Privacy Attacks on Image AutoRegressive Models

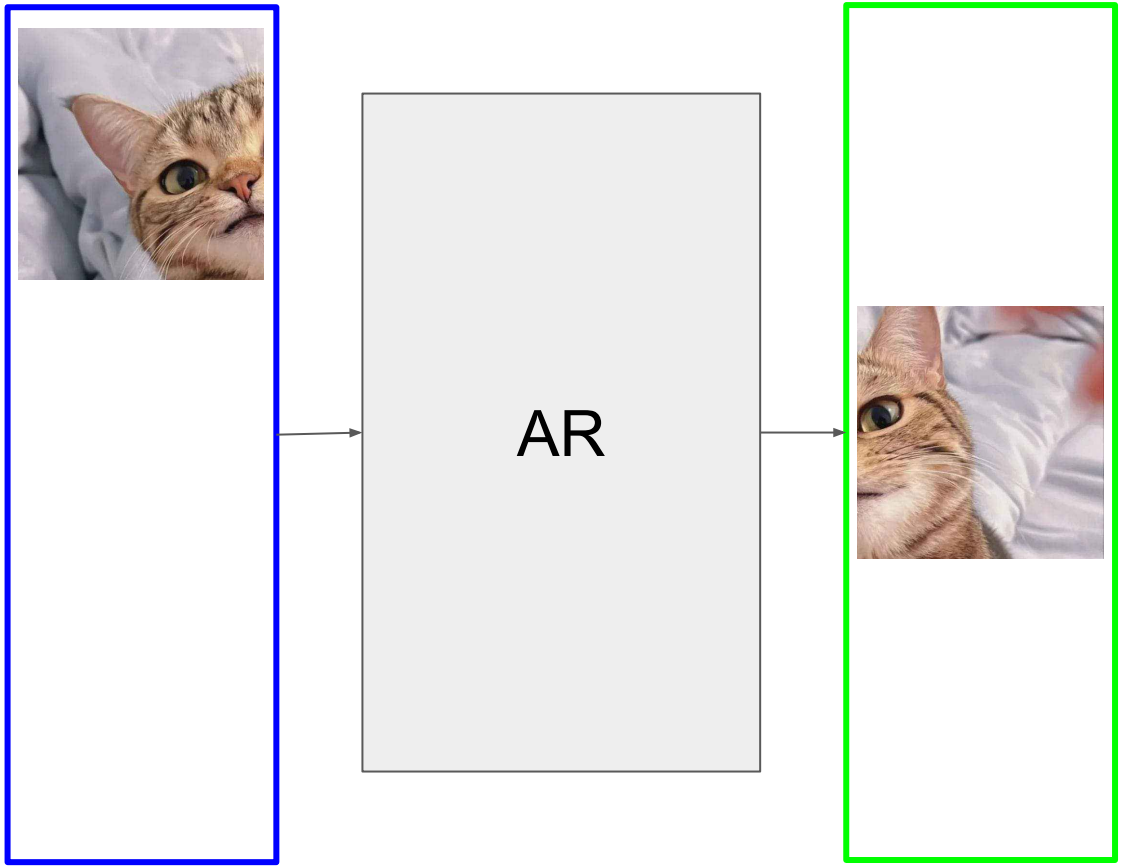
Antoni Kowalczuk*, Jan Dubiński*,
Franziska Boenisch, Adam Dziedzic

CISPA SprintML

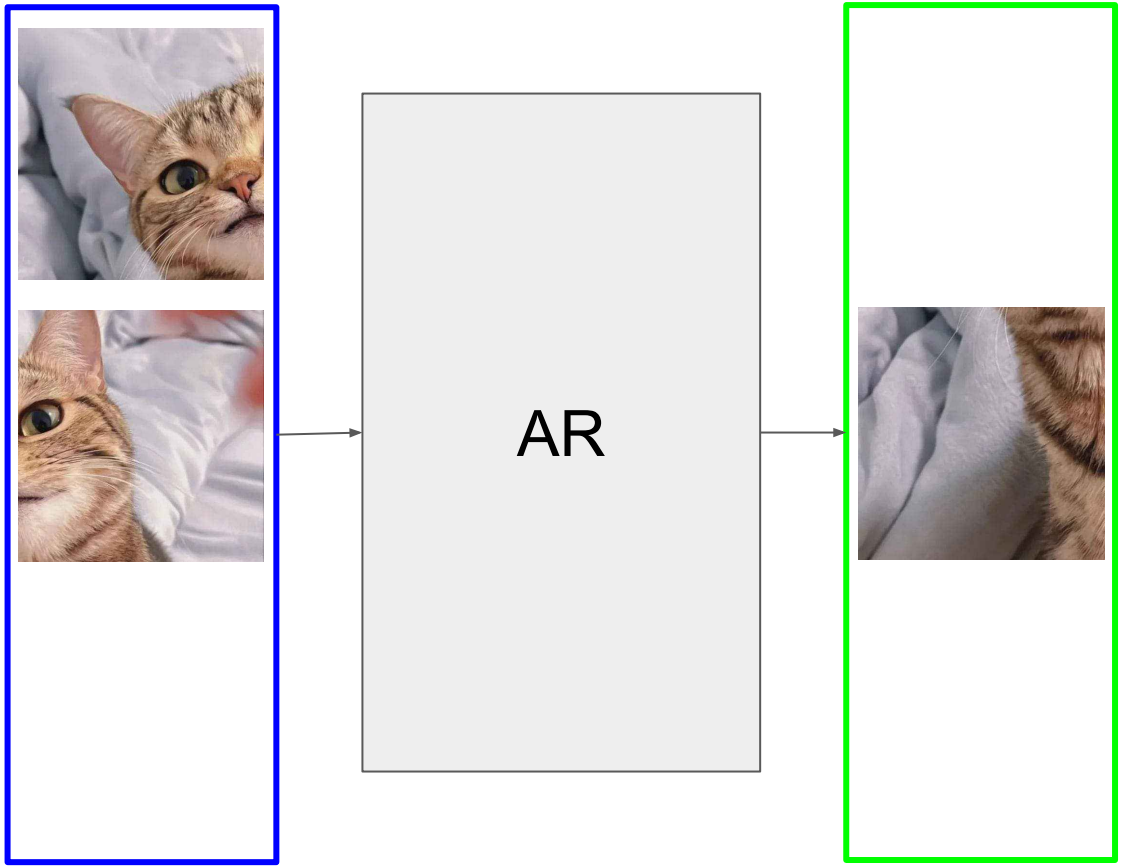
Warsaw University of Technology

NASK

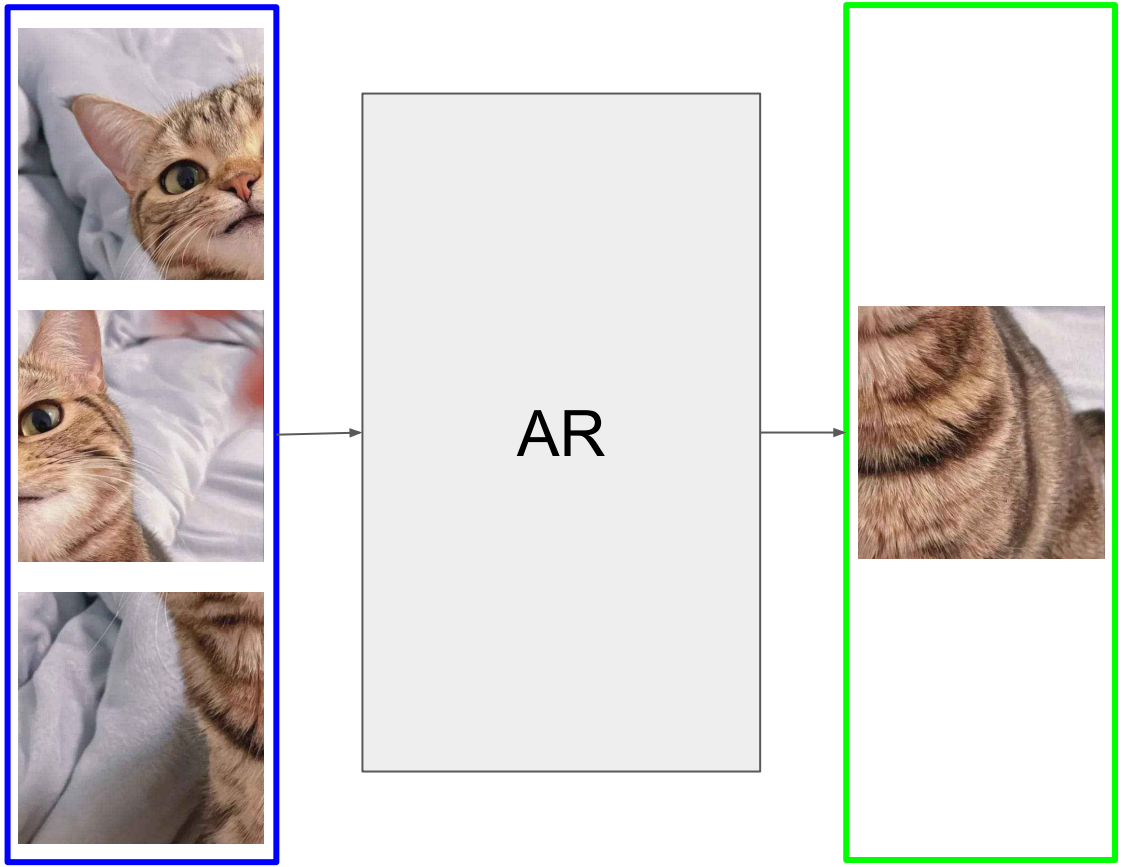
Image AutoRegressive Models (IARs)



