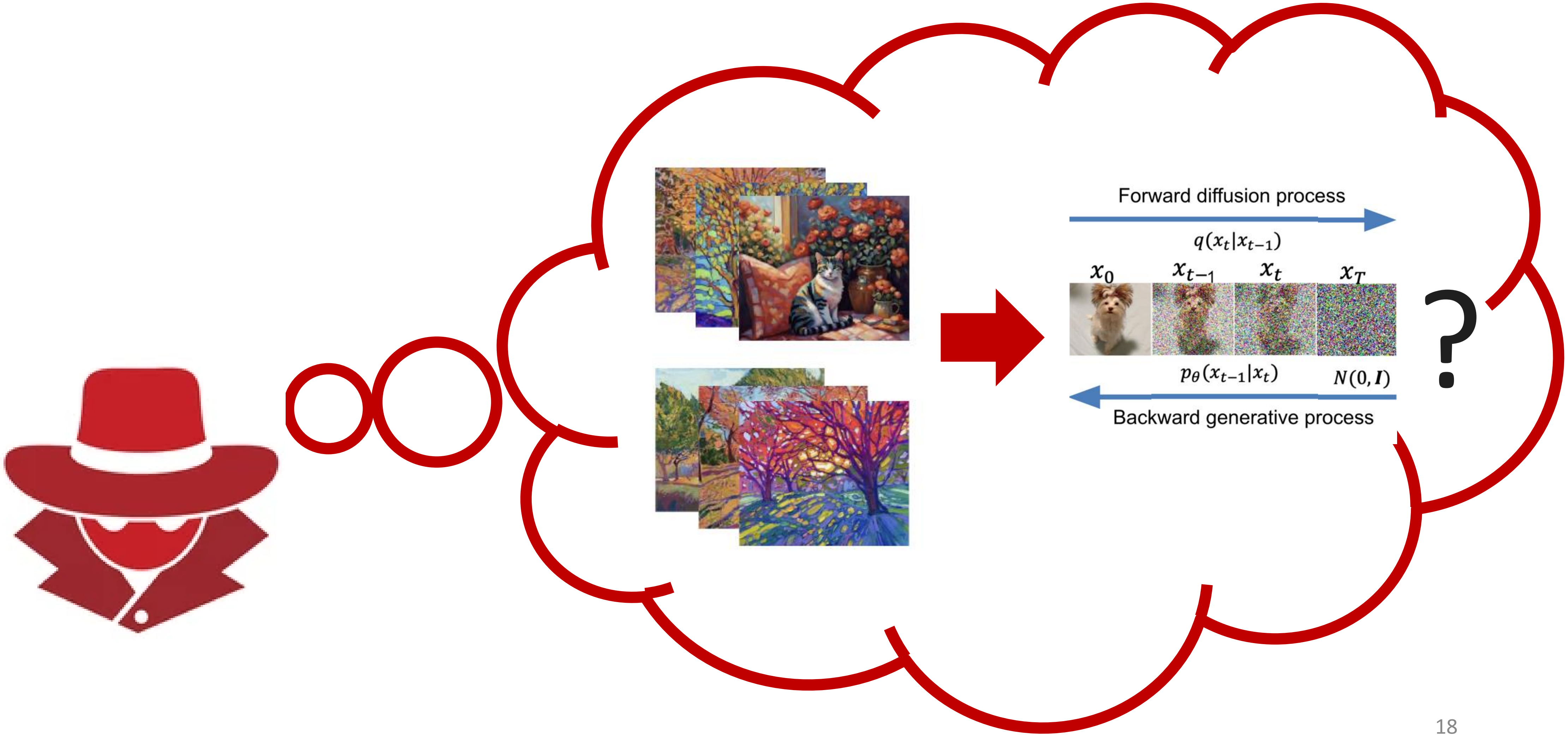
MIA	Max. TPR@FPR=1%
Denoising Loss	2.24%
Secmi	2.44%
PIA	5.57%
PIAN	1.53%

How can we do better?

Dataset Inference



Dataset Inference for Diffusion Models



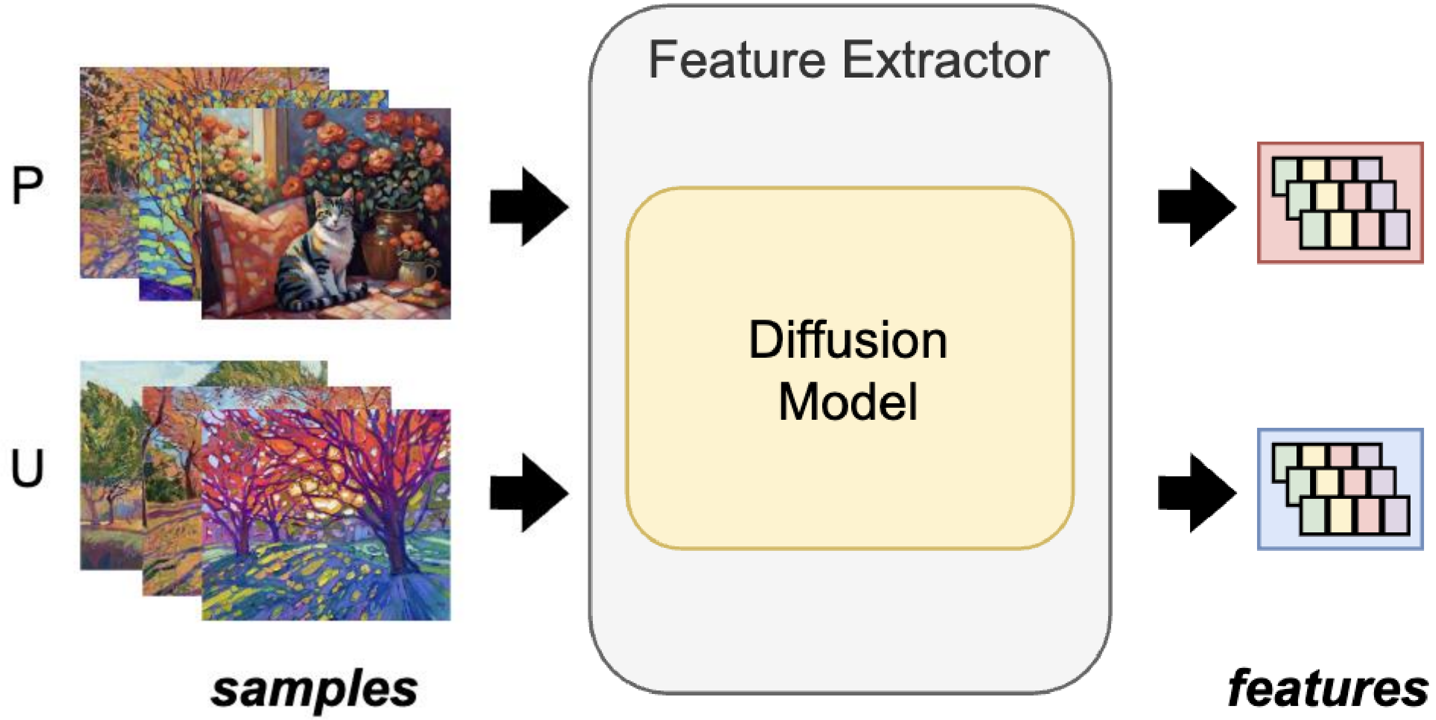
Step 1: Prepare Data and Model

P - published pictures – used for training?

