Privacy Attacks on Image AutoRegressive Models

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

Prague, 13.03.2025

Short intro about me

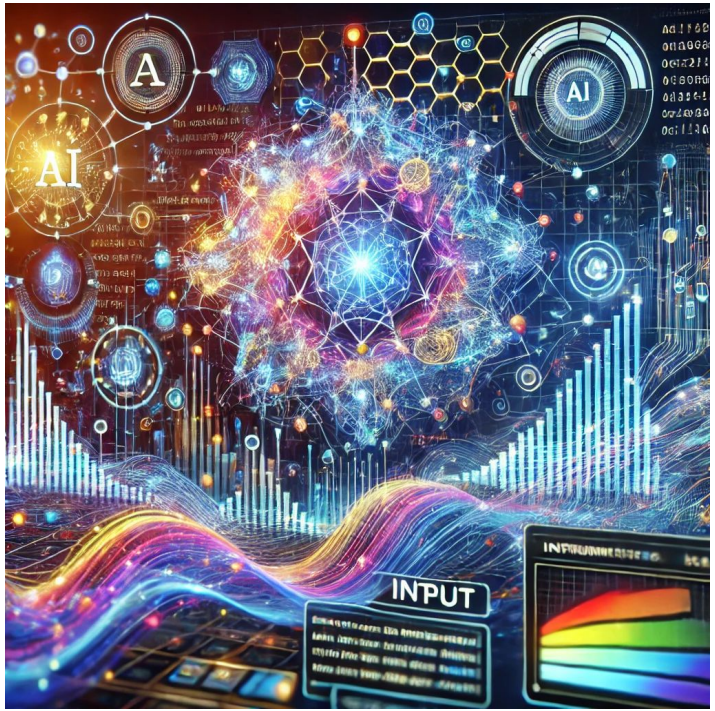
Short intro about me

CISPA
HELMHOLTZ CENTER FOR
INFORMATION SECURITY

SprintML

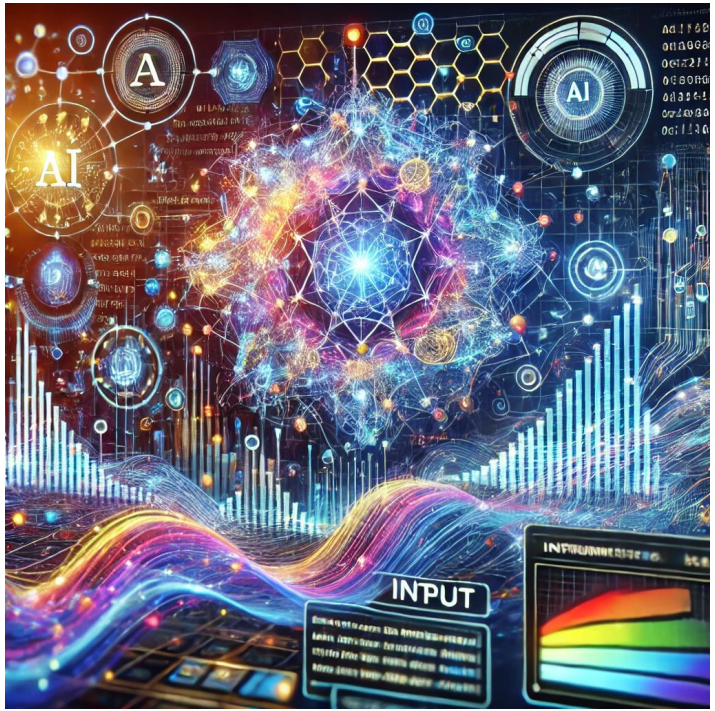
My research interests

Image generation



My research interests

Image generation



Data Privacy

Why data privacy?

Reuters

World Business Markets Sustainability Legal Breakingviews Technology In

Getty Images lawsuit says Stability AI misused photos to train AI

By **Blake Brittain**

February 6, 2023 6:32 PM GMT+1 · Updated 2 years ago

Aa

Why data privacy?

Reuters

World Business Markets Sustainability Legal Breakingviews Technology In

Lawsuits accuse AI content creators of misusing copyrighted work

By **Blake Brittain**

January 17, 2023 9:05 PM GMT+1 · Updated 2 years ago

Aa

Why data privacy?

Reuters

World Business Markets Sustainability Legal Breakingviews Technology Investigati

Artists take new shot at Stability, Midjourney in updated copyright lawsuit

By **Blake Brittain**

November 30, 2023 8:47 PM GMT+1 · Updated a year ago

Aa

Why data privacy?

Reuters

World Business Markets Sustainability Legal Breakingviews Technology

NY Times sues OpenAI, Microsoft for infringing copyrighted works

By **Jonathan Stempel**

December 28, 2023 12:50 AM GMT+1 · Updated a year ago

Image generation is cool :)

[Han J., et al. 2024]

What happens next:

1. Diffusion Models
2. LLMs -> Image AutoRegressive Models
3. Three models we attack!

Diffusion Models (DMs)

[Esser P., et al. 2024]

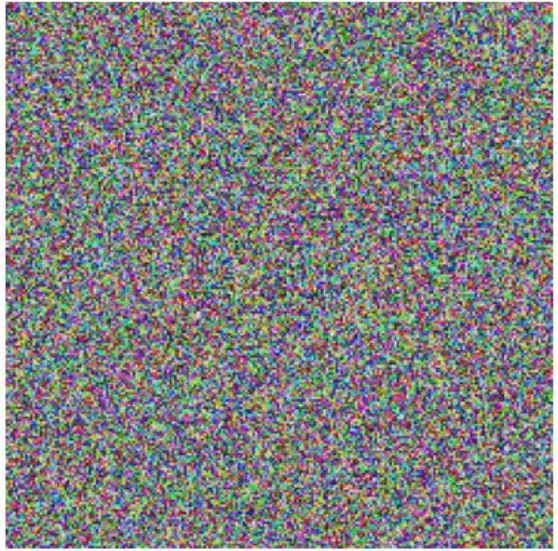
Key idea: iterative noising & de-noising