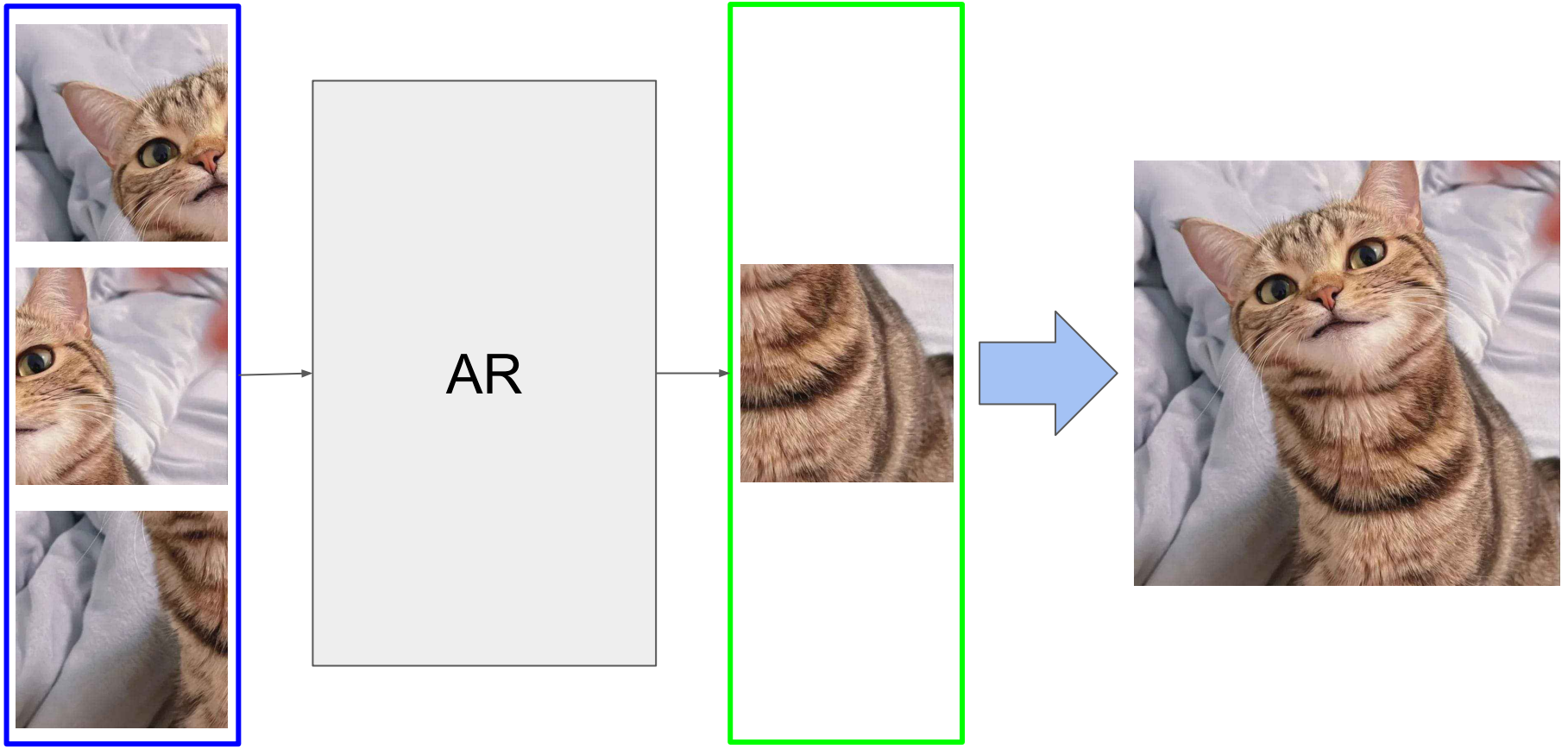
IARs



IARs

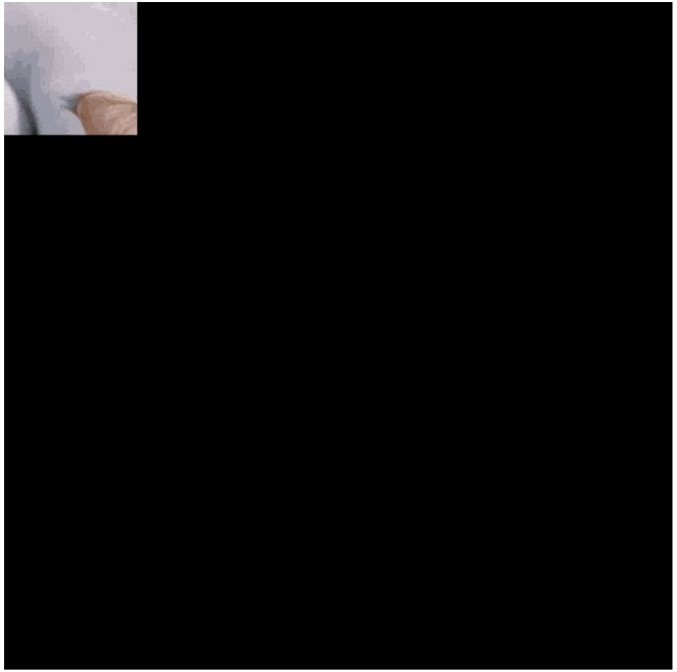


IARs

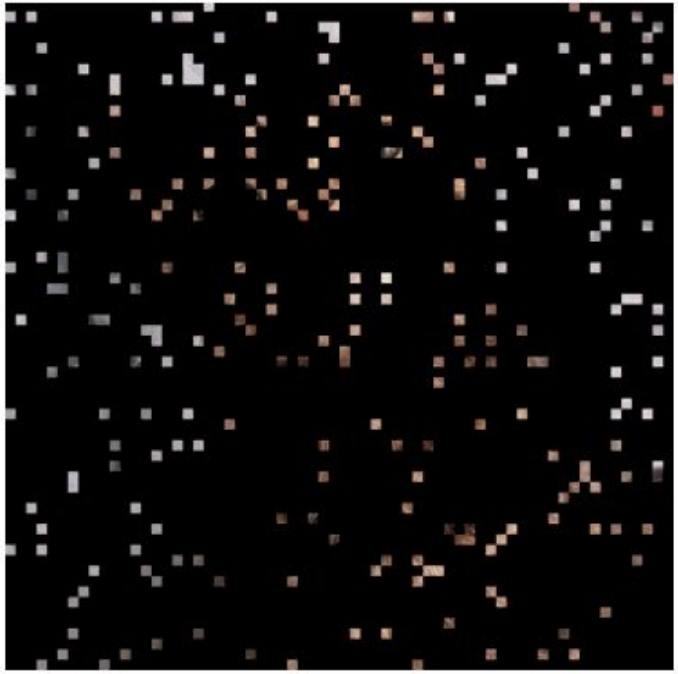


VAR (Tian et al., 2024)

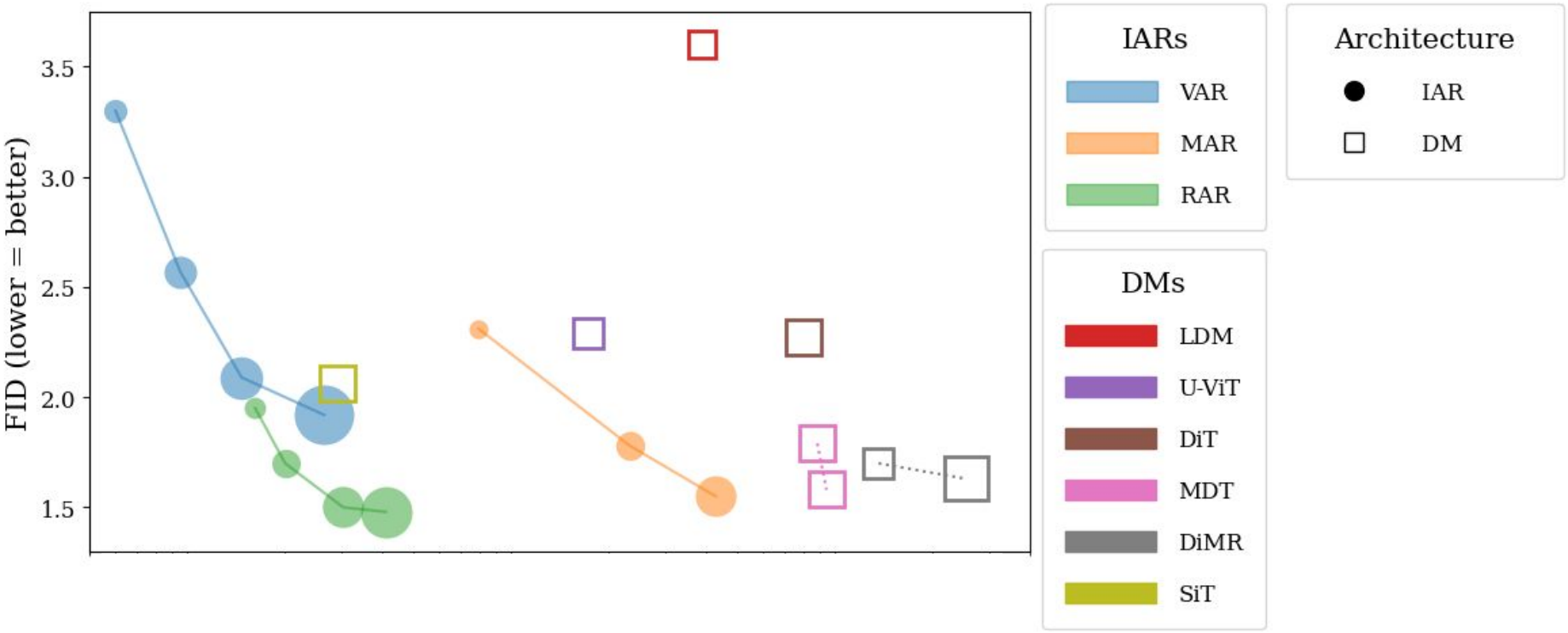
RAR (Yu et al., 2024)



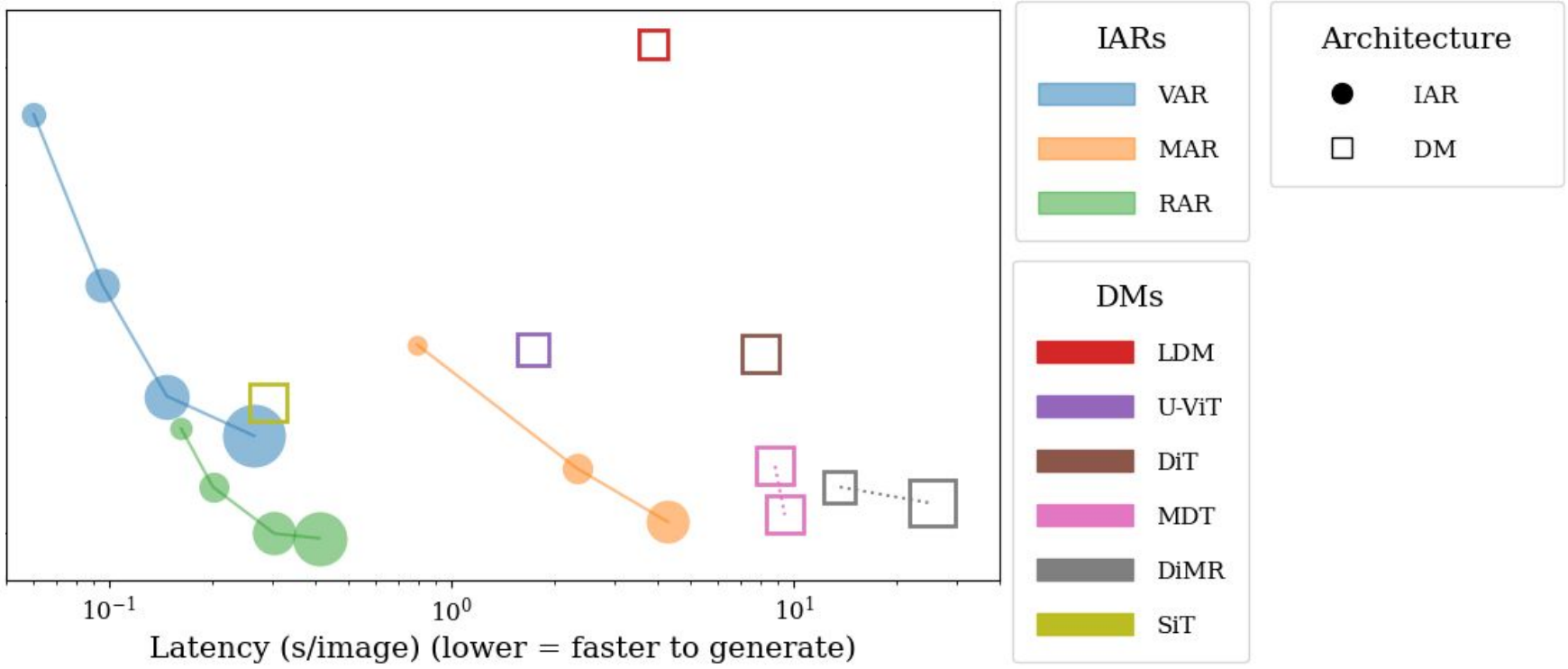
MAR (Li et al., 2024)



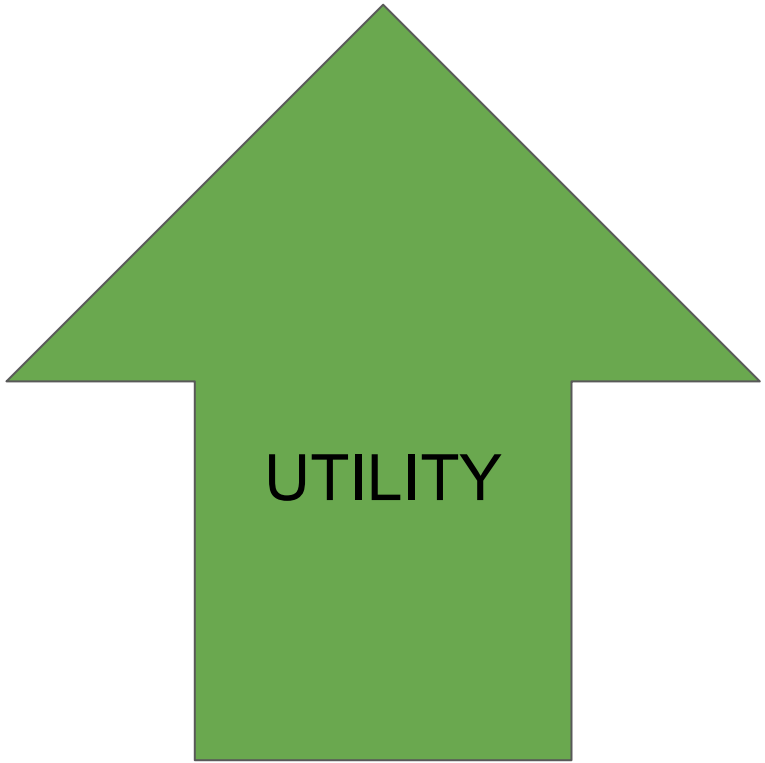
IARs generate **better** images than DMs