U - unpublished pictures

samples

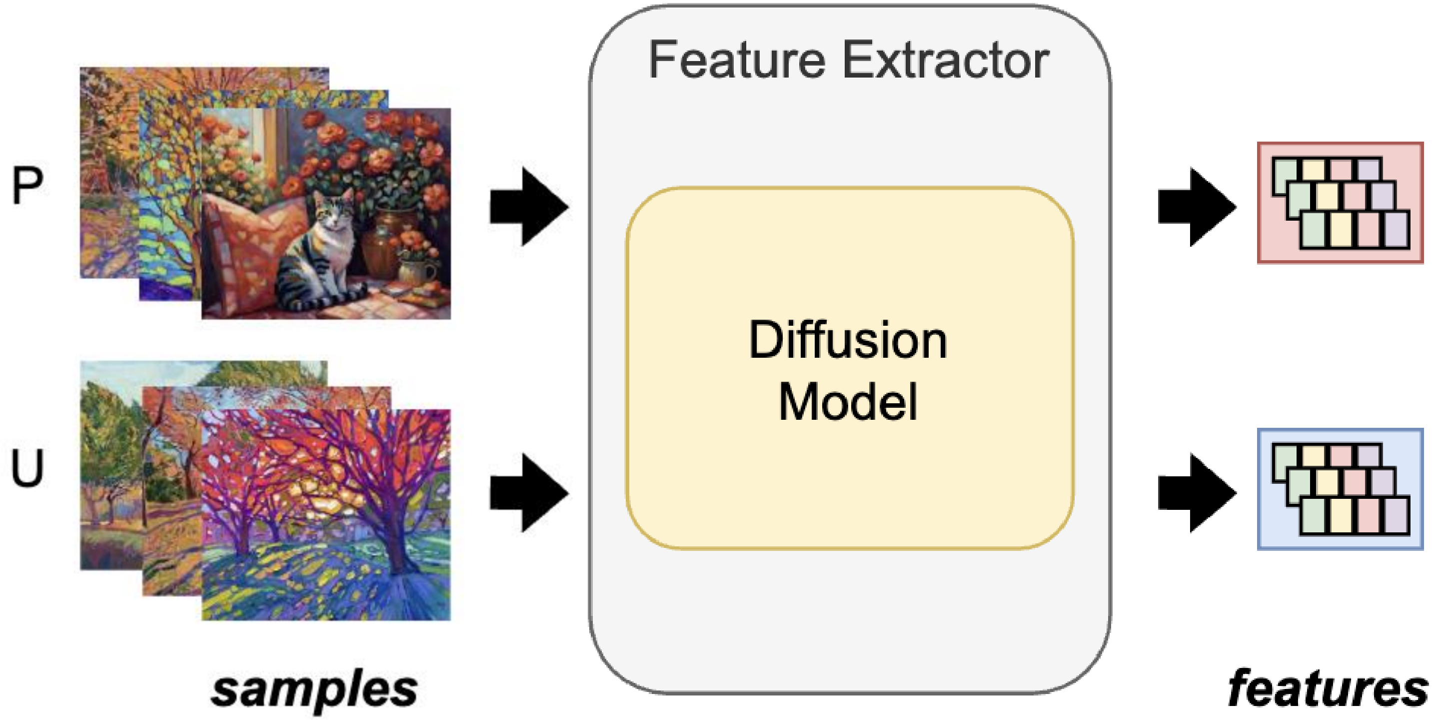
Step 2: Feature Extraction



Existing MIA Features:

1. Denoising Loss

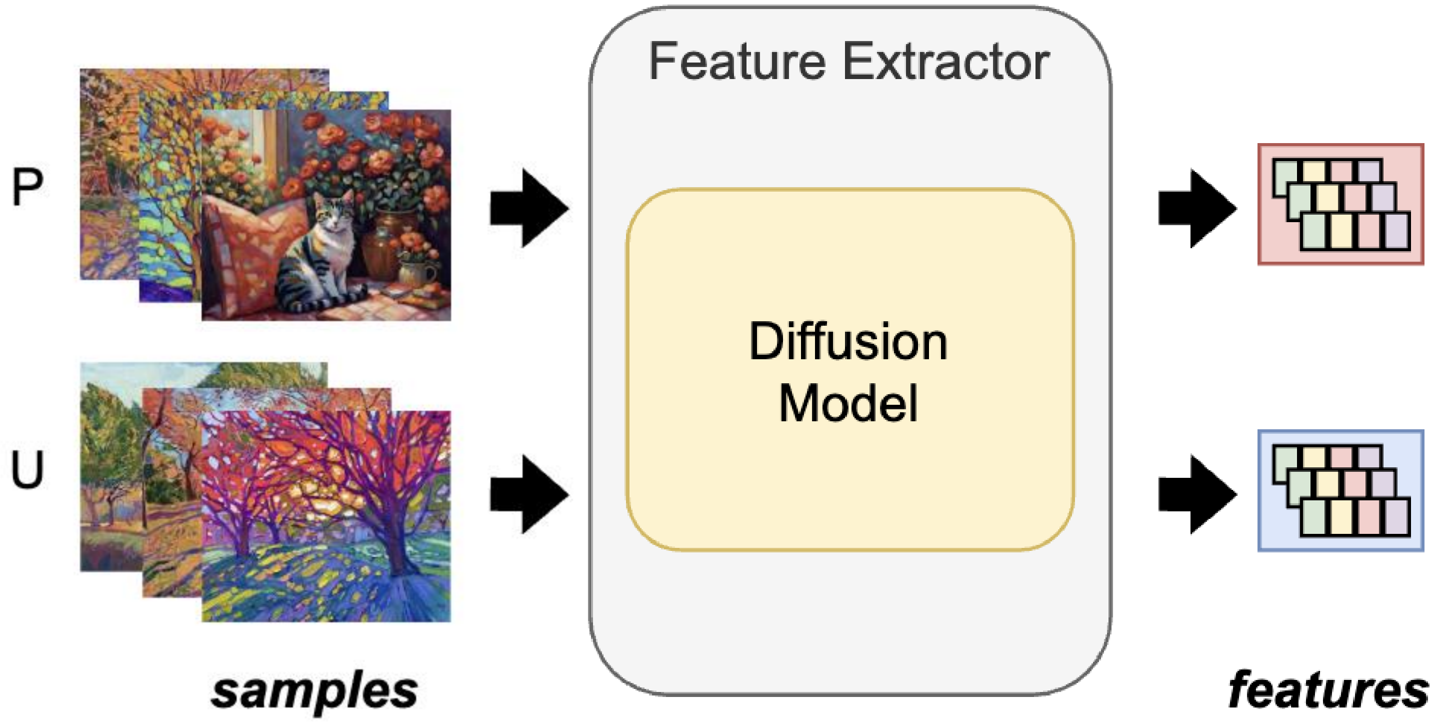
Step 2: Feature Extraction



Existing MIA Features:

1. Denoising Loss
2. SecMI

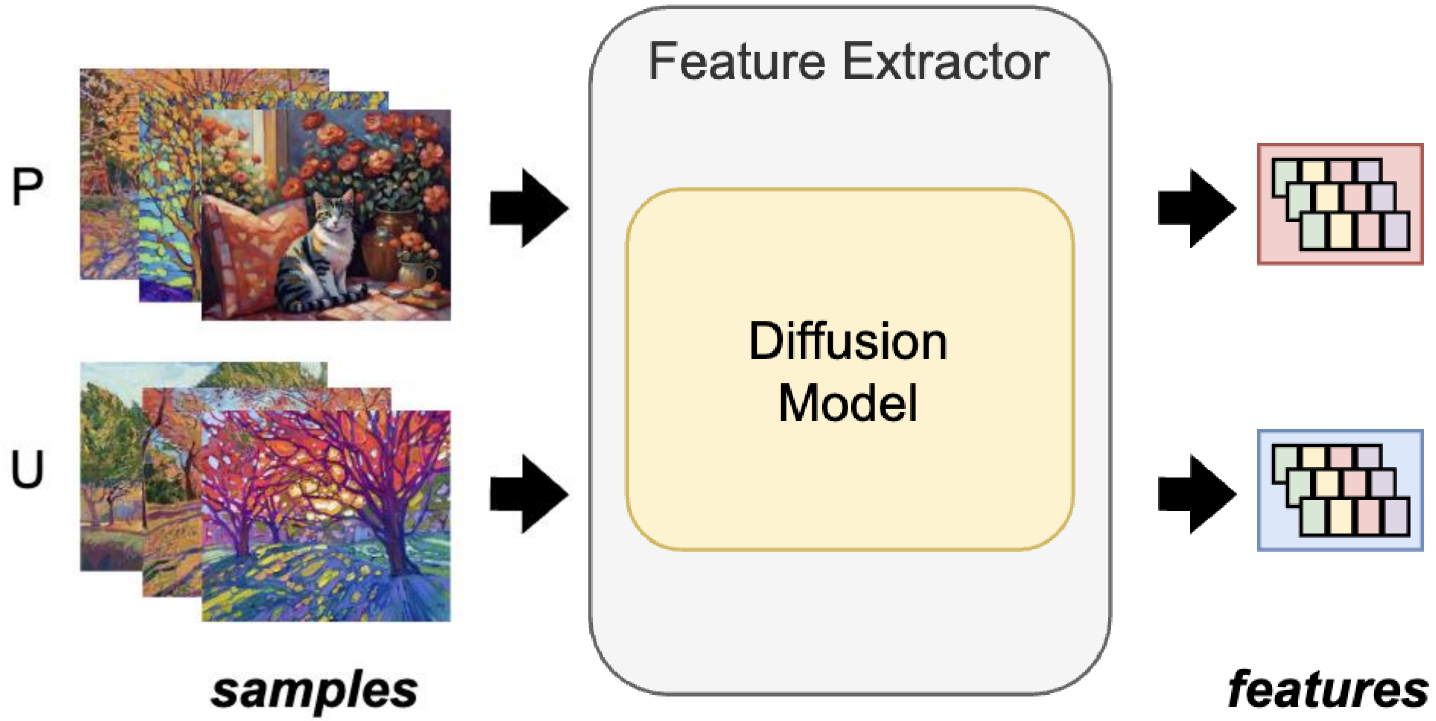
Step 2: Feature Extraction



Existing MIA Features:

1. Denoising Loss
2. SecMI
3. PIA/PIAN

Step 2: Feature Extraction



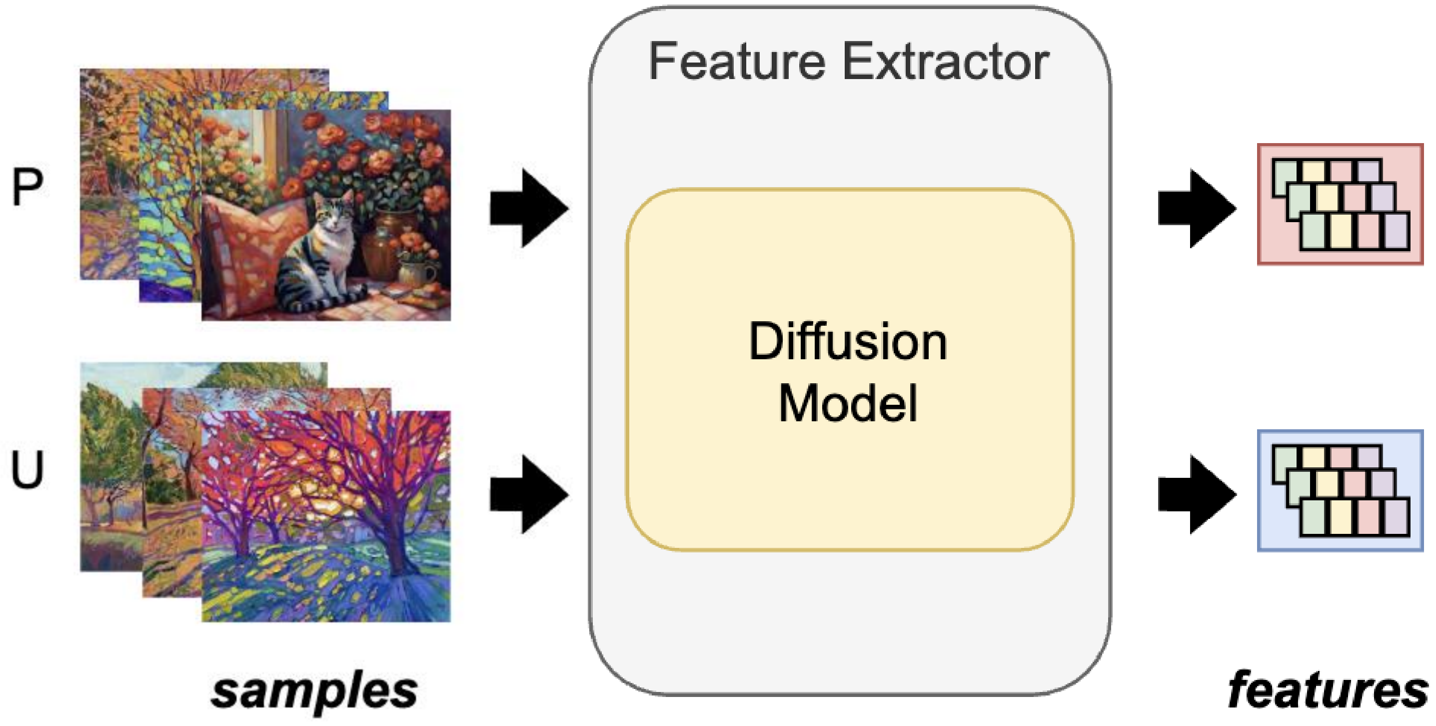
Existing MIA Features:

1. Denoising Loss
2. SecMI
3. PIA/PIAN

Our new features

1. Multiple Loss

Step 2: Feature Extraction



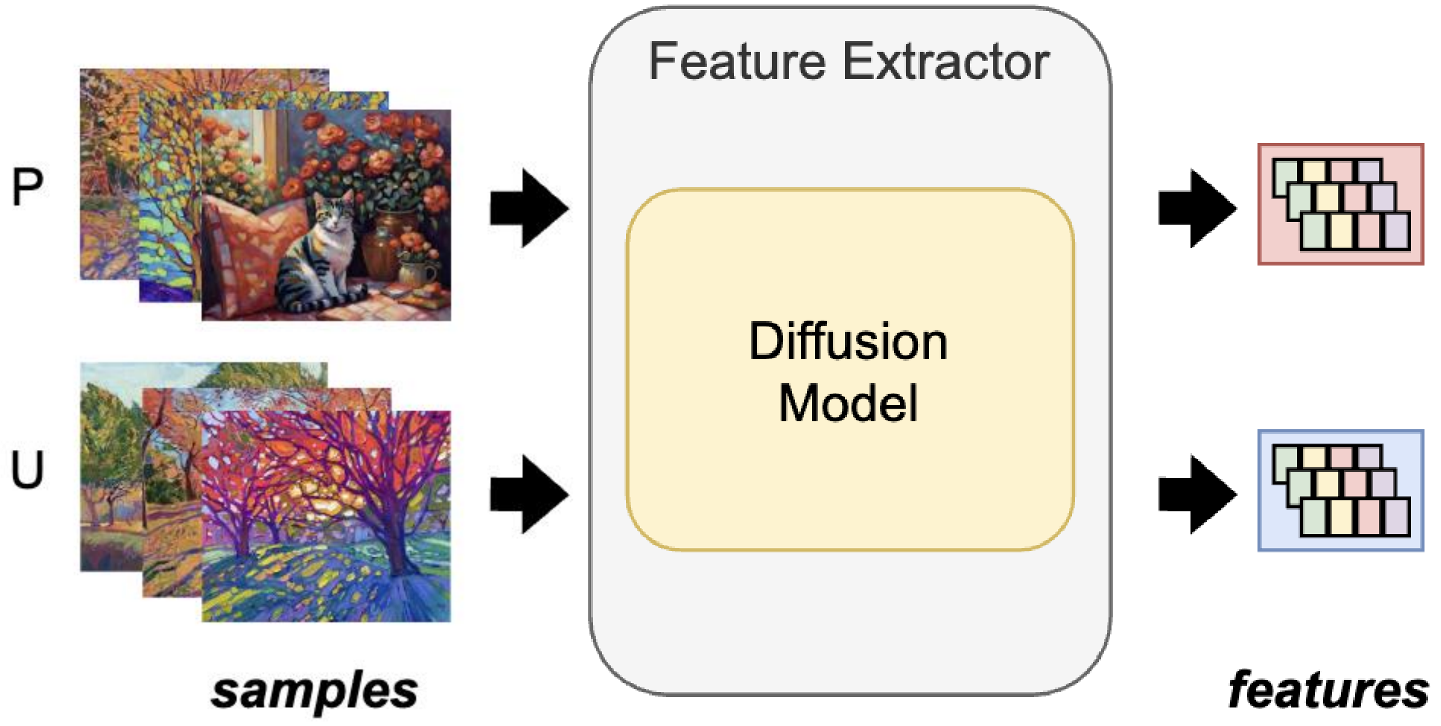
Existing MIA Features:

1. Denoising Loss
2. SecMI
3. PIA/PIAN

Our new features

1. Multiple Loss
2. Noise Optimisation

Step 2: Feature Extraction



Existing MIA Features:

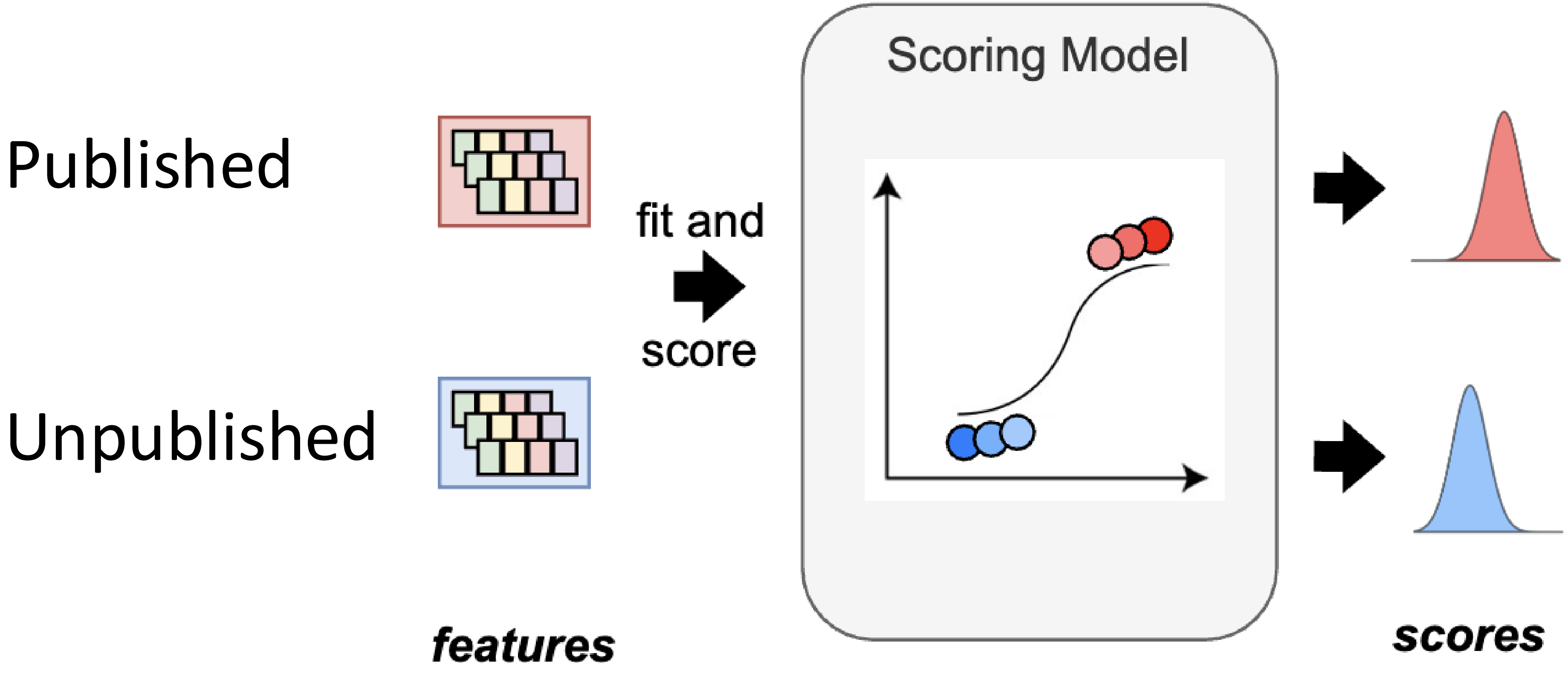
1. Denoising Loss
2. SecMI
3. PIA/PIAN

Our new features

1. Multiple Loss
2. Noise Optimisation
3. Gradient Masking

Step 3: Scoring model

$$s: Rk \to [0,1]$$

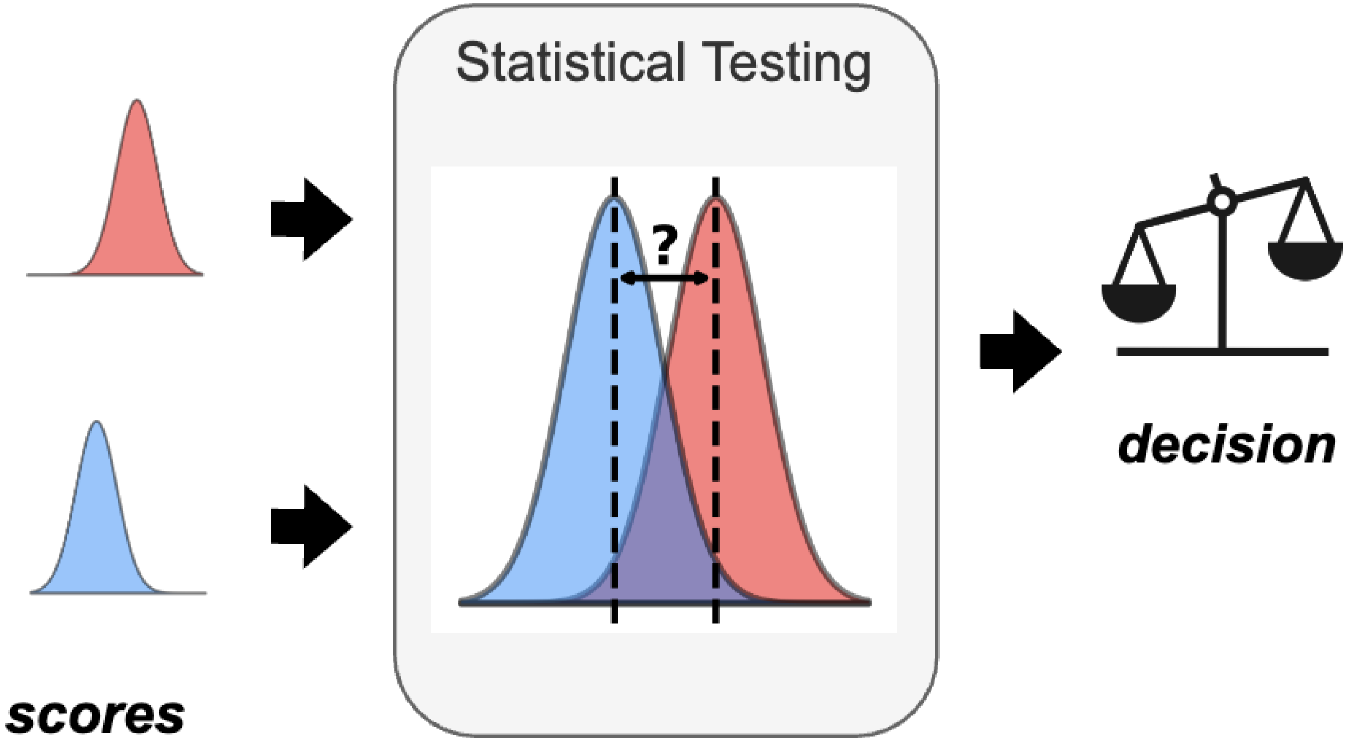


Step 4: Statistical testing

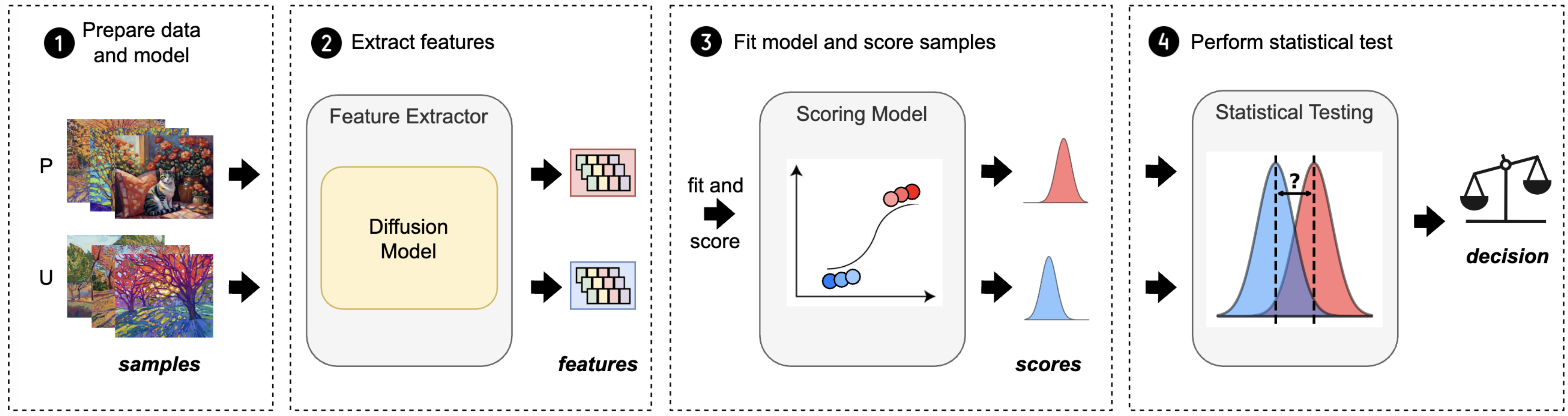
$$H_0\text{: } scores(Published) \; <= \; scores(Unpublished)$$