Noising

Key idea: iterative noising & de-noising

Noising

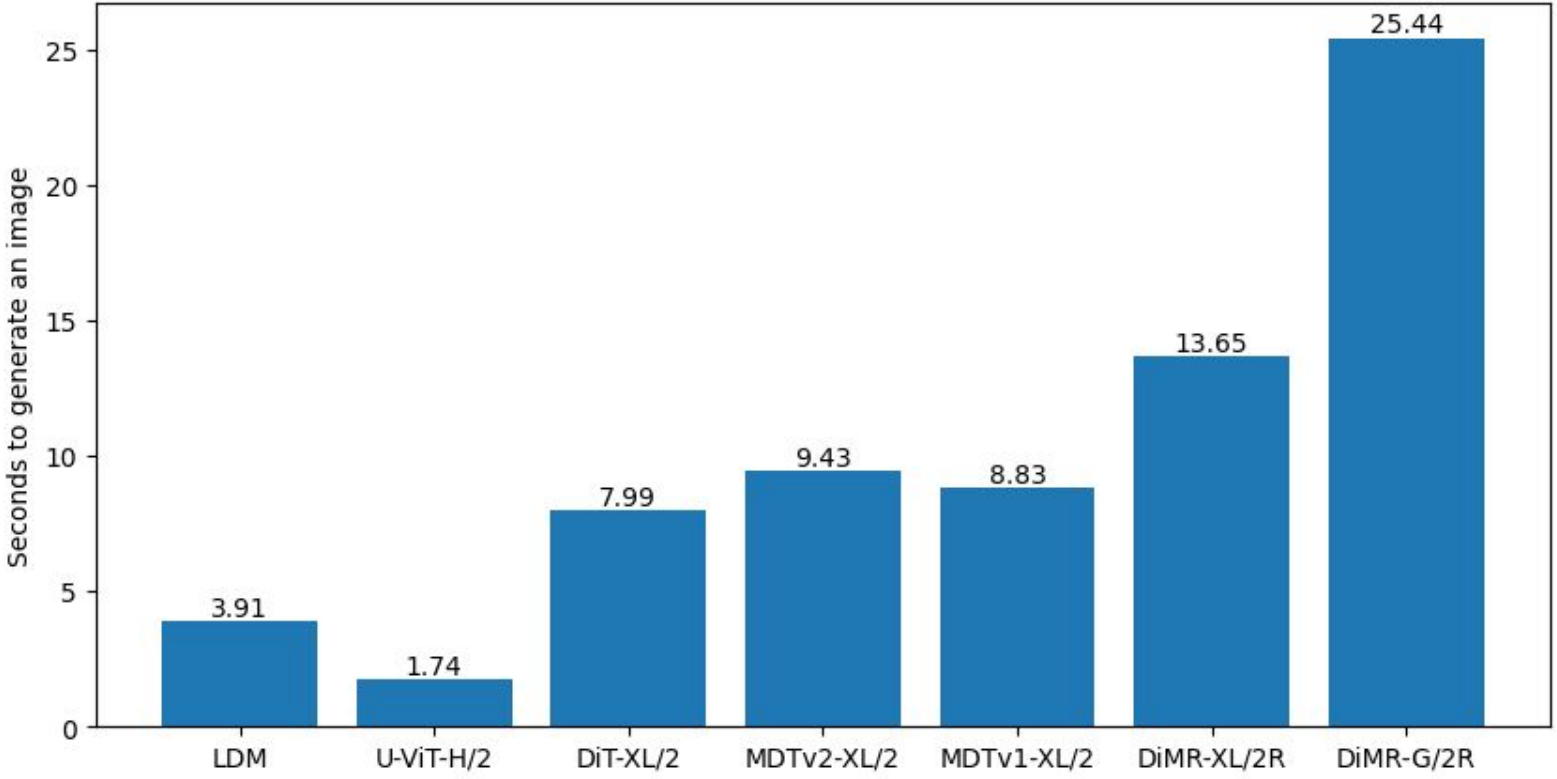
Denoising



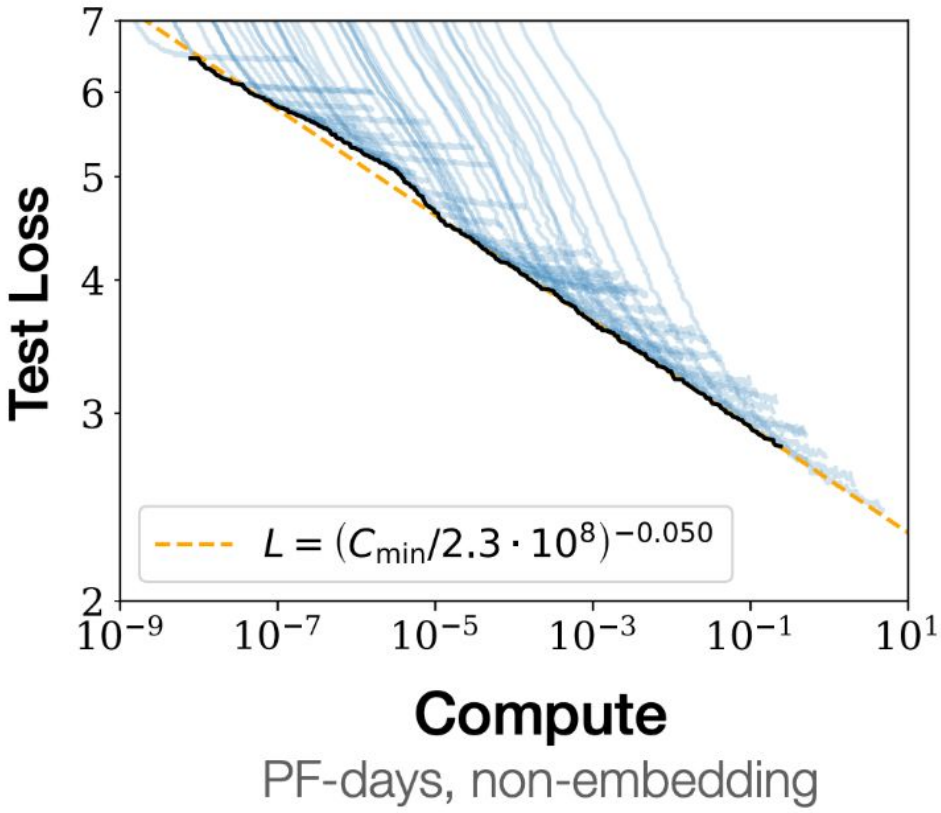
DMs are widely adopted

stability.ai

DMs are slow

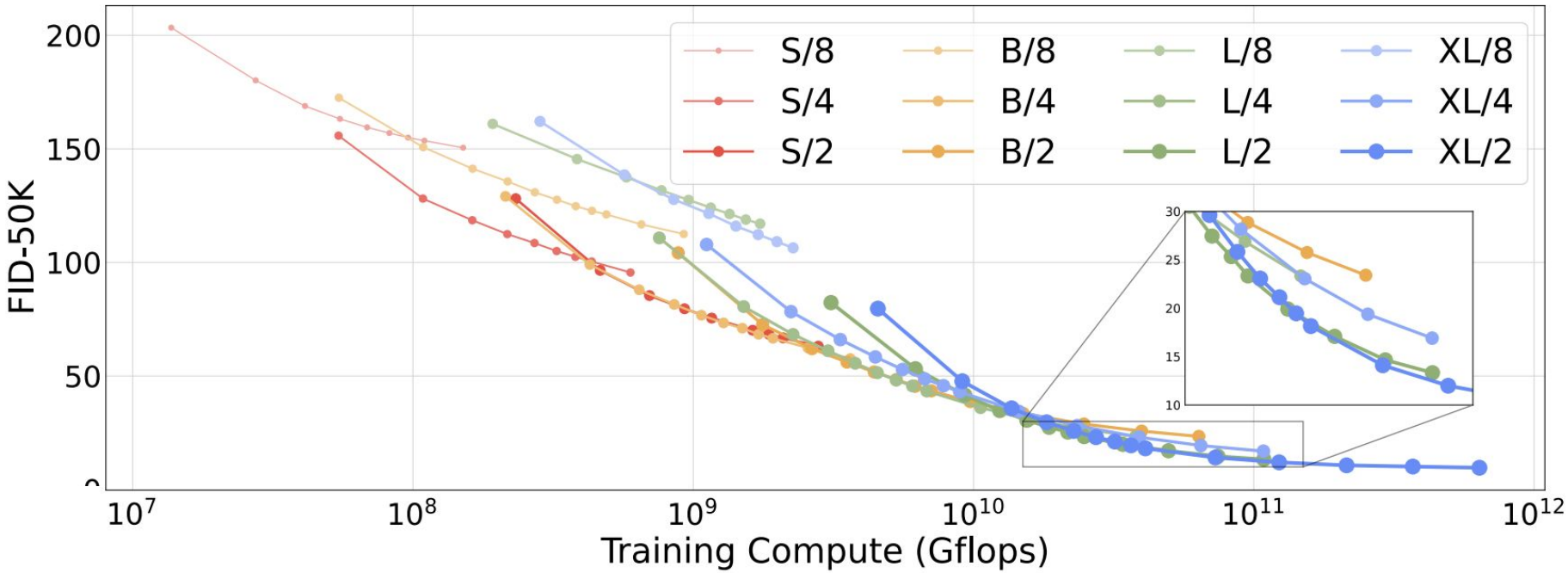

Seconds to generate an image
0
5
10
15
20
25
3.91
1.74
7.99
9.43
8.83
13.65
25.44
LDM
U-ViT-H/2
DiT-XL/2
MDTv2-XL/2
MDTv1-XL/2
DiMR-XL/2R
DiMR-G/2R

DMs do not scale well: expectation



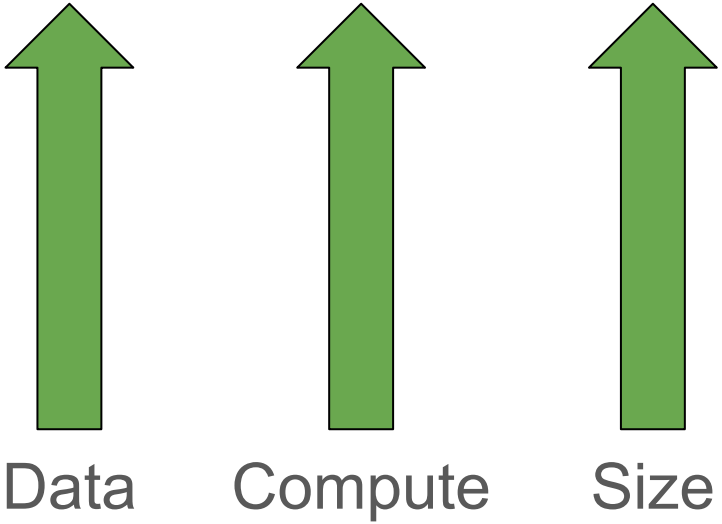
[Kaplan J., et al. 2020]

DMs do not scale well: reality

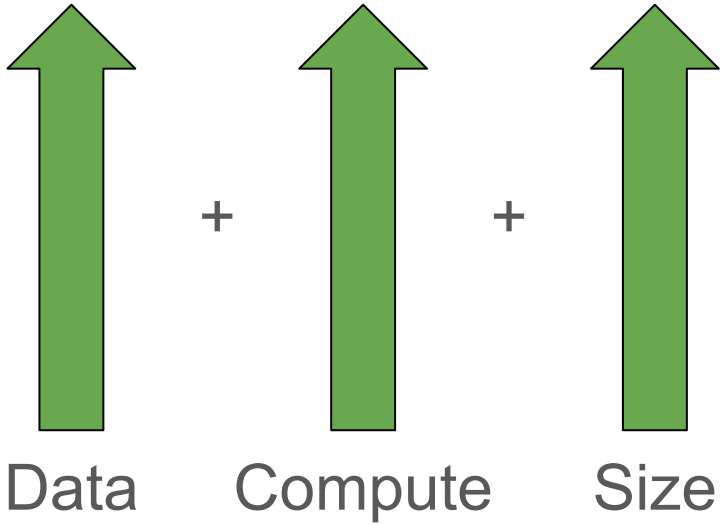


[Peebles W., et al. 2023]