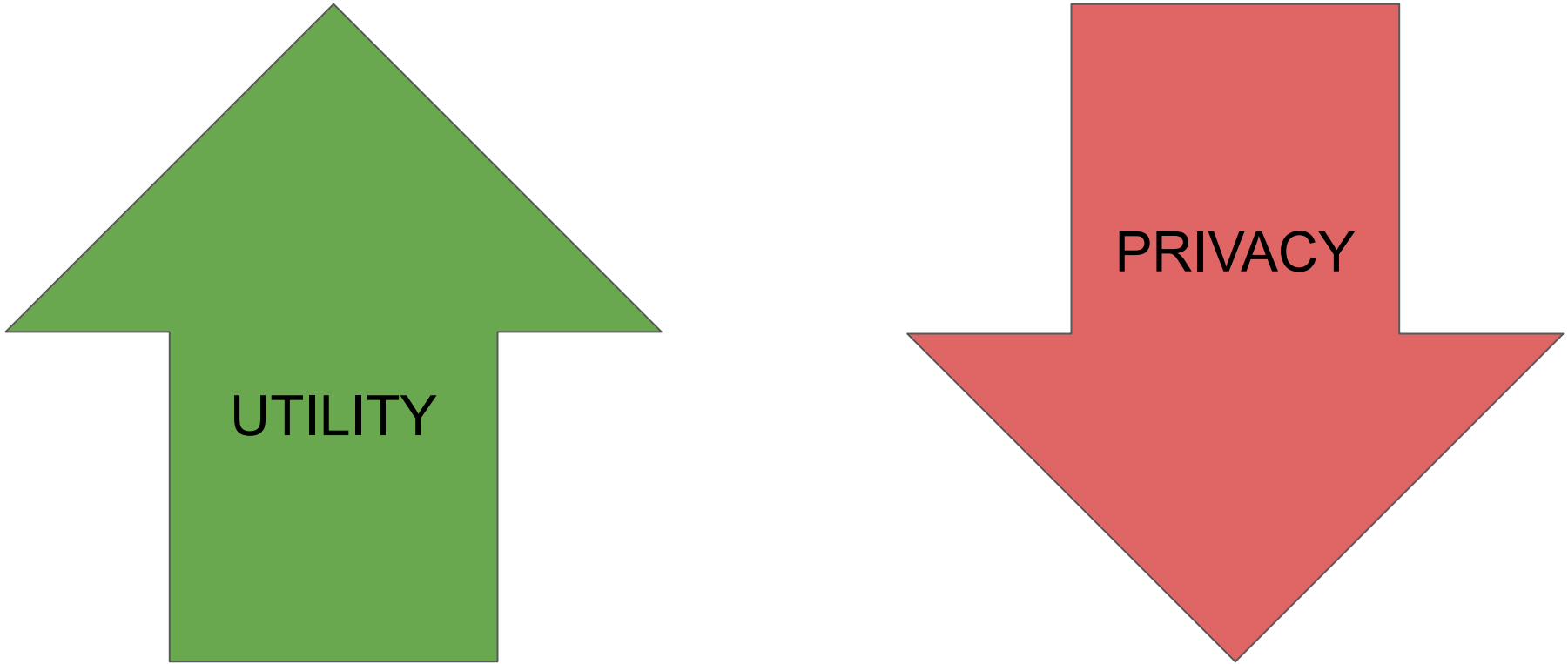
IARs images **faster** than DMs



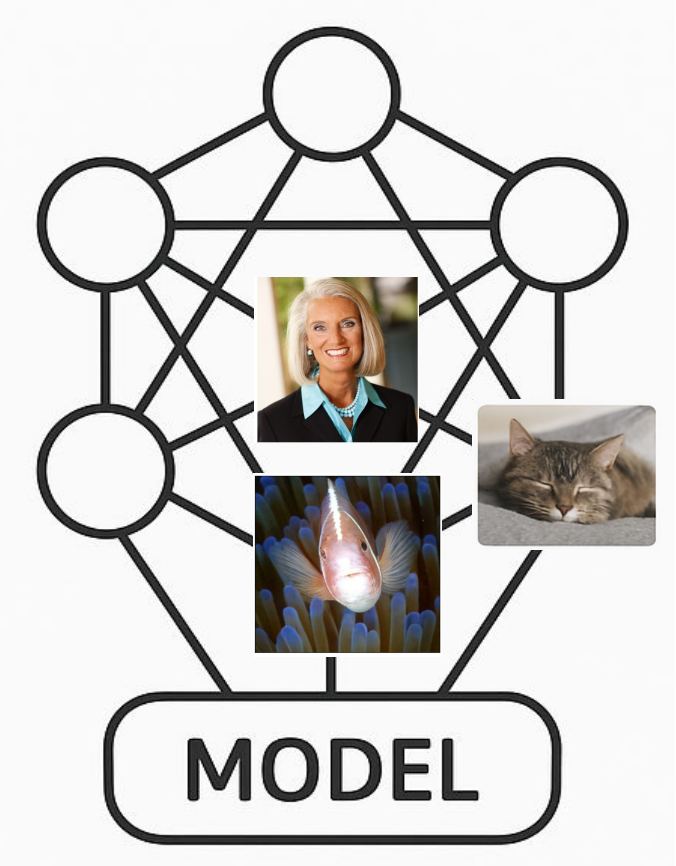
Privacy-utility trade-off



Privacy-utility trade-off



Privacy Leakage

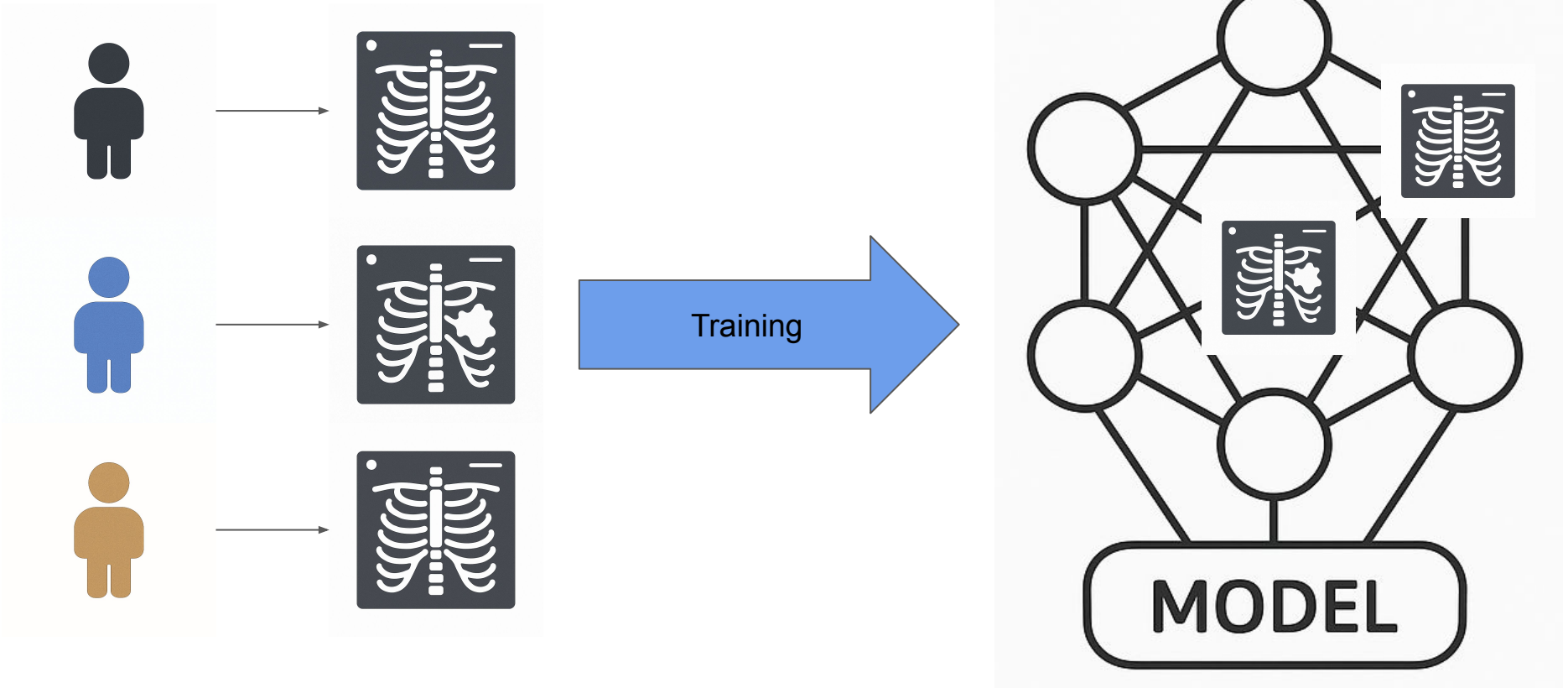


Leak Data?

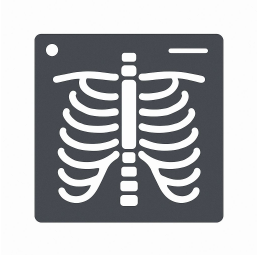
OK

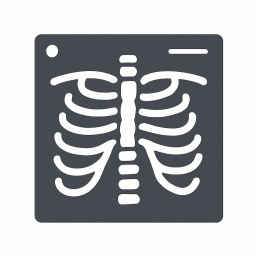
Sensitive Data

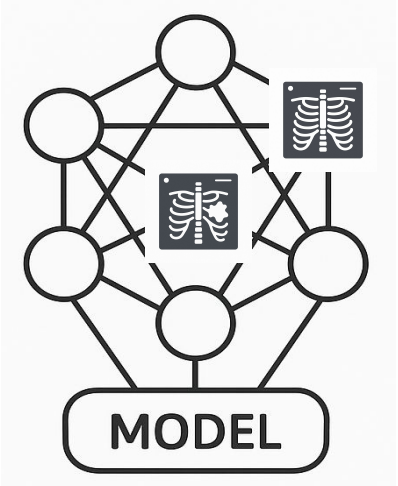
Privacy Leakage



Privacy Leakage

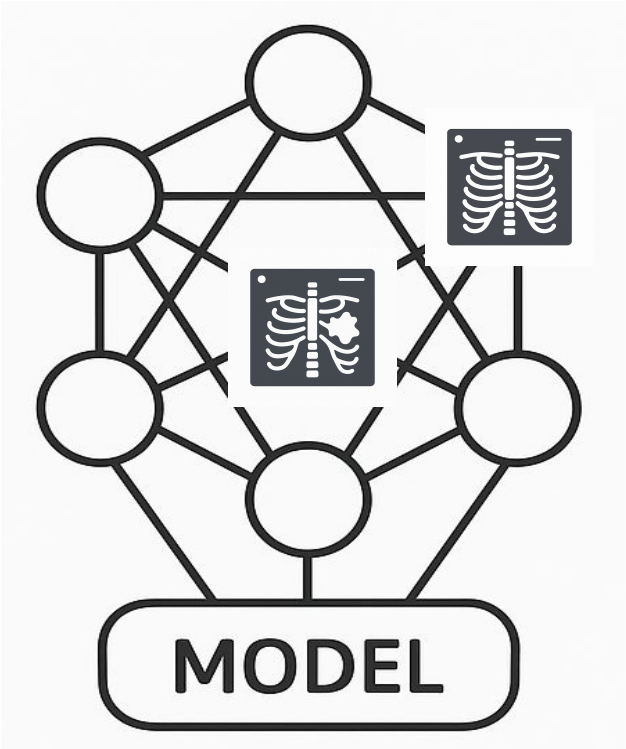






Privacy Leakage

in

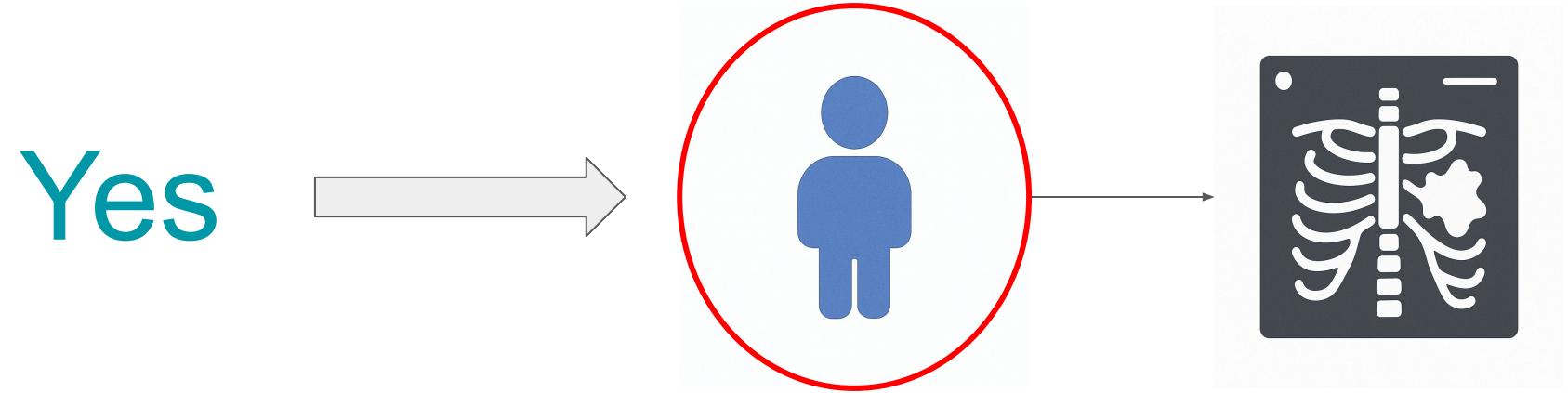


?