Published

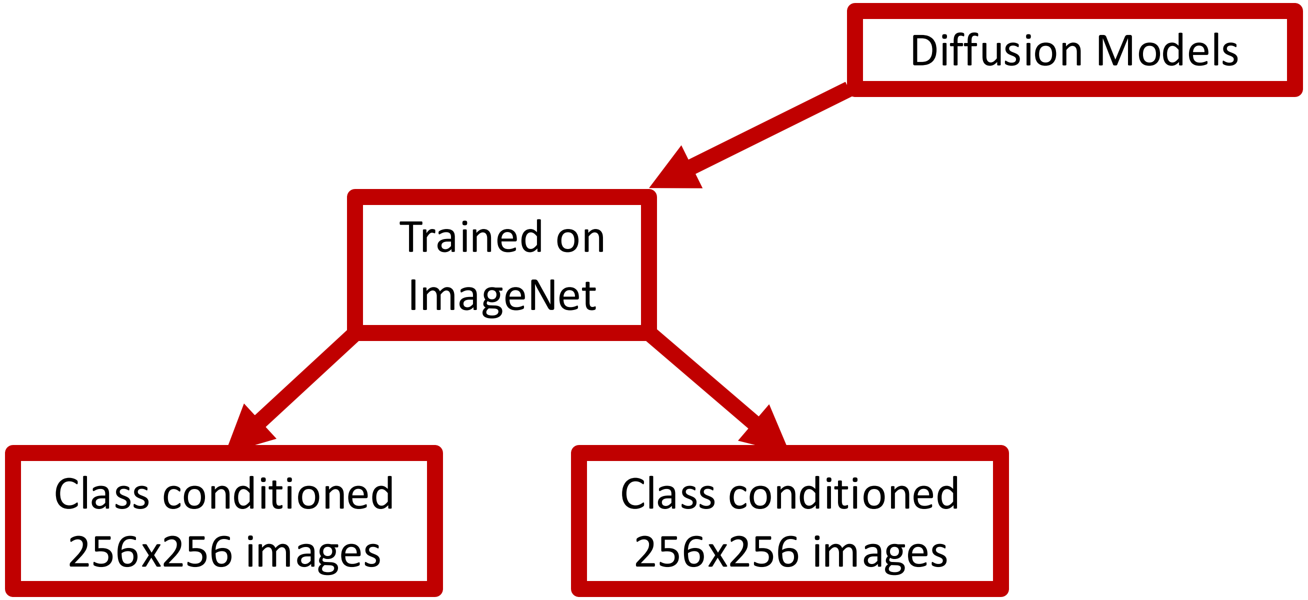
Unpublished



End-to-End solution



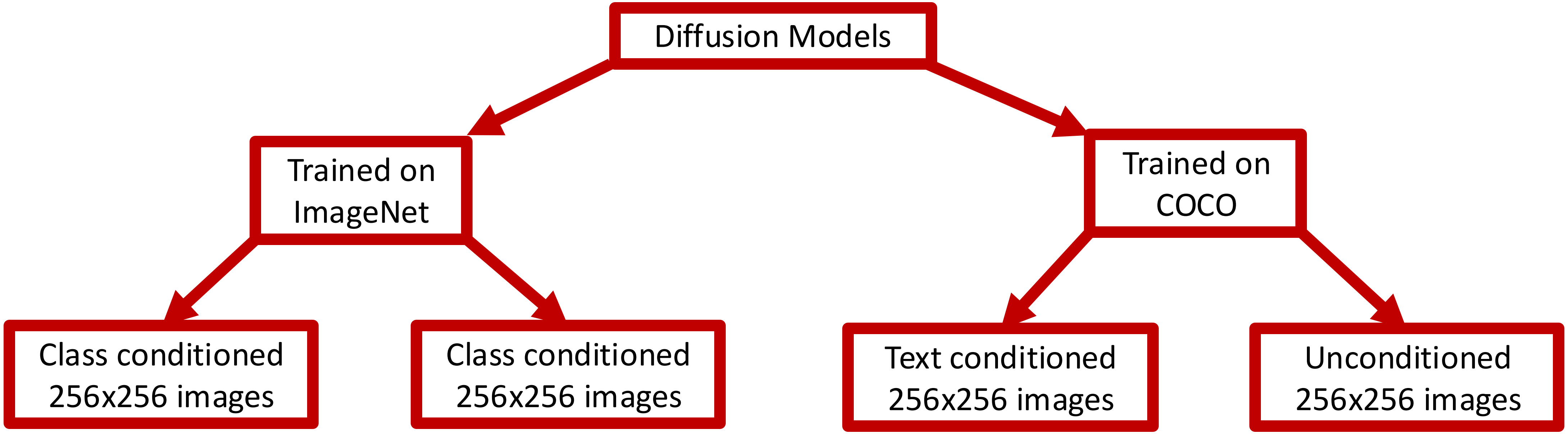
Experimental set-up



LDM
UViT-256
DiT-256

UViT-512
DiT-512

Experimental set-up



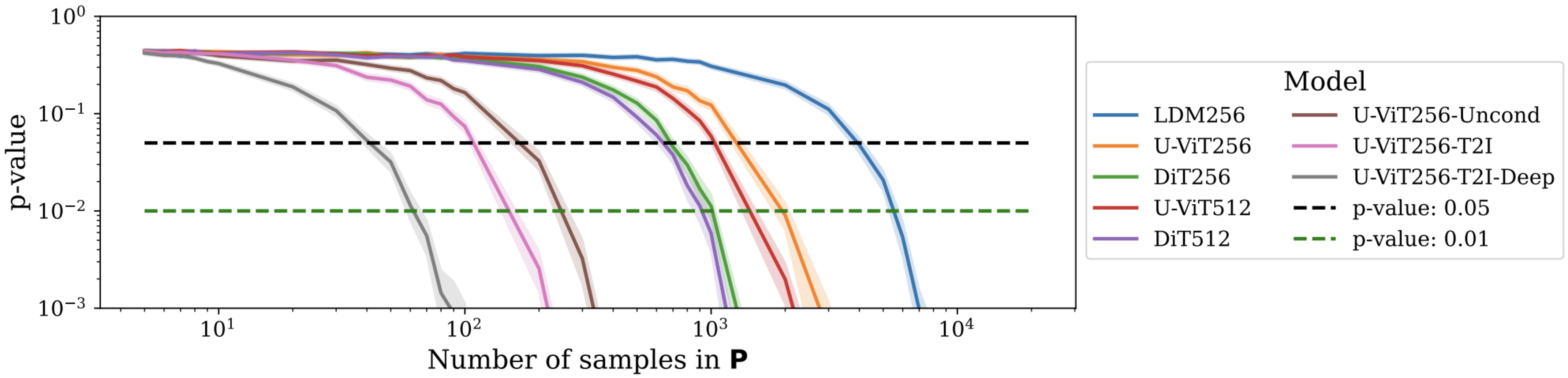
LDM
UViT-256
DiT-256

UViT-512
DiT-512

UViT-T2I
UViT-T2I-Deep

UViT-Uncond

CDI works



In some case we need <100 samples