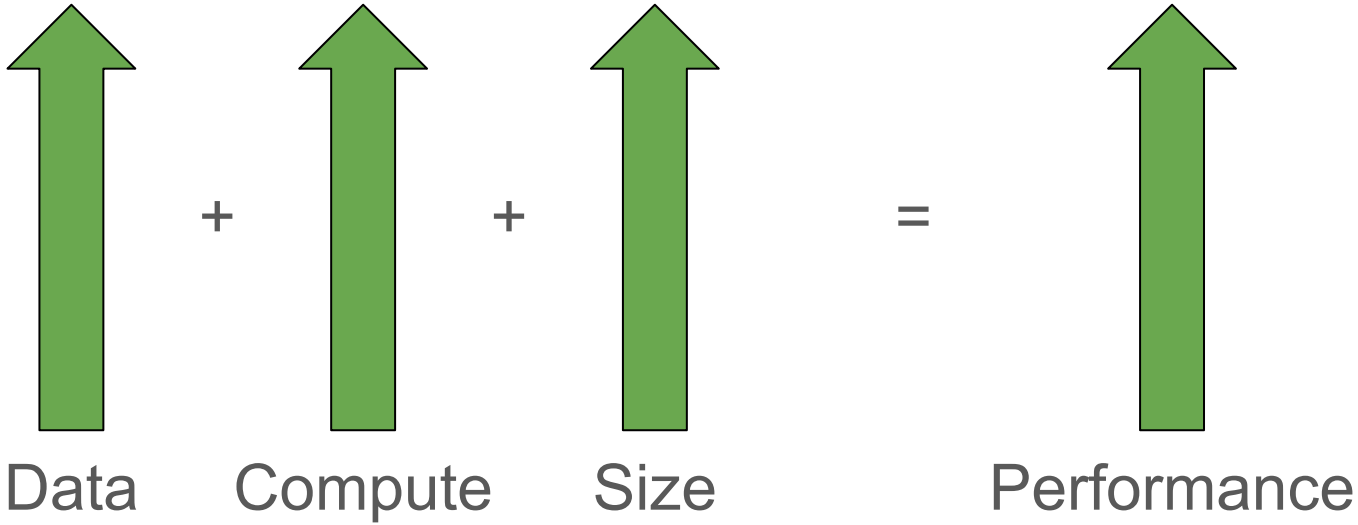
Why scaling is important?



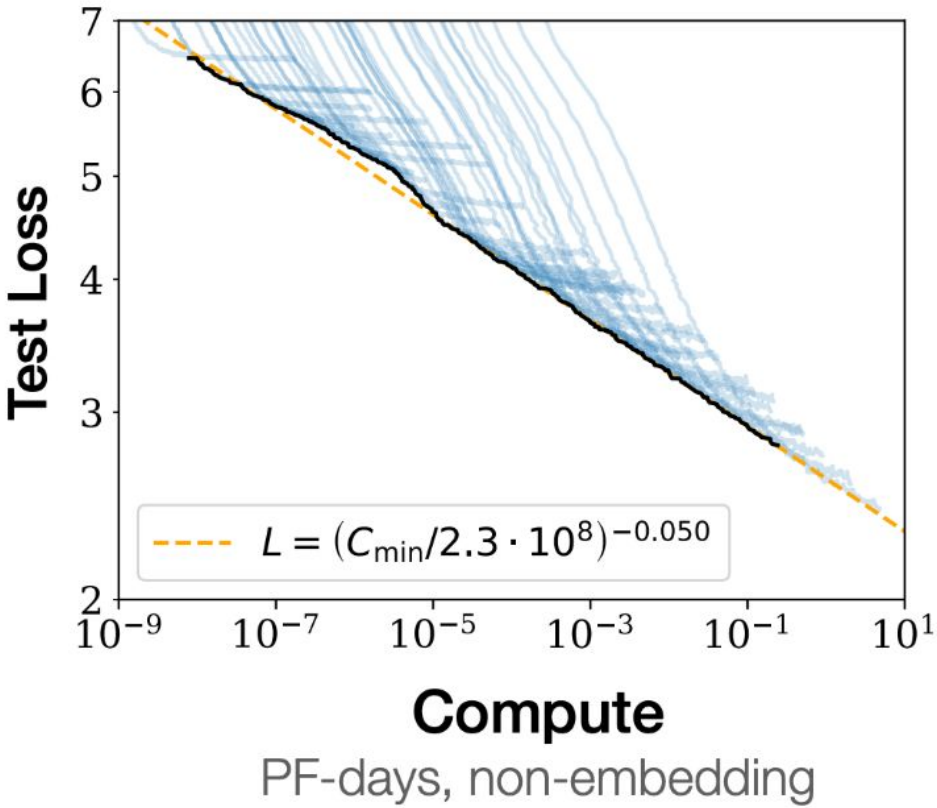
Why scaling is important?



Why scaling is important?

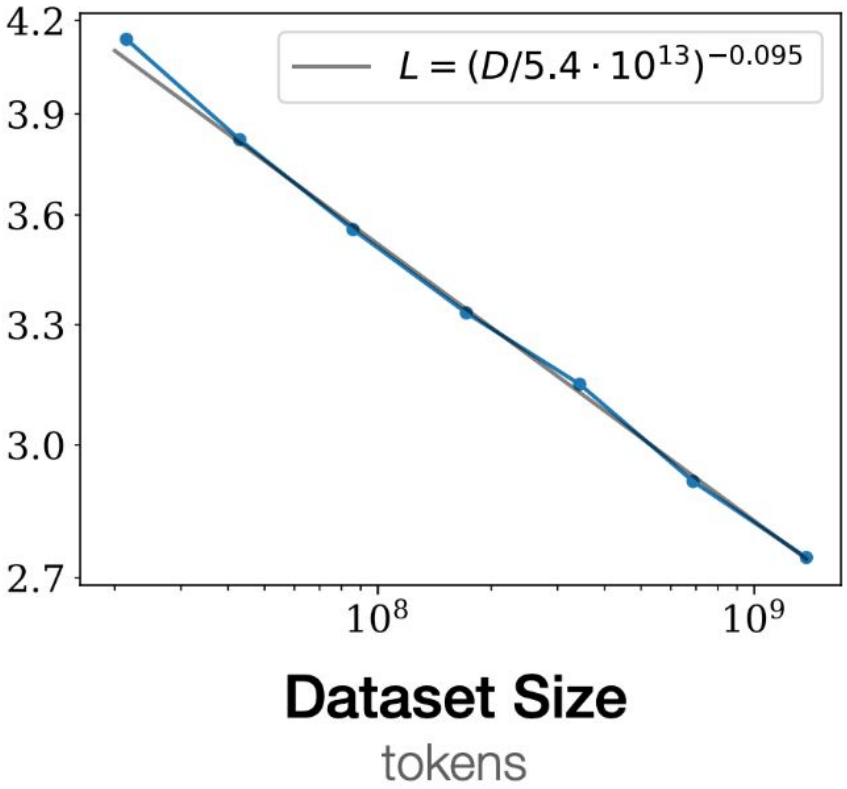


What scales well? LLMs!



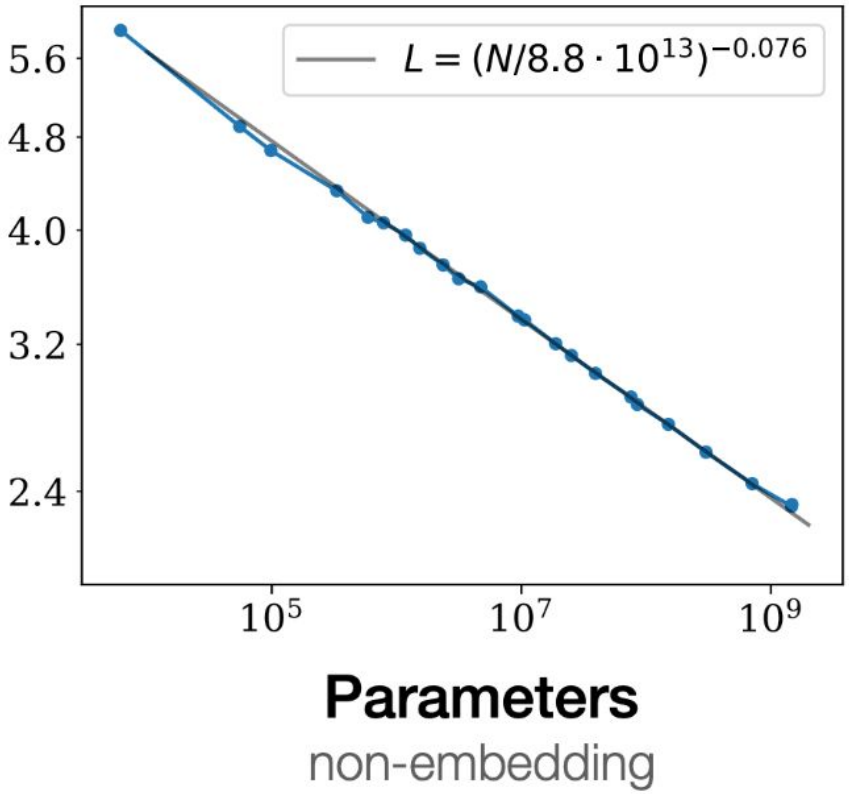
[Kaplan J., et al. 2020]

What scales well? LLMs!