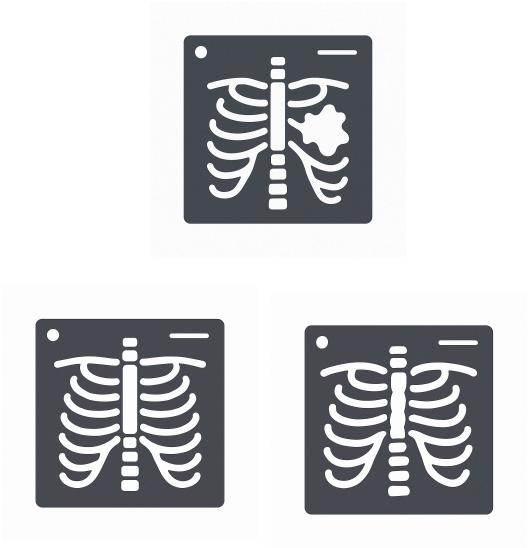
Membership Inference Attack (MIA)

Has lung cancer

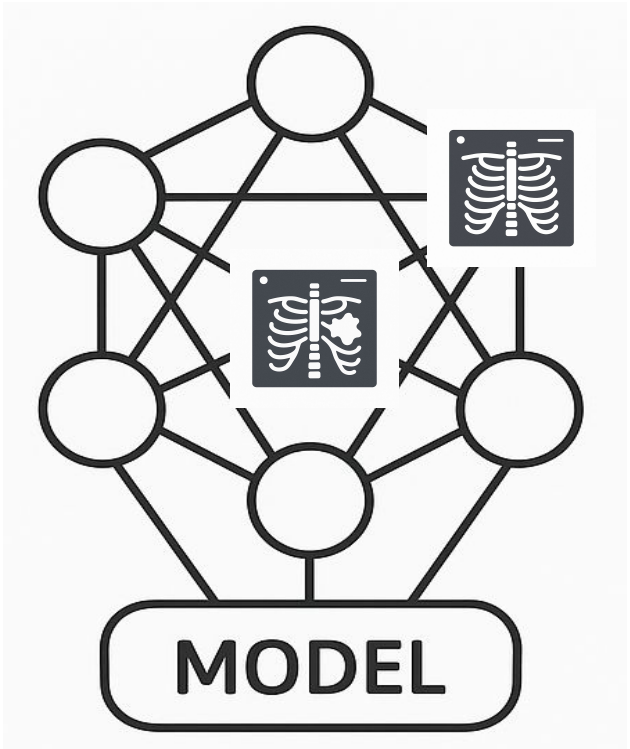
Yes



Dataset Inference (DI)

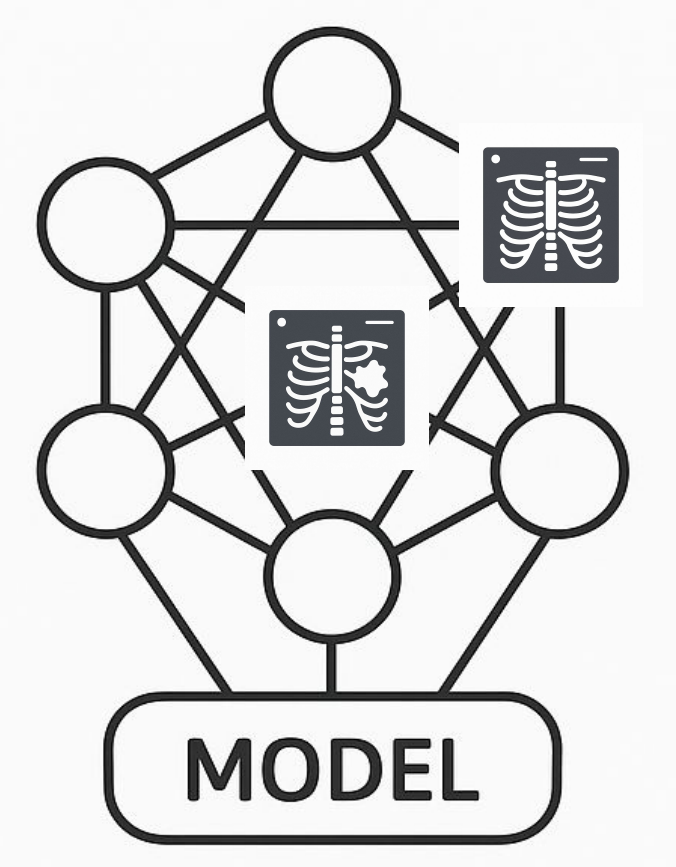


in



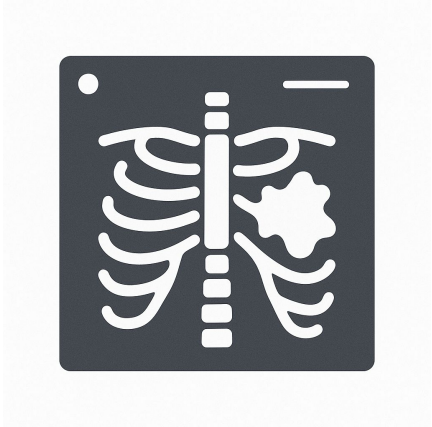
?