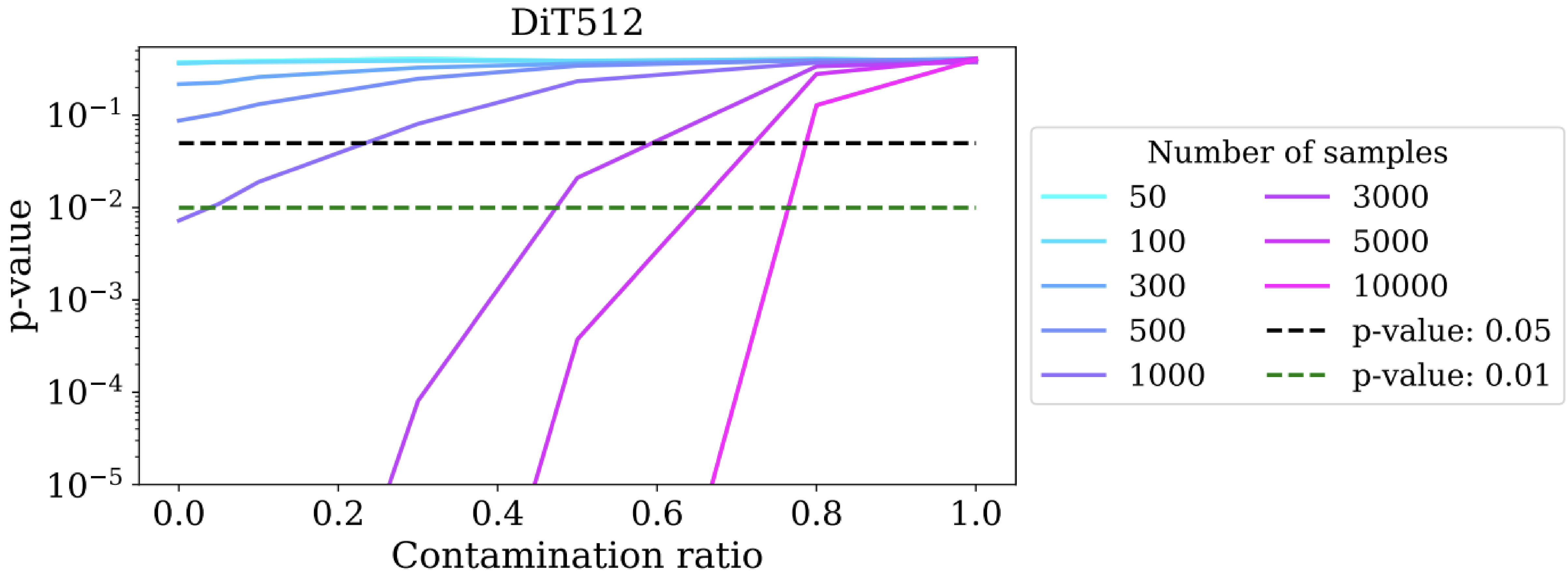
Our new features lower the number of samples

LDM256	13000	6000
U-ViT256	3000	3000
DiT256	2000	2000
U-ViT512	12000	2000
DiT512	8000	1000
U-ViT256-Uncond	400	300
U-ViT256-T2I	200	200
U-ViT256-T2I-Deep	70	70

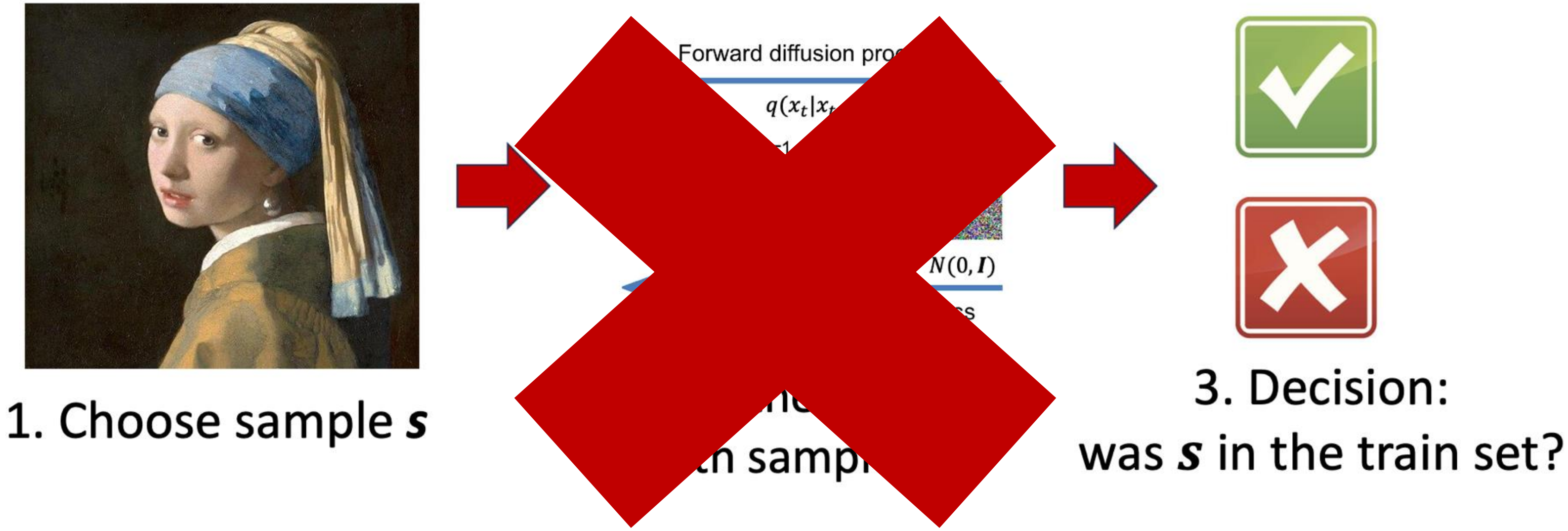
No False Positives

	LDM	DiT-256	UViT-512	UViT-T2i
Members	10^{-6}	10^{-59}	10^{-31}	~0.0
Nonmembers	0.4	0.39	0.39	0.39