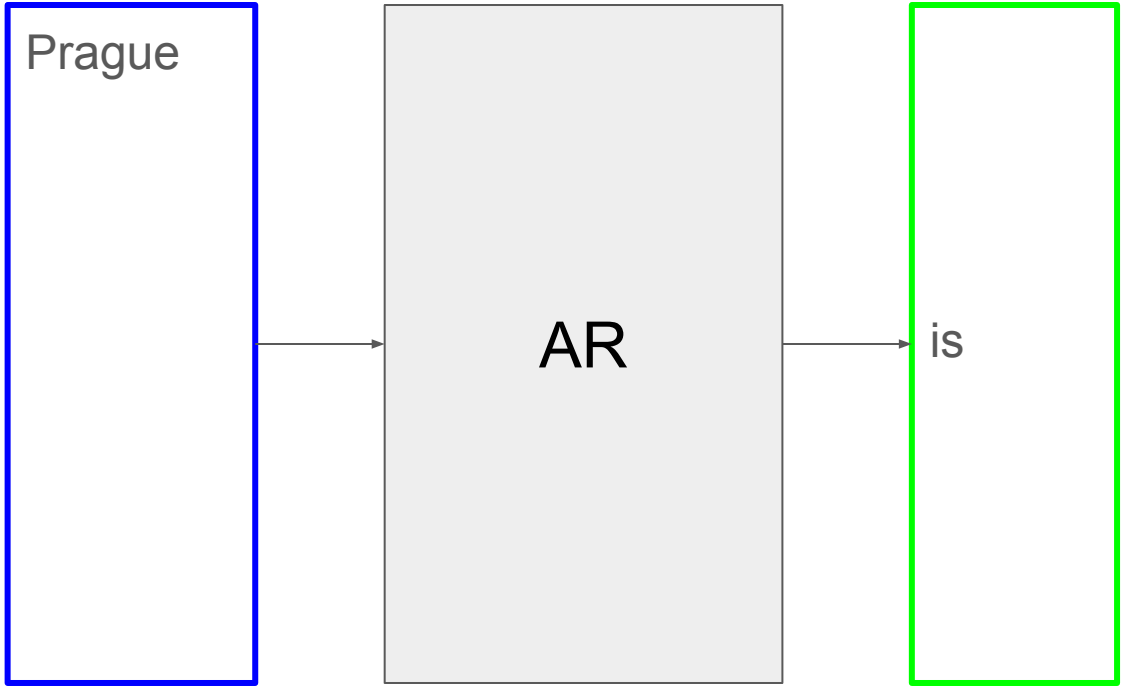
[Kaplan J., et al. 2020]

What scales well? LLMs!

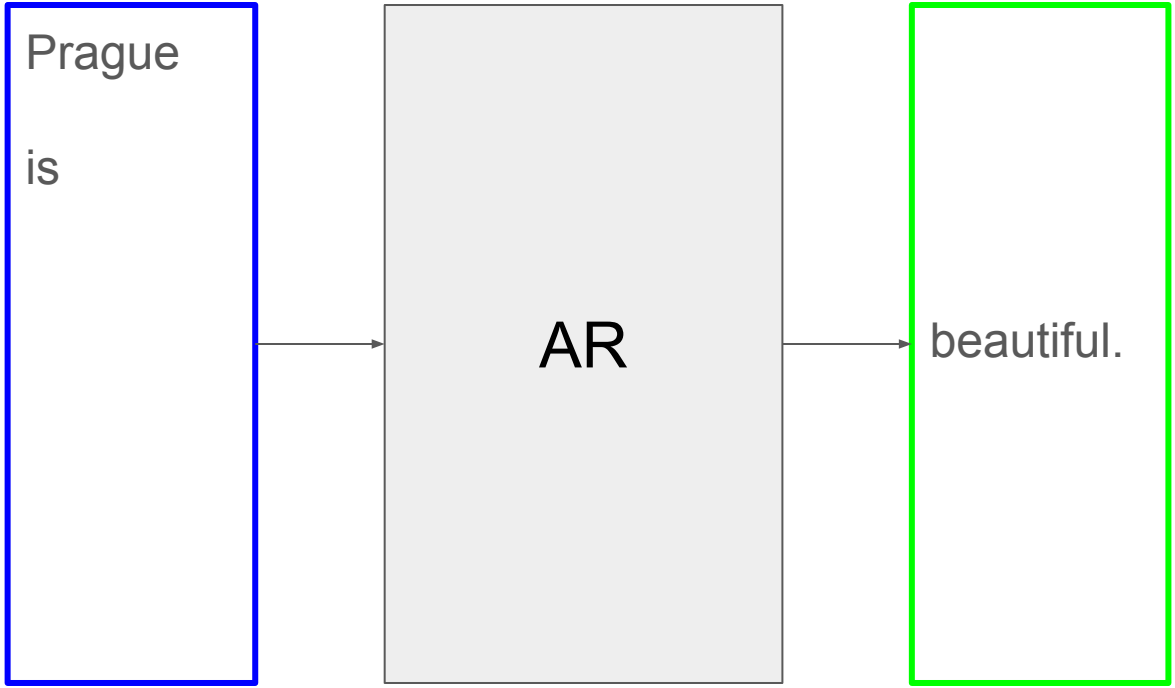


[Kaplan J., et al. 2020]