Data Extraction

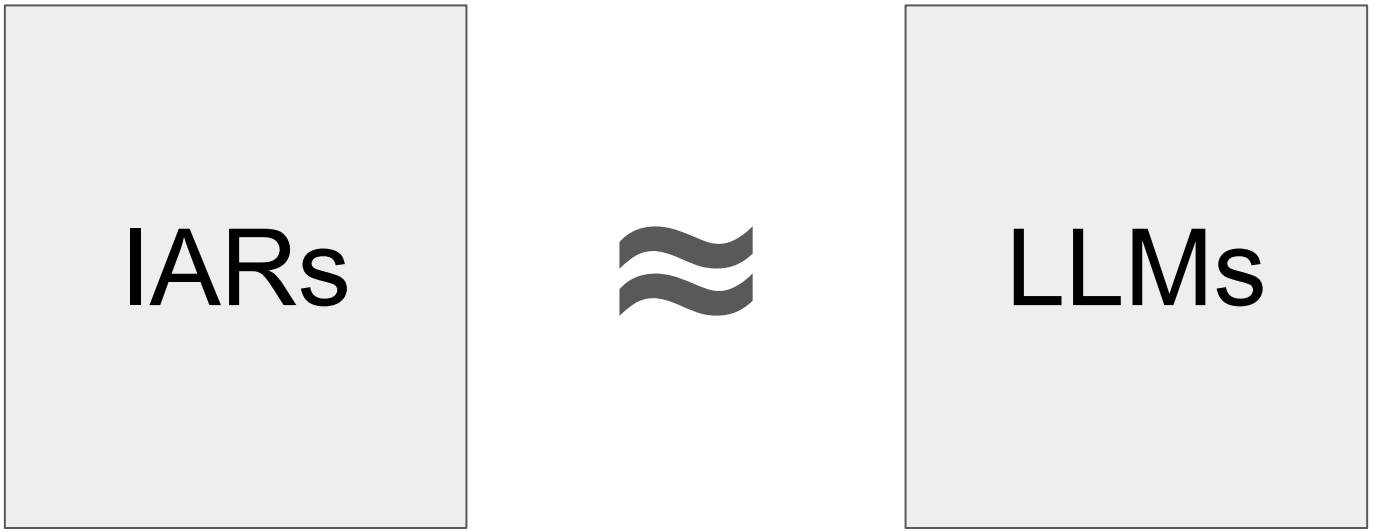


Generate a Scan

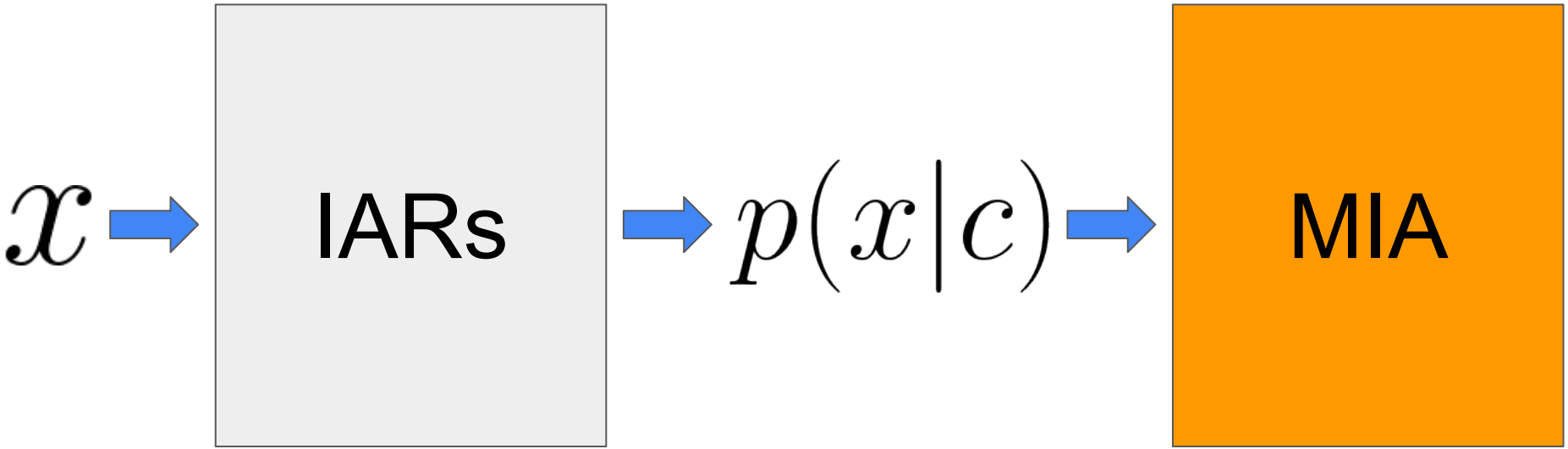
OK



Privacy Leakage: MIA

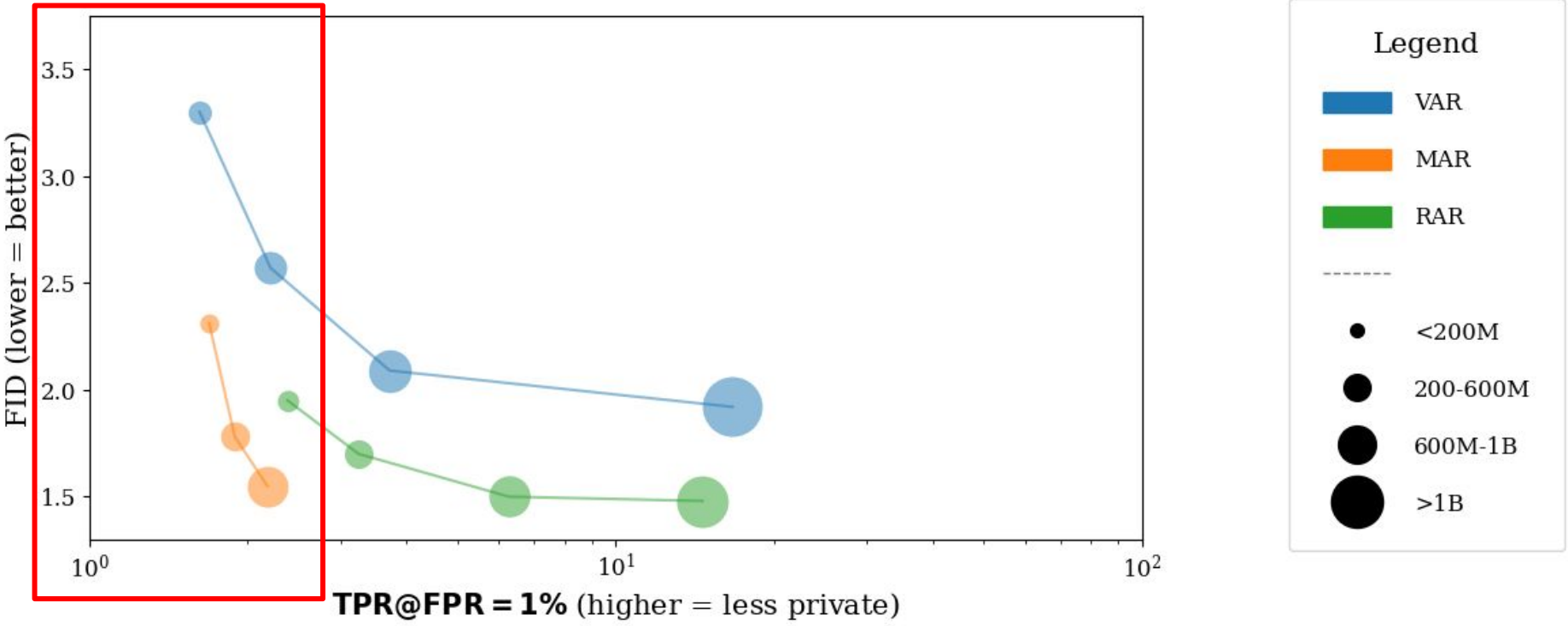


Privacy Leakage: MIA

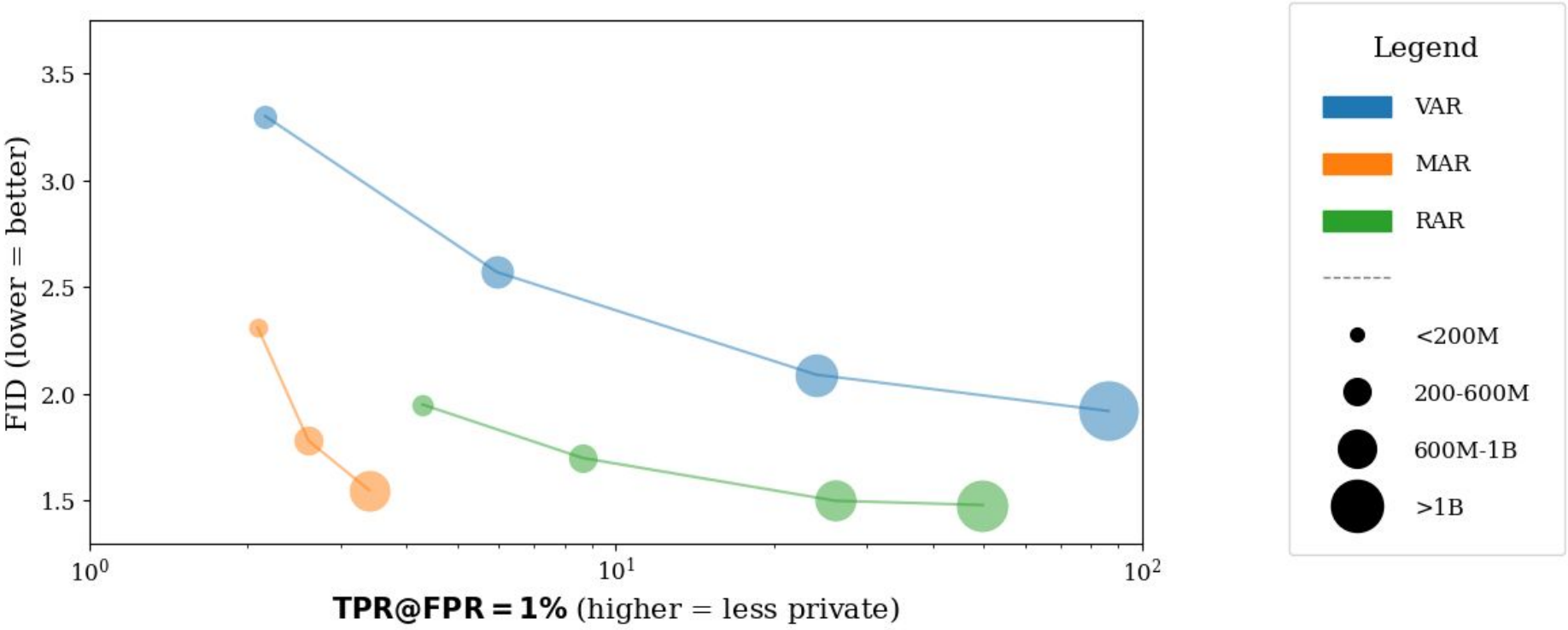


LLM MIA

Random Guessing



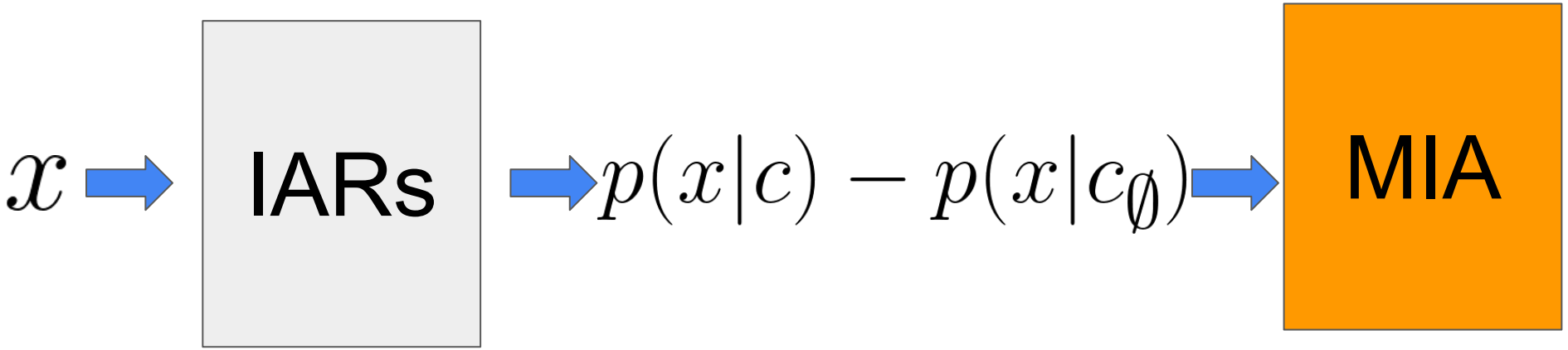
Our MIA


3.5
3.0
2.5
2.0
1.5
FID (lower = better)
10^0
10^1
10^2
TPR@FPR = 1% (higher = less private)
Legend
VAR
MAR
RAR
<200M
200-600M
600M-1B
>1B

Our MIA

VAR & RAR
$p(x\|c) - p(x\|c_\emptyset)$

Our MIA



Our MIA

VAR & RAR	MAR
$p(x\|c) - p(x\|c_\emptyset)$	DM-specific modifications

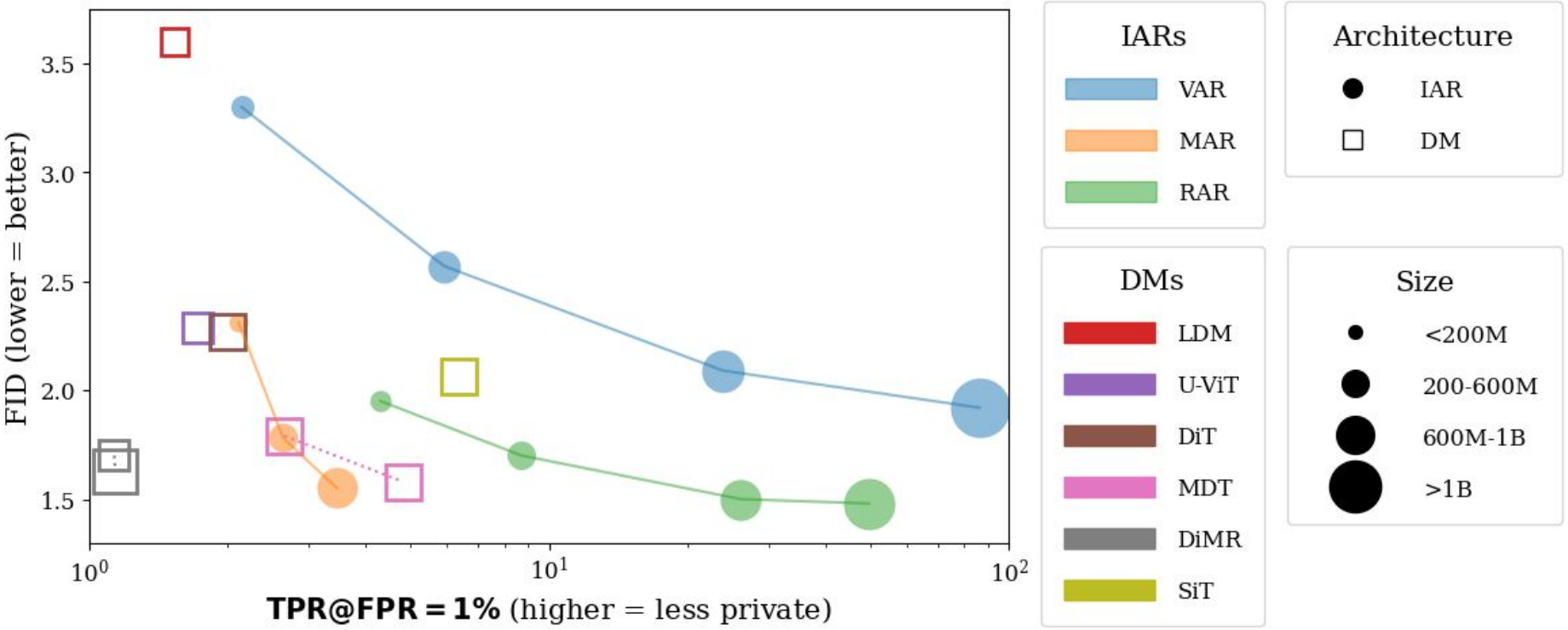
Our improvement

	VAR-*d*30	**RAR-XXL**	**MAR-H**
Baseline TPR@FPR=1%	16.68	14.62	2.18

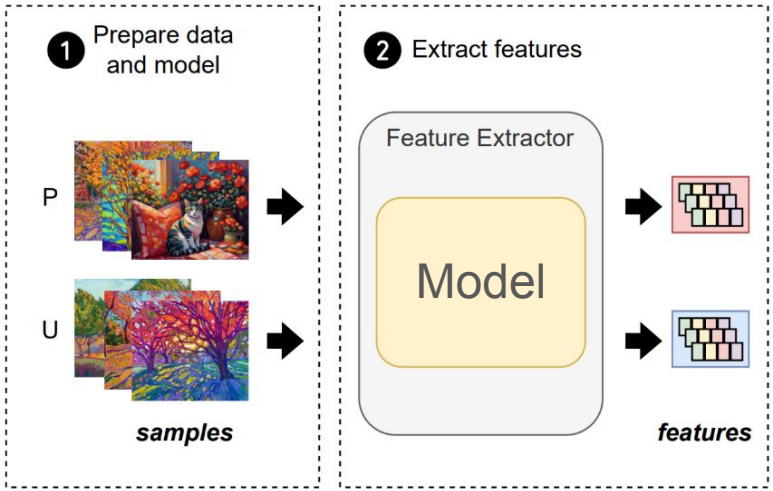
Our improvement

	VAR-*d*30	**RAR-XXL**	**MAR-H**
Baseline TPR@FPR=1%	16.68	14.62	2.18
Ours TPR@FPR=1%	**86.38**	**49.80**	**3.40**

IARs leak more than DMs

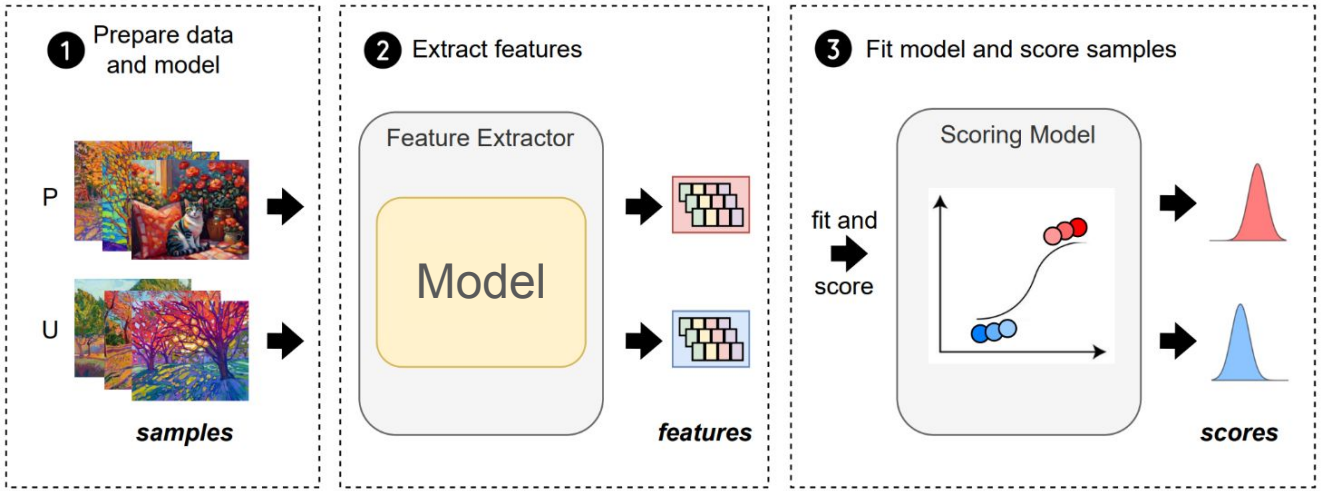


Dataset Inference (DI)