Works if only part of data was used in training

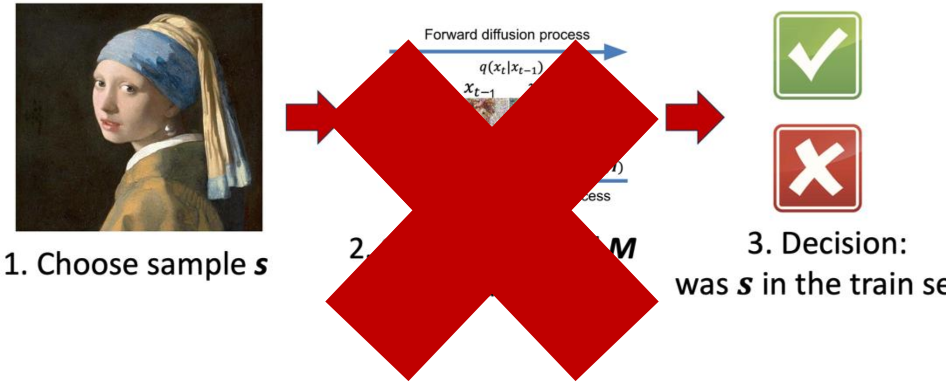


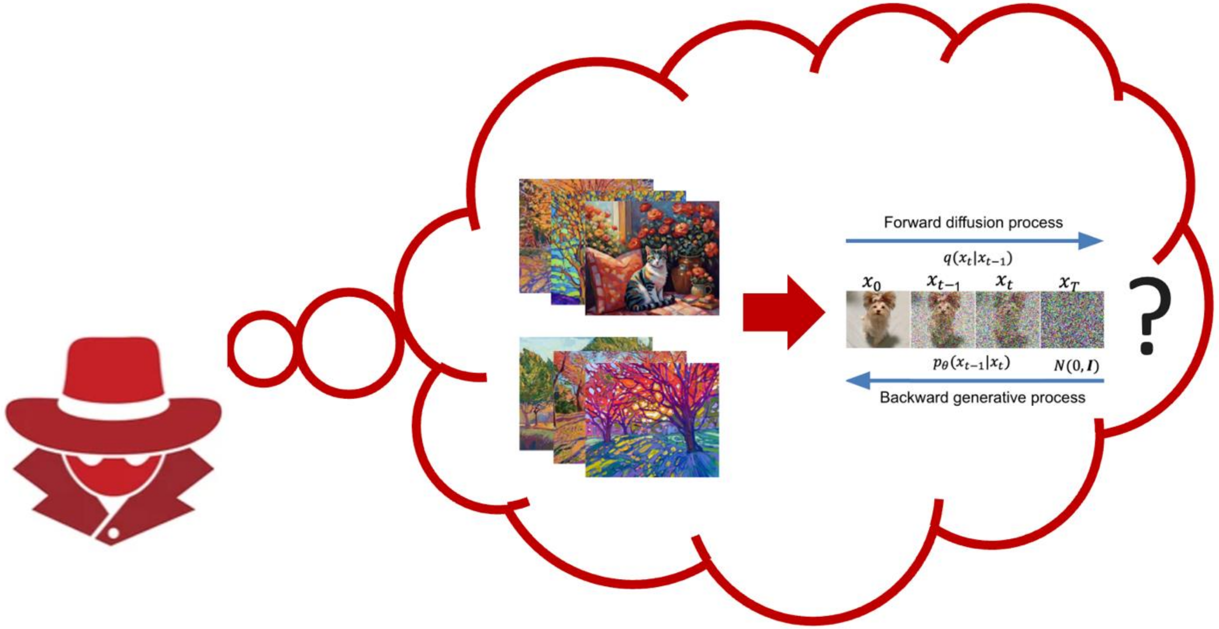
Key findings



State of the art membership inference methods fail on large diffusion models!

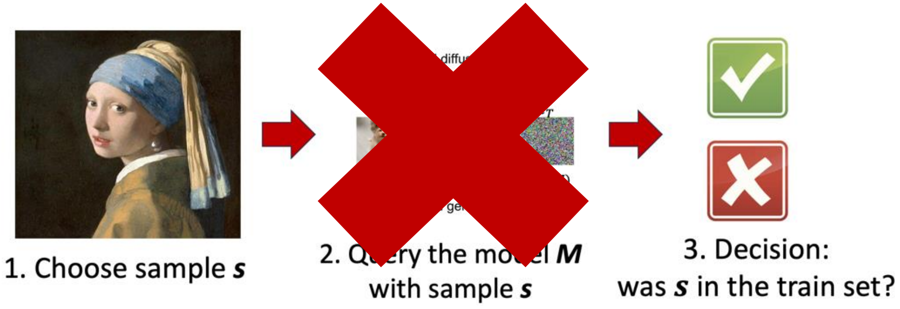
Key findings

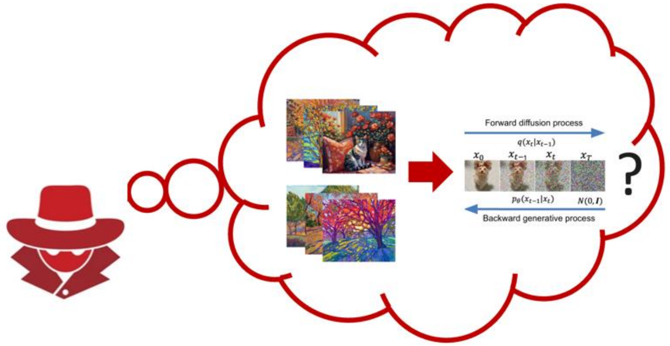


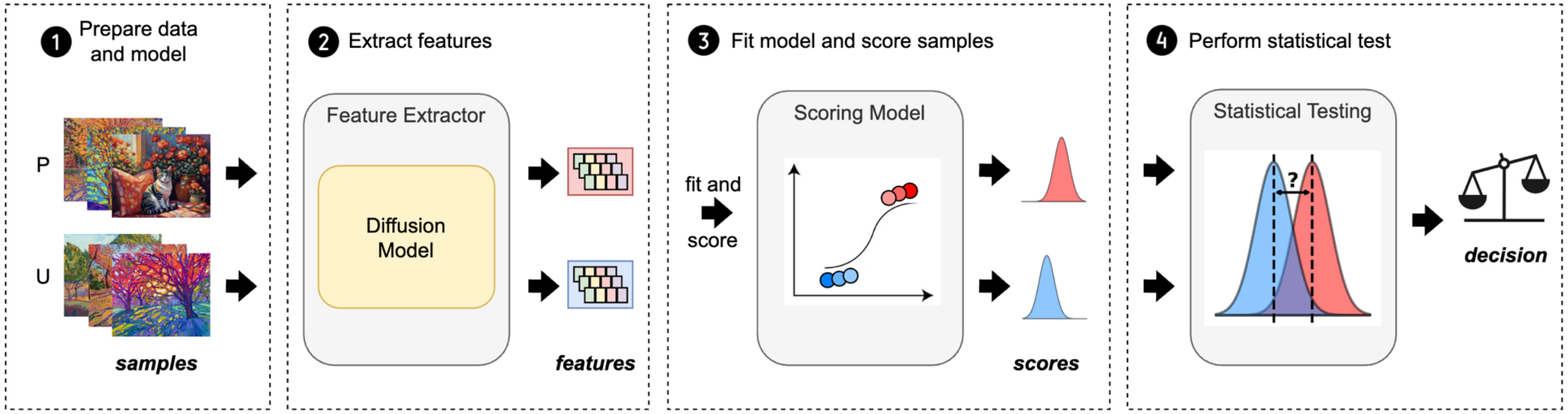


We shift to Dataset Inference to protect the Intellectual Property in data collections

Key findings







CDI successfully identifies data collections used in training of large diffusion models