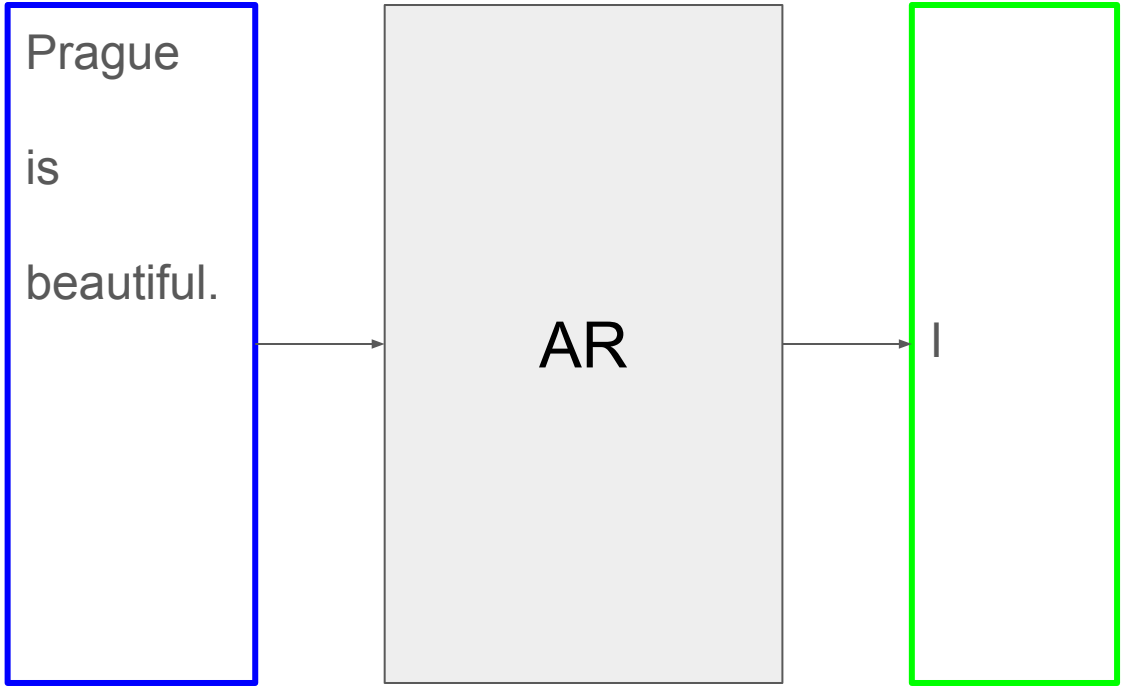
AutoRegressive (AR) prediction



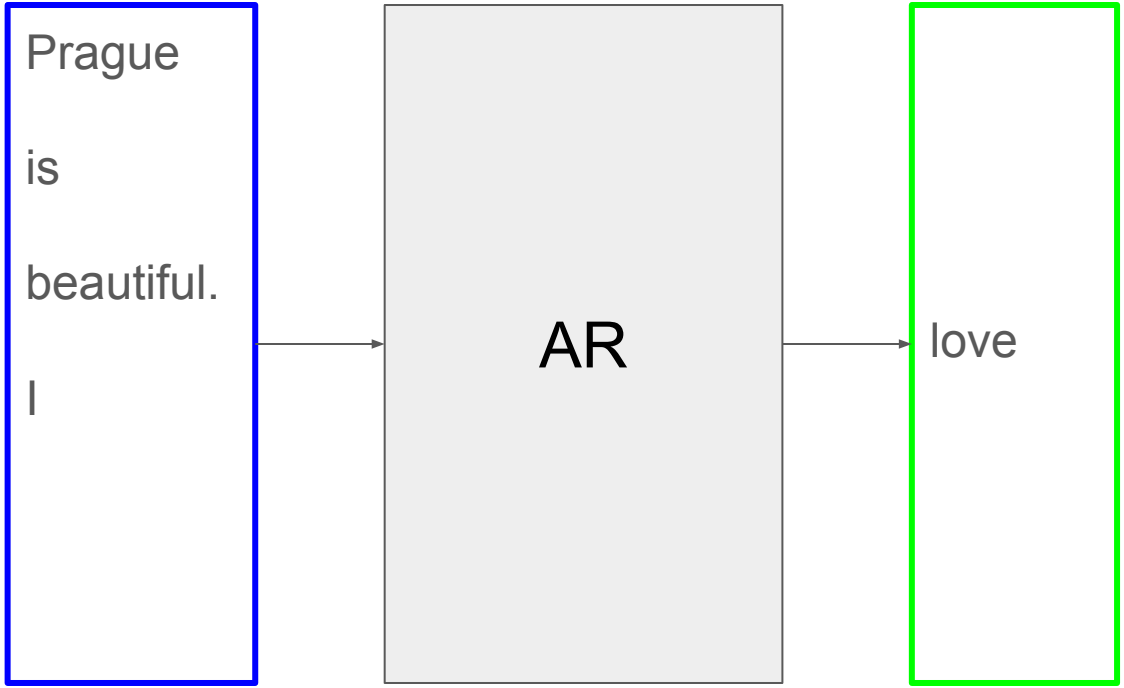
AutoRegressive (AR) prediction



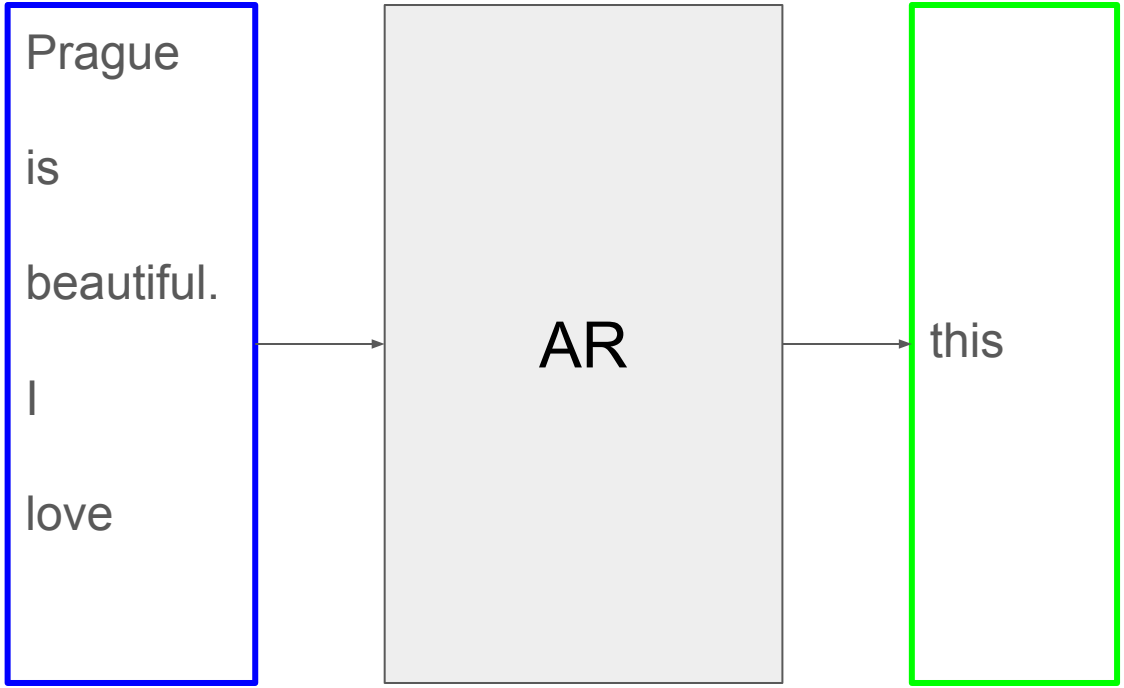
AutoRegressive (AR) prediction



AutoRegressive (AR) prediction



AutoRegressive (AR) prediction



AutoRegressive (AR) prediction

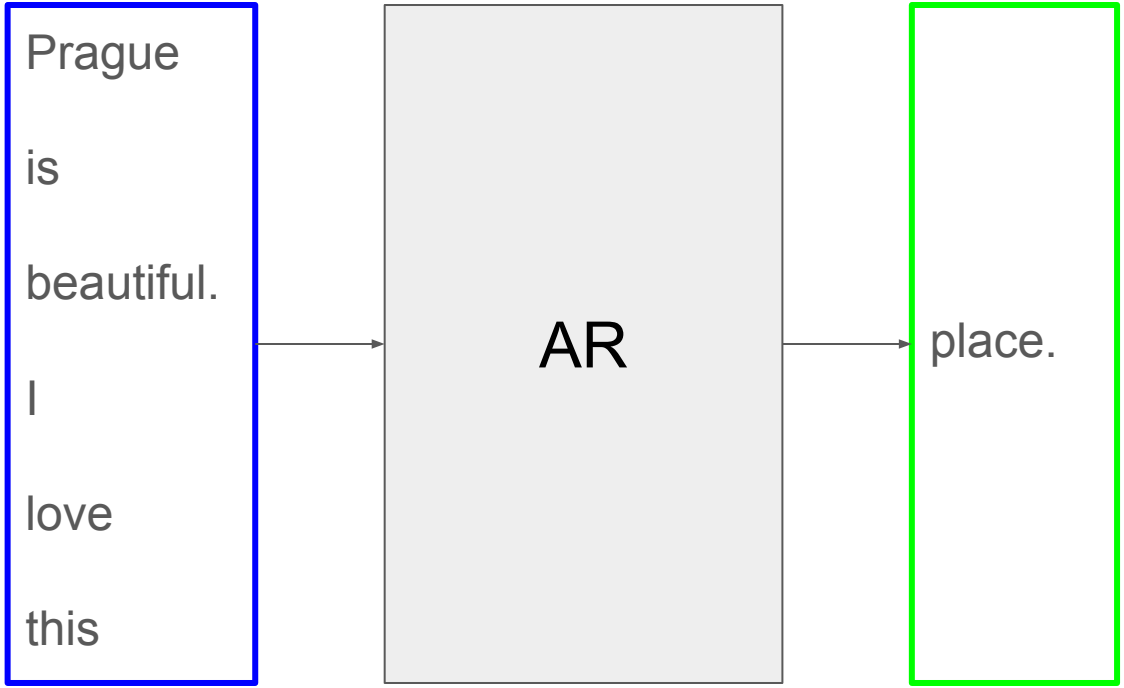


Image AutoRegressive Models (IARs)

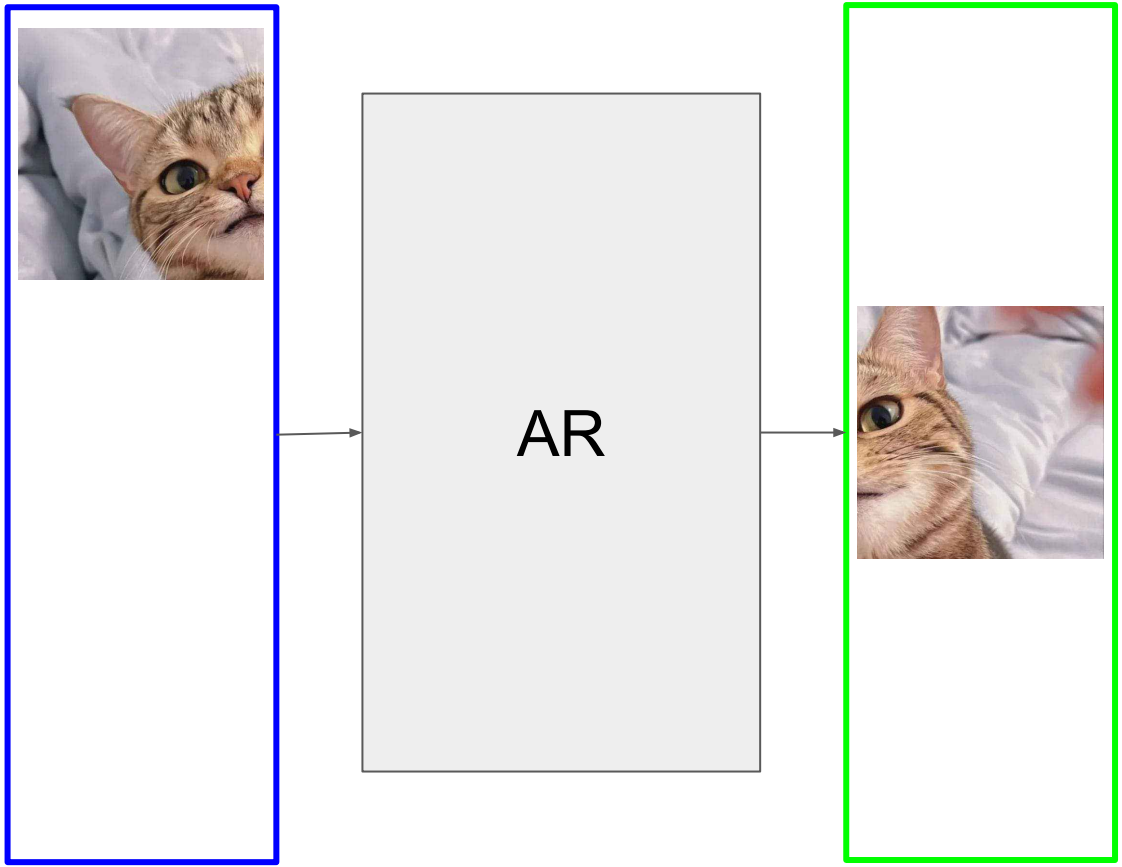


Image AutoRegressive Models (IARs)