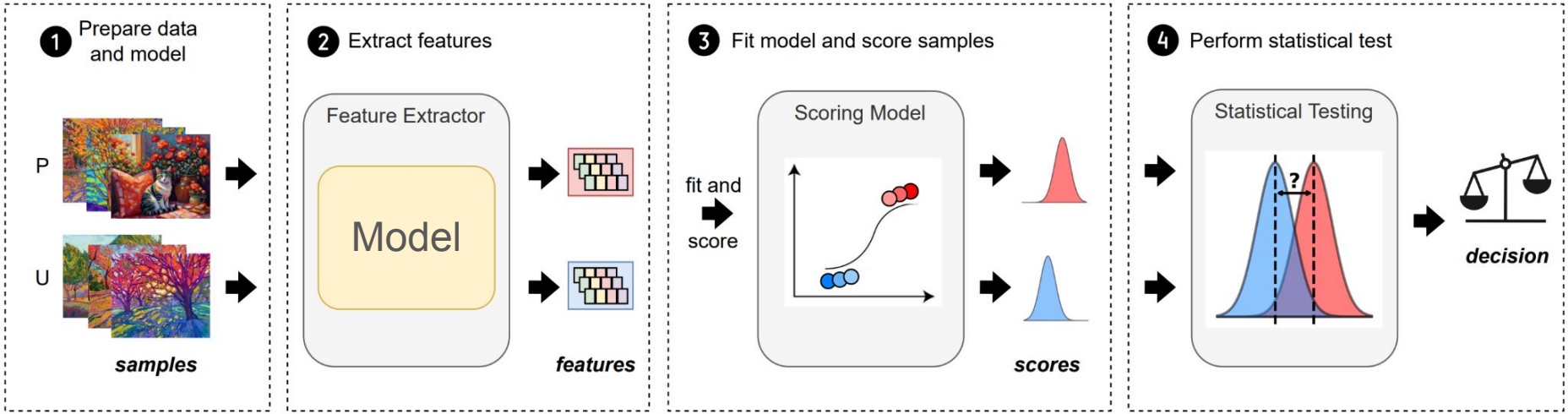
Jan Dubiński*, Antoni Kowalczuk*, Franziska Boenisch, Adam Dziedzic
"CDI: Copyrighted Data Identification in Diffusion Models", [CVPR 2025]

Original DI



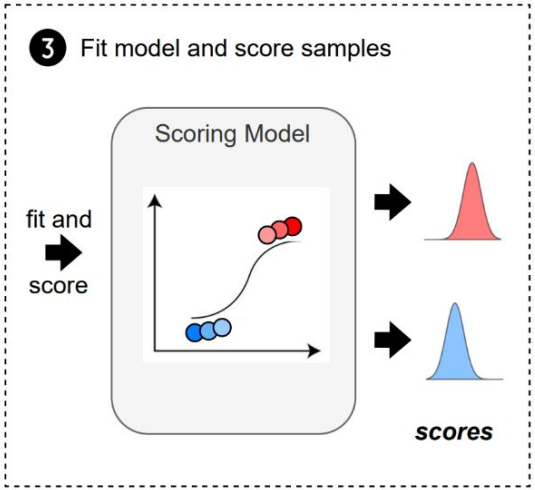
Jan Dubiński*, Antoni Kowalczuk*, Franziska Boenisch, Adam Dziedzic
"CDI: Copyrighted Data Identification in Diffusion Models", [CVPR 2025]

Original DI



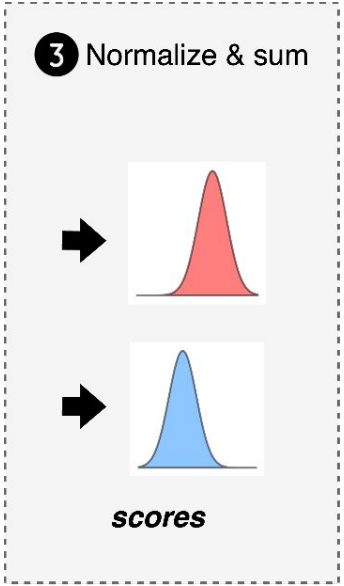
Jan Dubiński*, Antoni Kowalczuk*, Franziska Boenisch, Adam Dziedzic
"CDI: Copyrighted Data Identification in Diffusion Models", [CVPR 2025]

DI

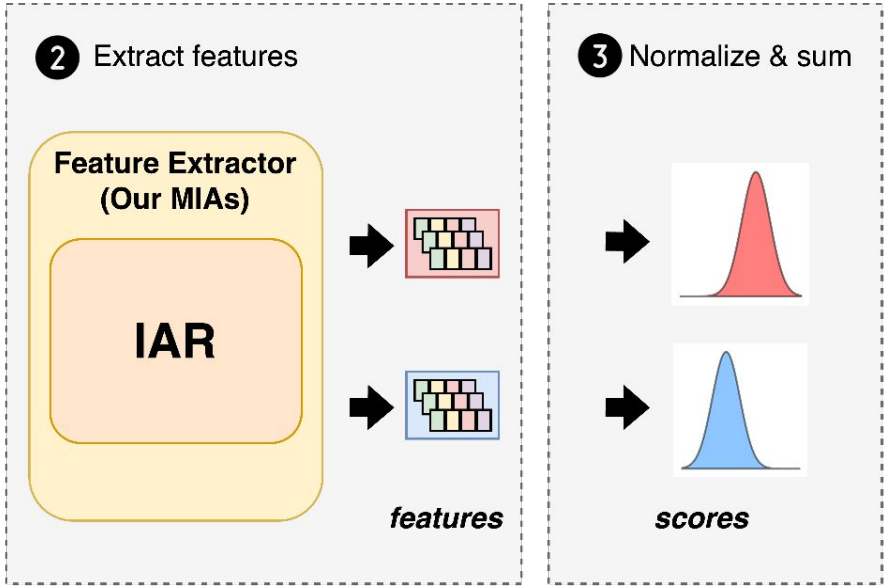


Jan Dubiński*, Antoni Kowalczuk*, Franziska Boenisch, Adam Dziedzic
"CDI: Copyrighted Data Identification in Diffusion Models", [CVPR 2025]

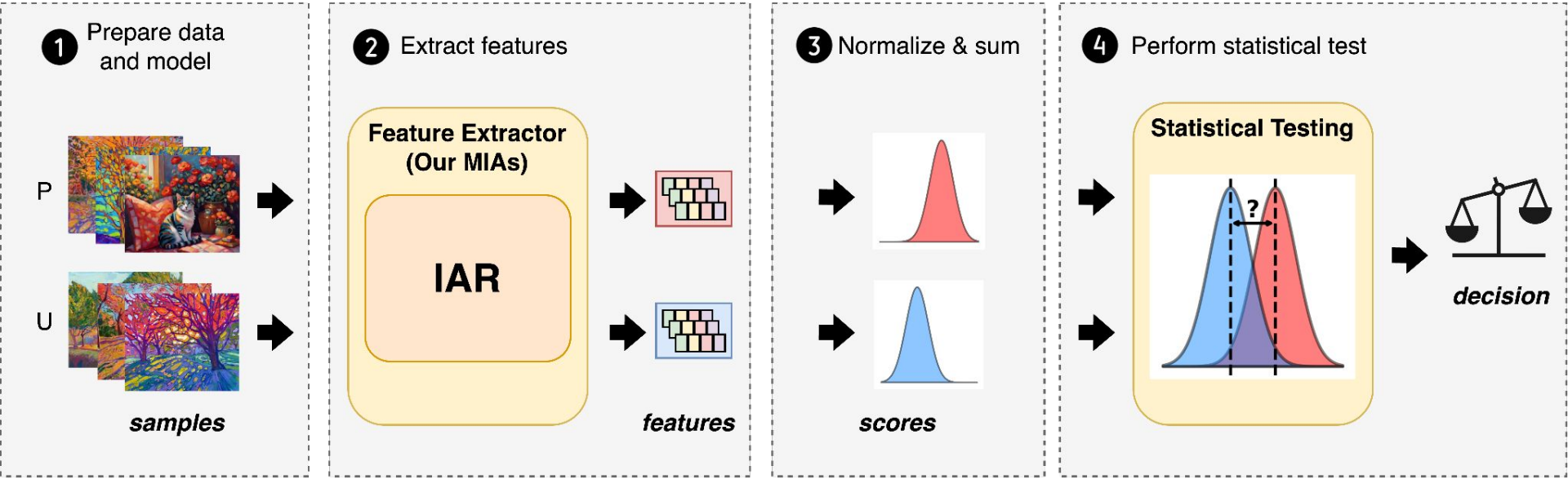
Our DI



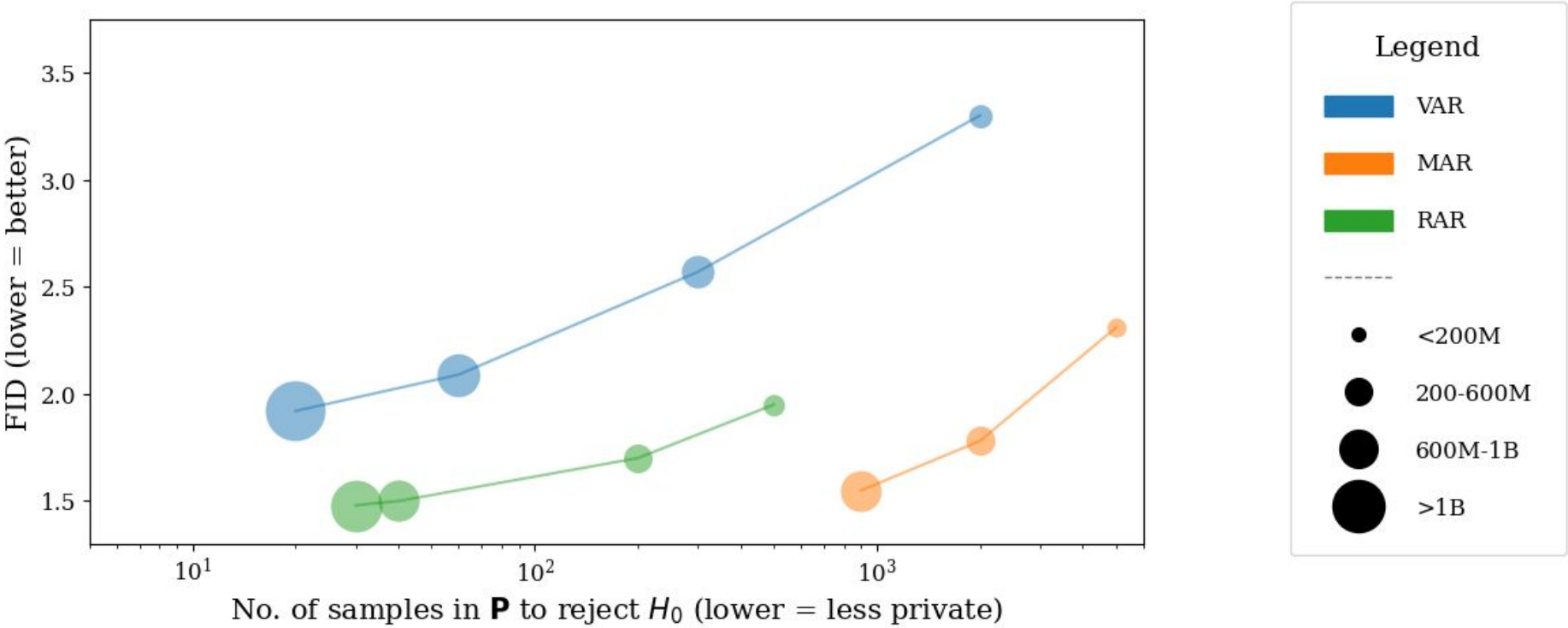
Our DI


2 Extract features
Feature Extractor
(Our MIAs)
IAR
features
3 Normalize & sum
scores

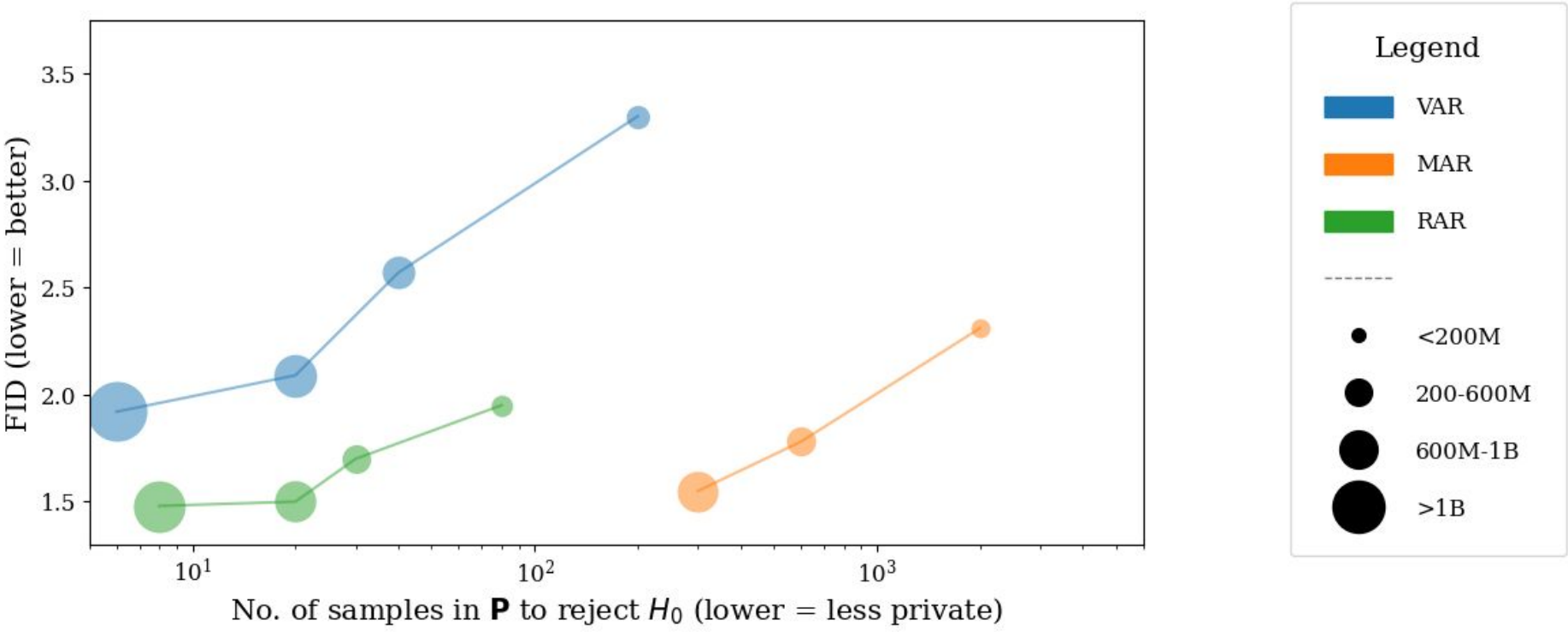
Our DI


1 Prepare data and model
P
U
samples
2 Extract features
Feature Extractor (Our MIAs)
IAR
features
3 Normalize & sum
scores
4 Perform statistical test
Statistical Testing
?
decision