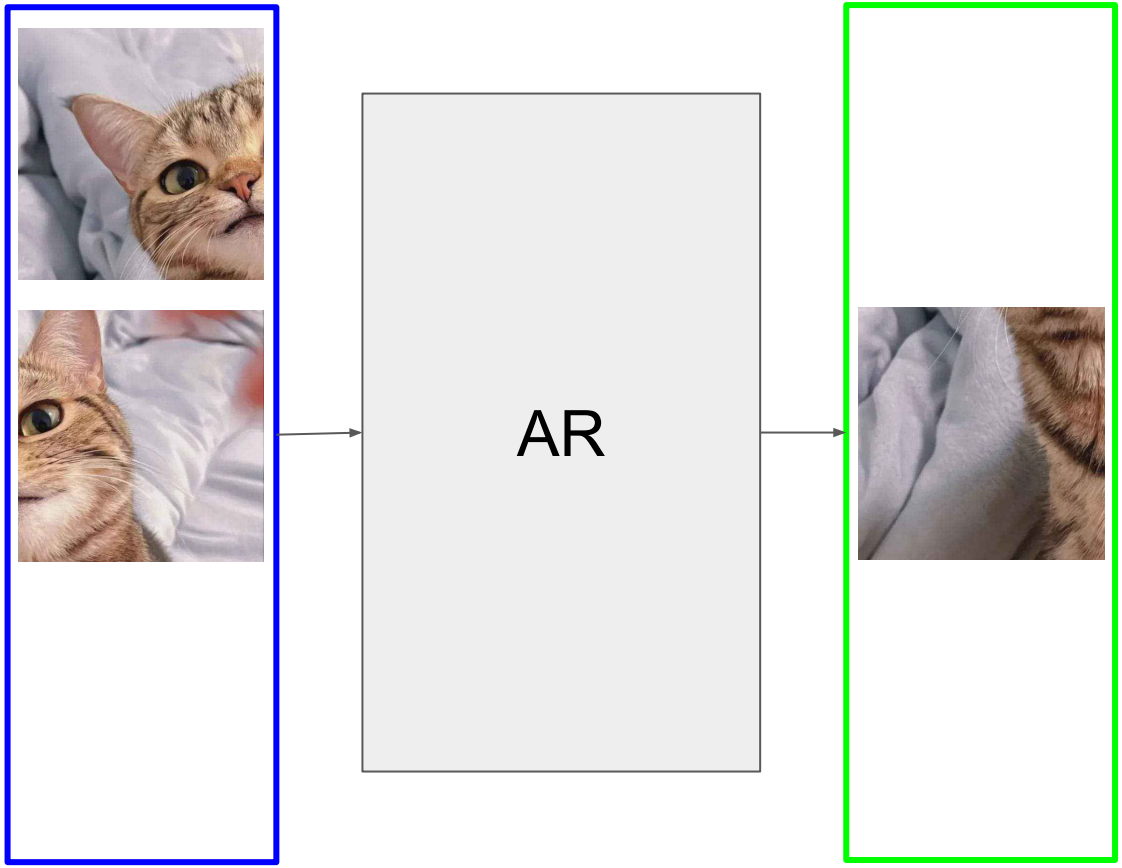


Image AutoRegressive Models (IARs)

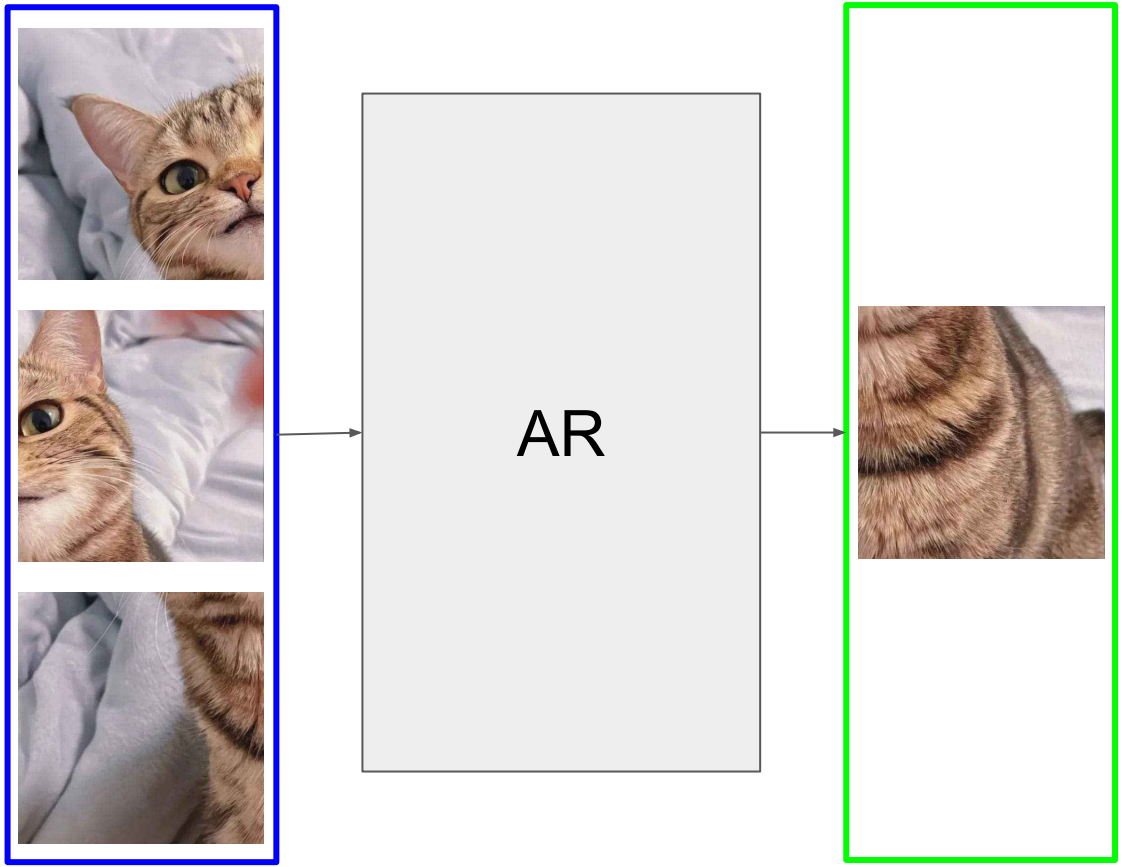
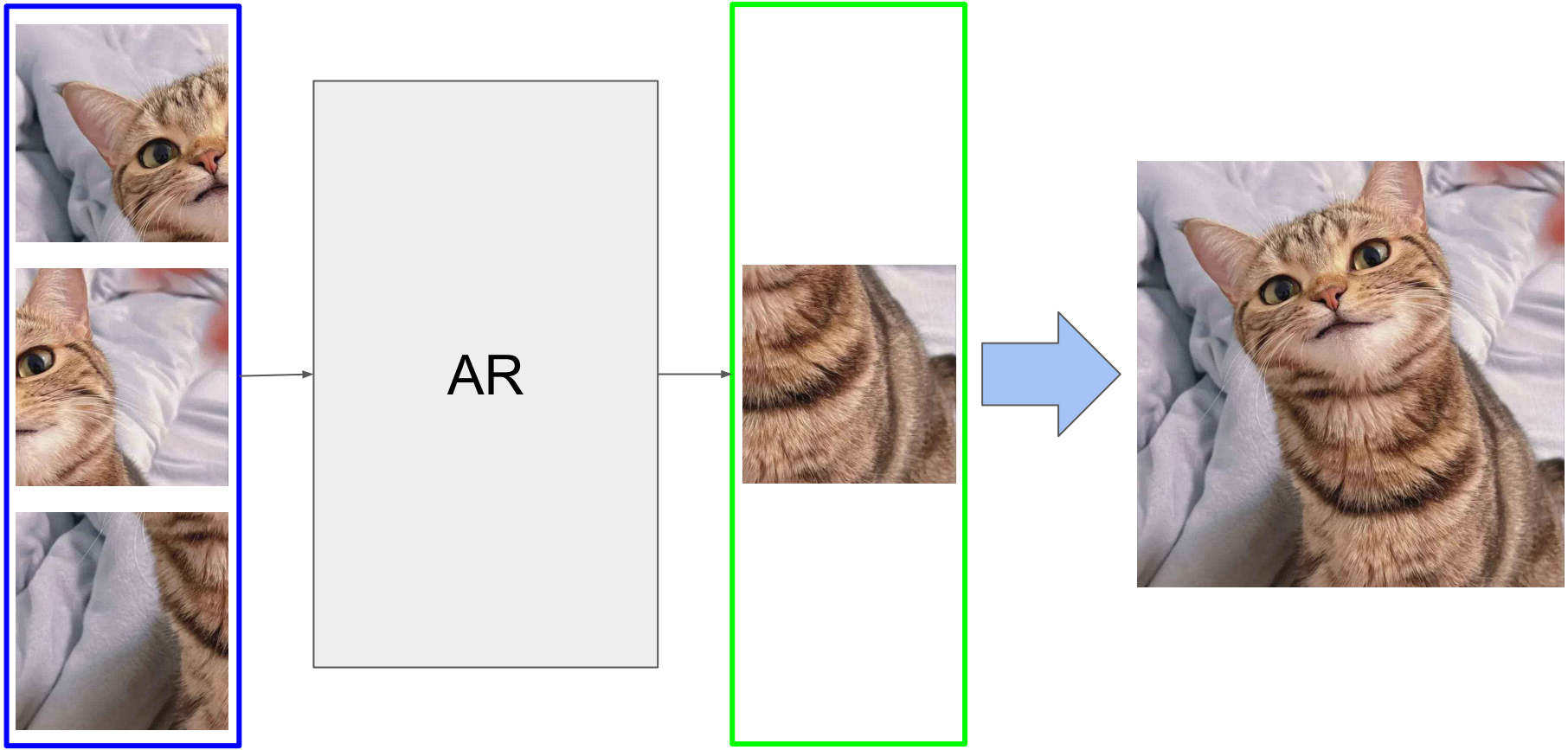
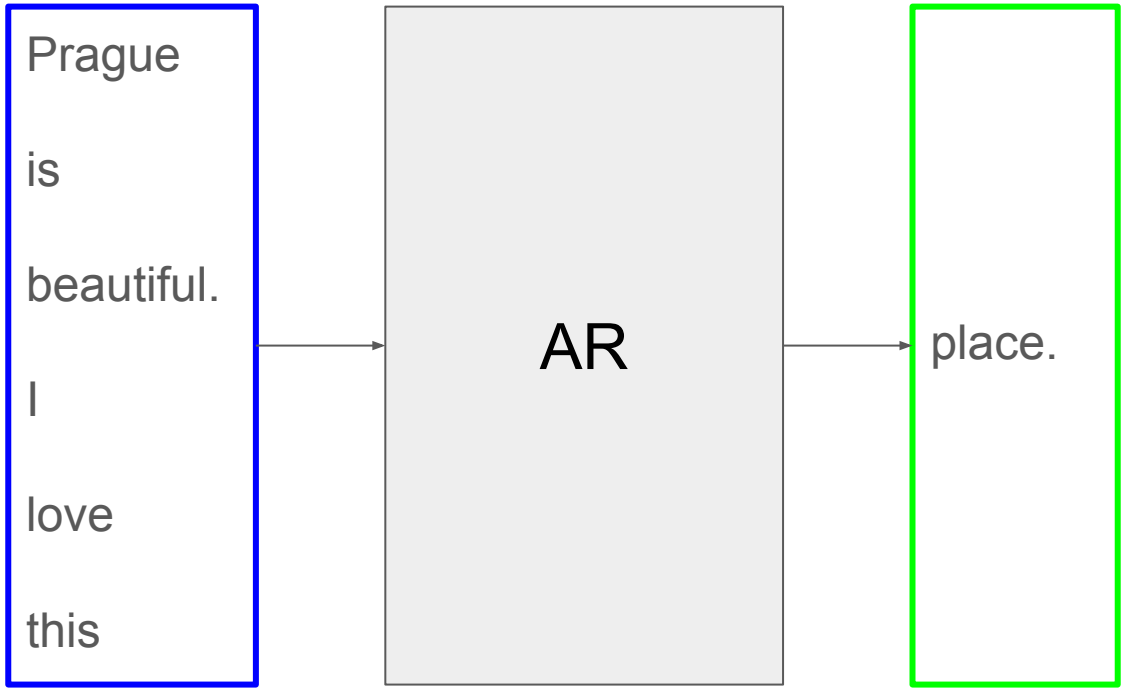


Image AutoRegressive Models (IARs)



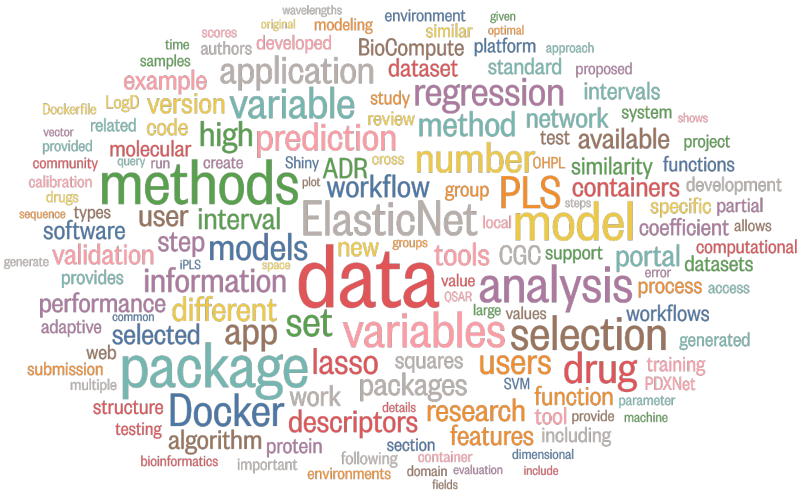
Problem: images are continuous



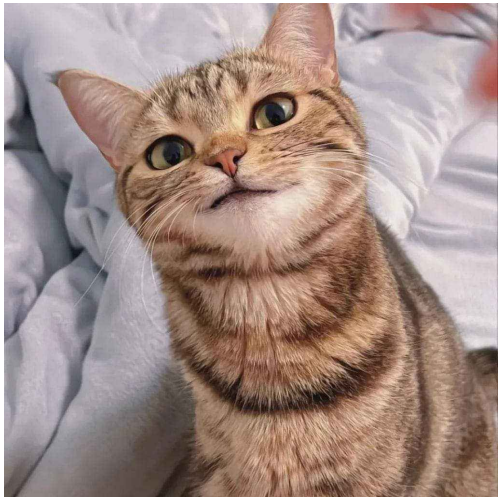
Problem: images are continuous

place.

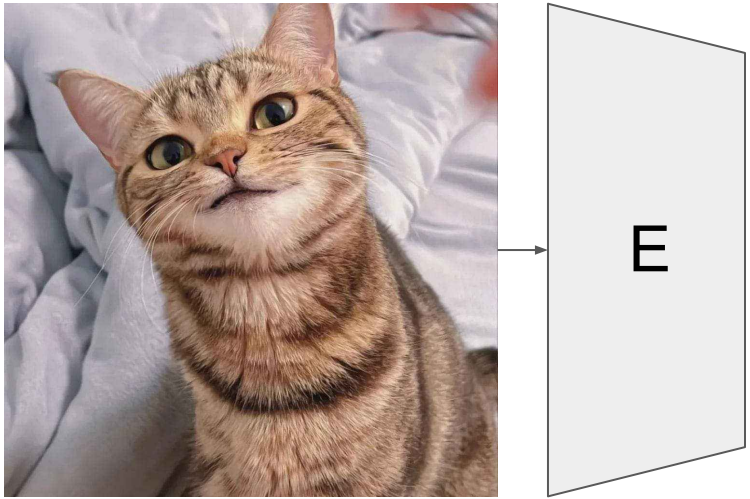
$\in$



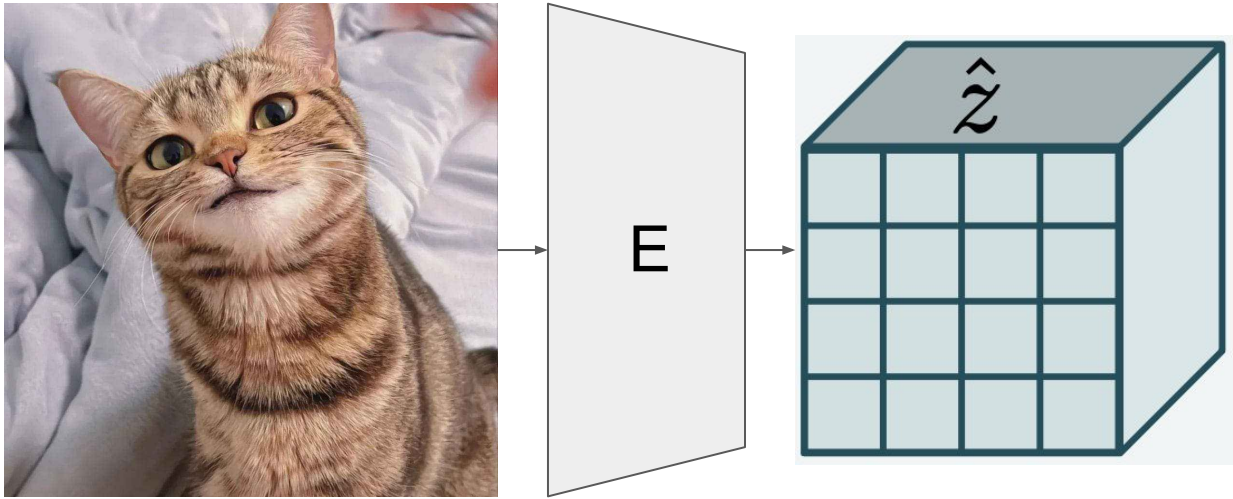
Solution: tokenize them!