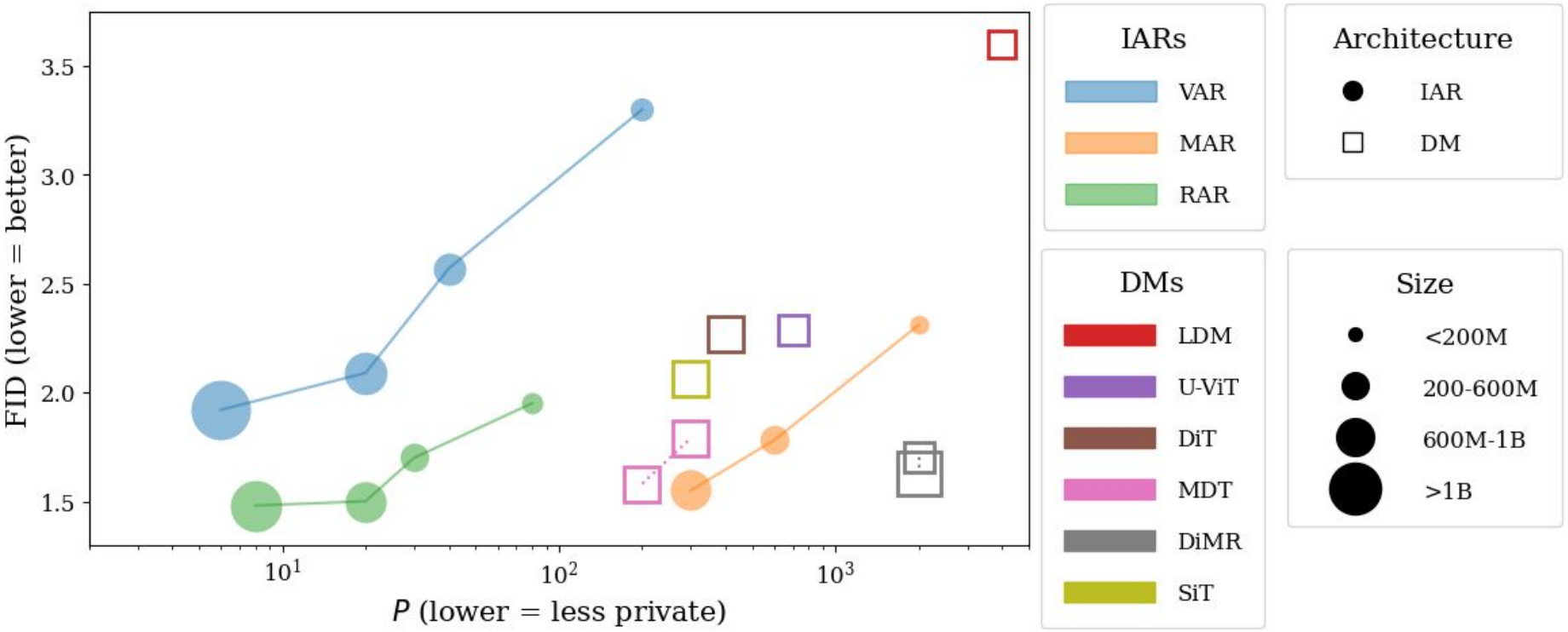
DI: before


FID (lower = better)
3.5
3.0
2.5
2.0
1.5
10^1
10^2
10^3
No. of samples in **P** to reject H_0 (lower = less private)
Legend
VAR
MAR
RAR
<200M
200-600M
600M-1B
>1B

DI: after


FID (lower = better)
3.5
3.0
2.5
2.0
1.5
10^1
10^2
10^3
No. of samples in **P** to reject H_0 (lower = less private)
Legend
VAR
MAR
RAR
<200M
200-600M
600M-1B
>1B

DMs are less prone to DI

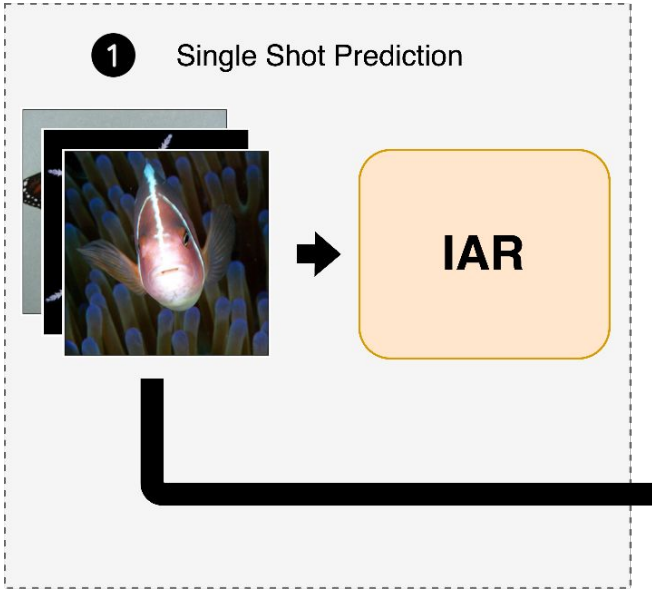


Data Extraction

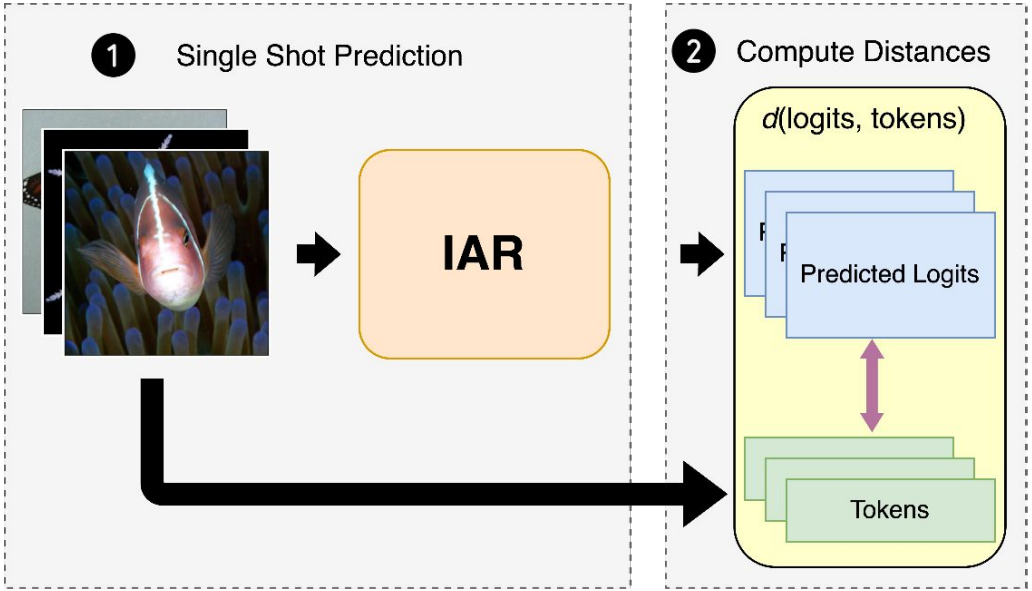
Training Sample

Generated Image

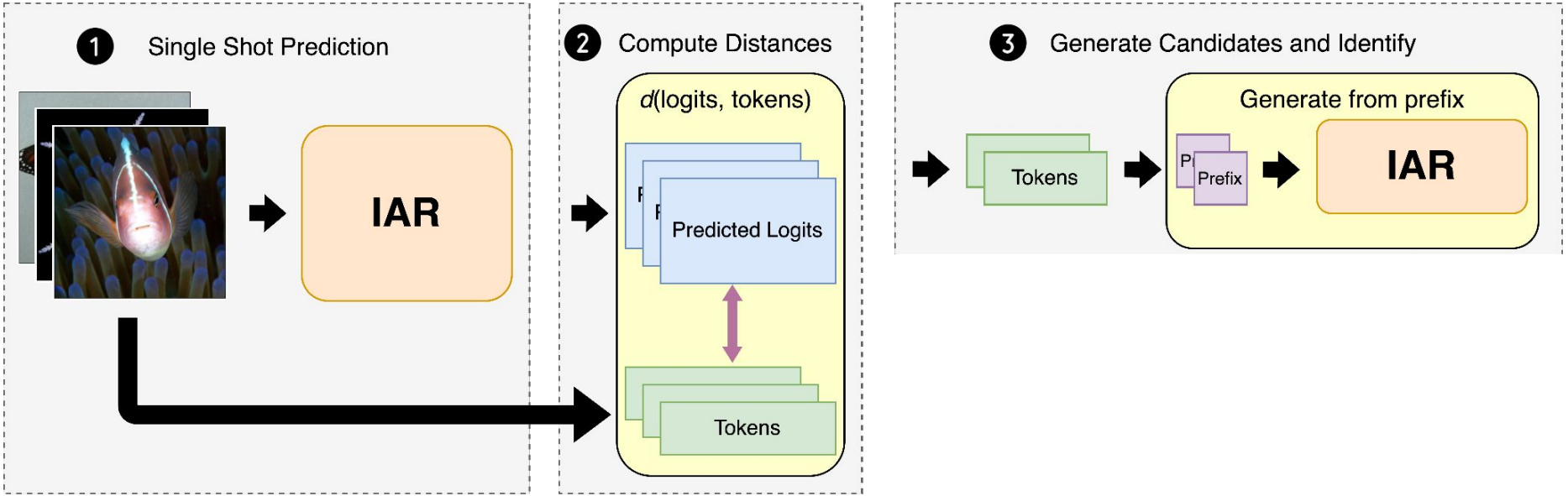
Our Data Extraction



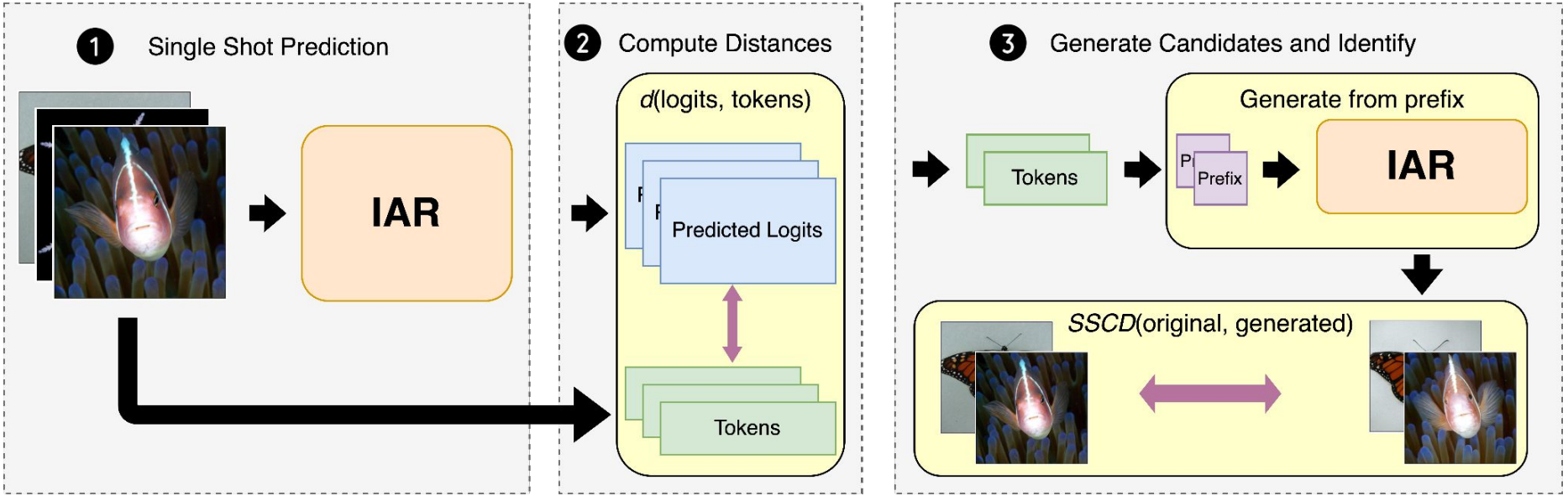
Our Data Extraction



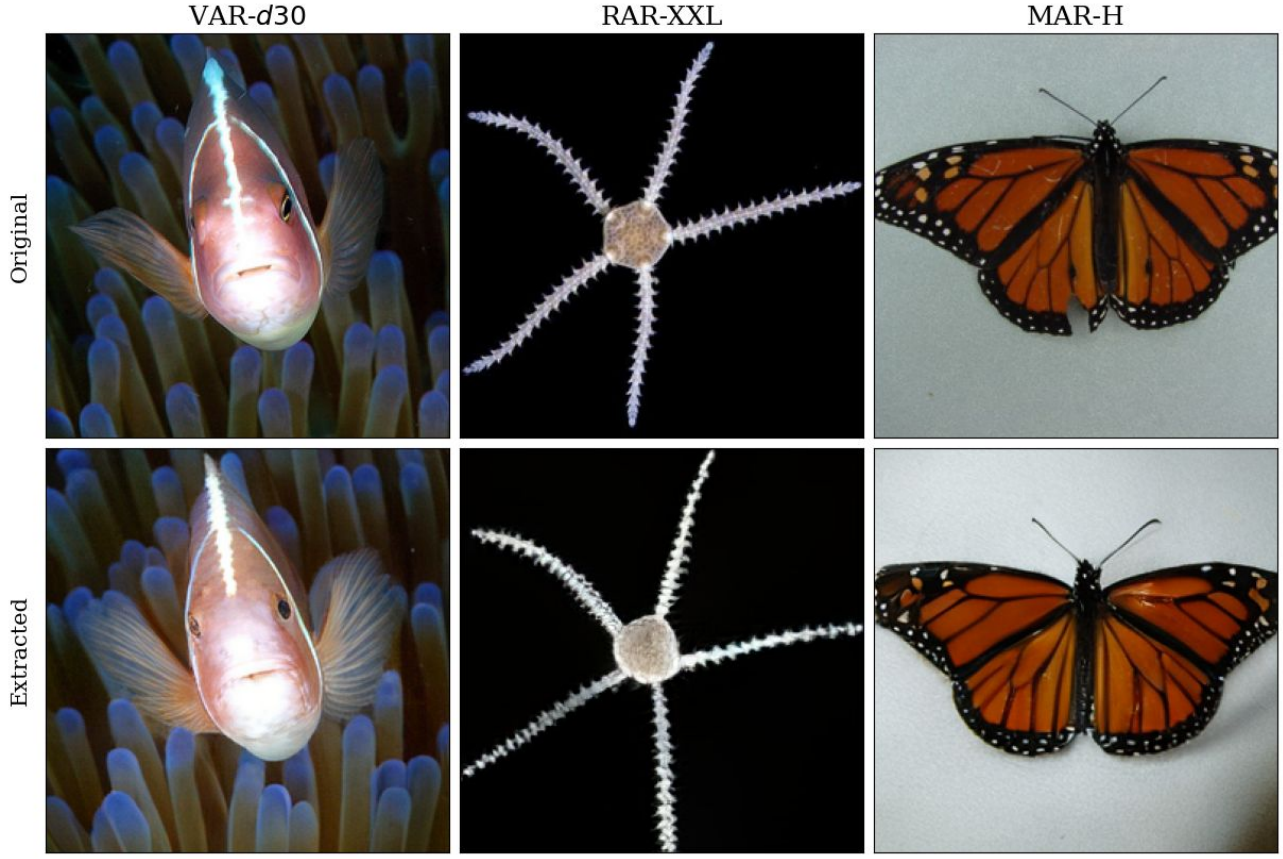
Our Data Extraction



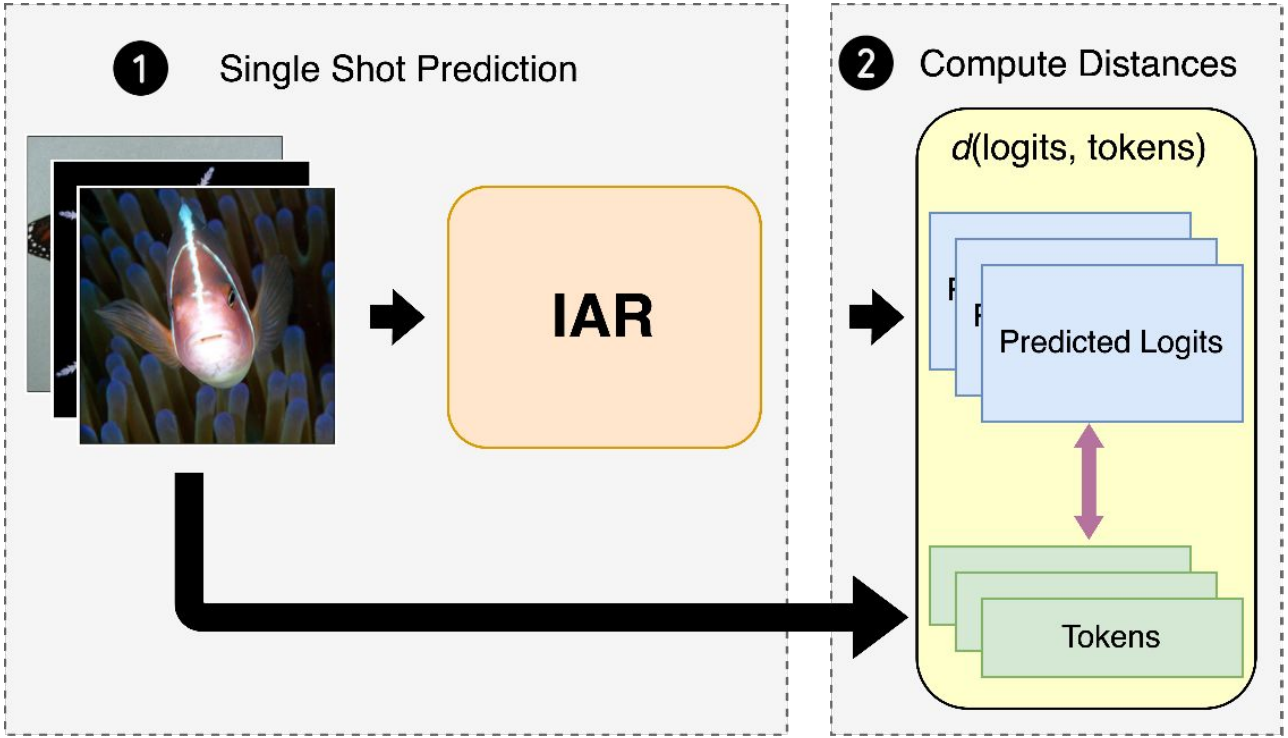
Our Data Extraction



Successful Extraction



Efficient Candidate Selection



Why do IARs leak more?

Reasons for Higher Leakage

IARs	**DMs**
Access to p(x)	No access to p(x)

Reasons for Higher Leakage

IARs	**DMs**
Access to p(x)	No access to p(x)
More data per sample	Single noisy image

Reasons for Higher Leakage

IARs	**DMs**
Access to p(x)	No access to p(x)
More data per sample	Single noisy image
Each token leaks	Single leakage

Contributions

1. **First** to investigate IARs' privacy

Contributions

1. **First** to investigate IARs' privacy
2. New, **SOTA** privacy attacks against IARs

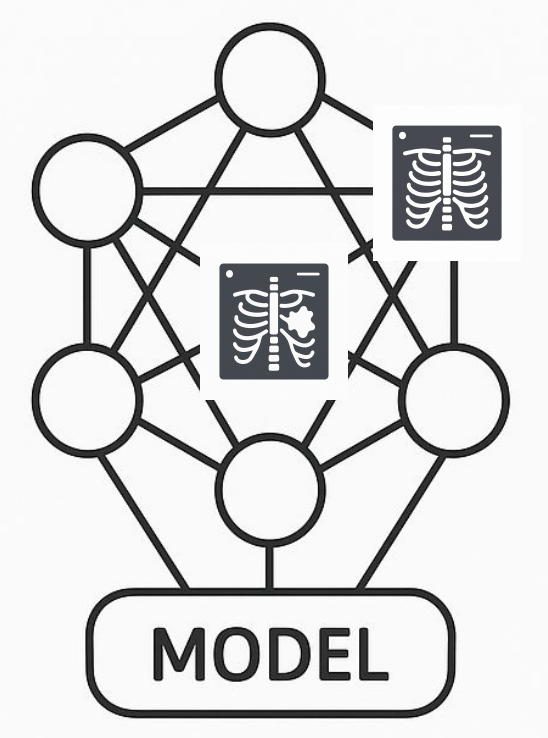
Contributions

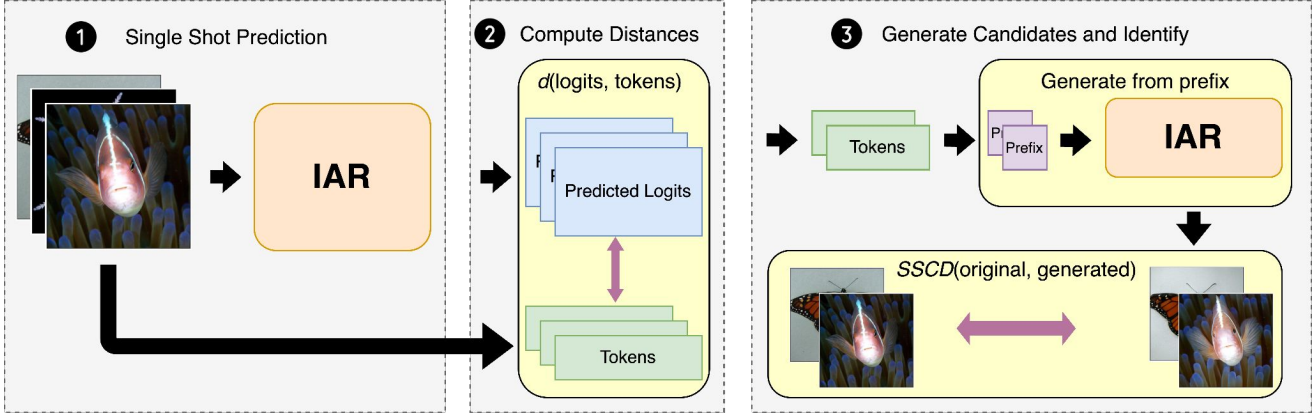
1. **First** to investigate IARs' privacy
2. New, **SOTA** privacy attacks against IARs
3. **First** successful data extraction from IARs

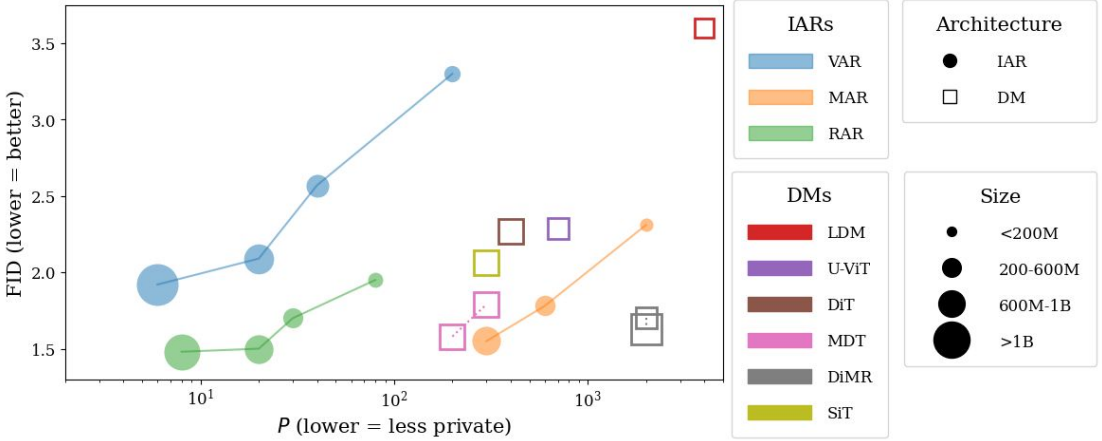
Contributions

1. **First** to investigate IARs' privacy
2. New, **SOTA** privacy attacks against IARs
3. **First** successful data extraction from IARs
4. Explanation **why** IARs leak more than DMs

Thank you!


MODEL


1 Single Shot Prediction
IAR
2 Compute Distances
d(logits, tokens)
Predicted Logits
Tokens
3 Generate Candidates and Identify
Tokens
Generate from prefix
Prefix
IAR
SSCD(original, generated)


FID (lower = better)
3.5
3.0
2.5
2.0
1.5
10^1
10^2
10^3
P (lower = less private)
IARs
VAR
MAR
RAR
Architecture
IAR
DM
DMs
LDM
U-ViT
DiT
MDT
DiMR
SiT
Size
<200M
200-600M
600M-1B
>1B