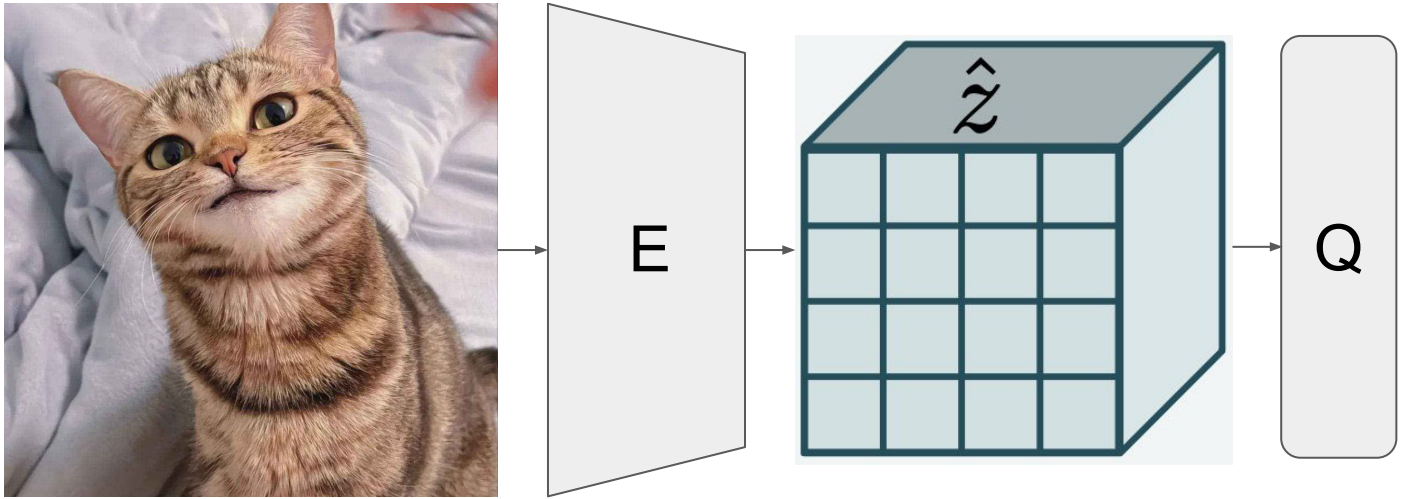
Solution: tokenize them!



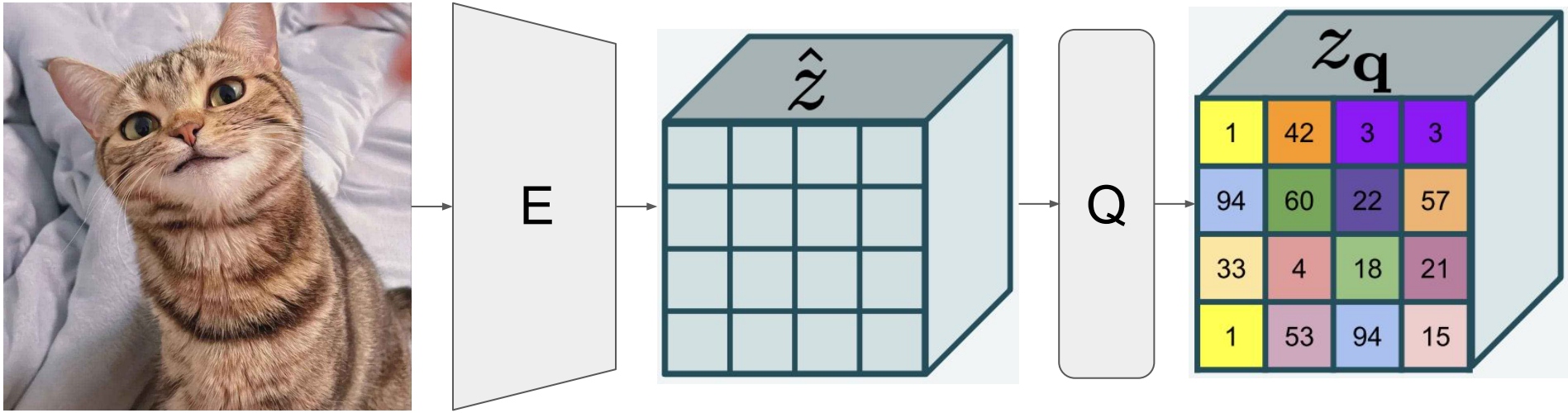
Solution: tokenize them!



Solution: tokenize them!

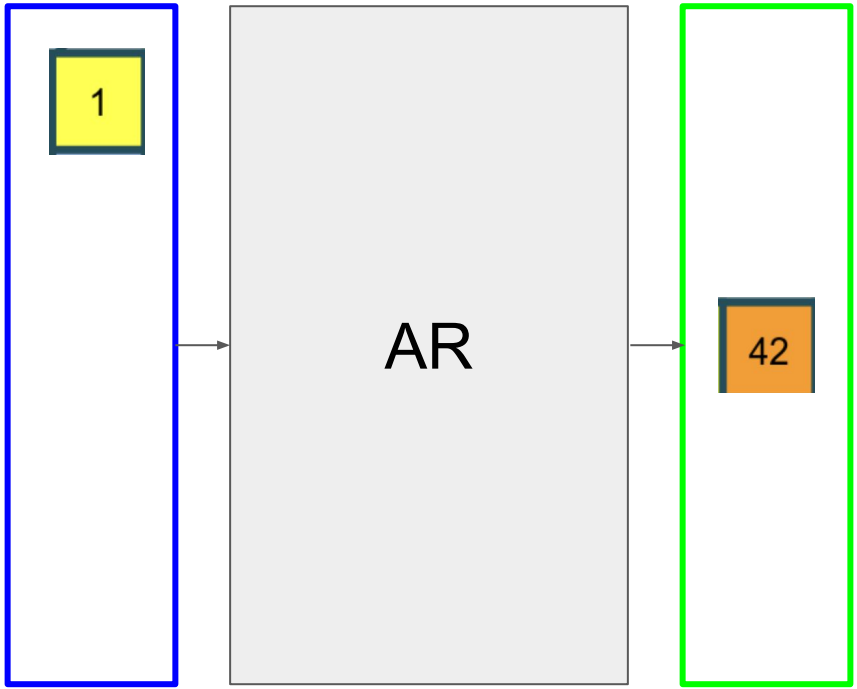


Solution: tokenize them!

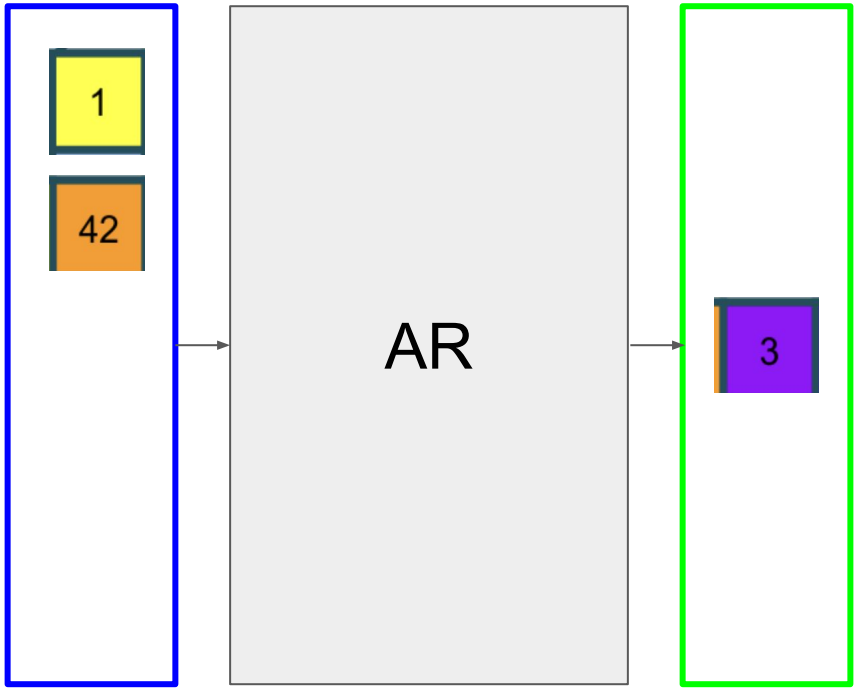


[Esser P., et al. 2021]

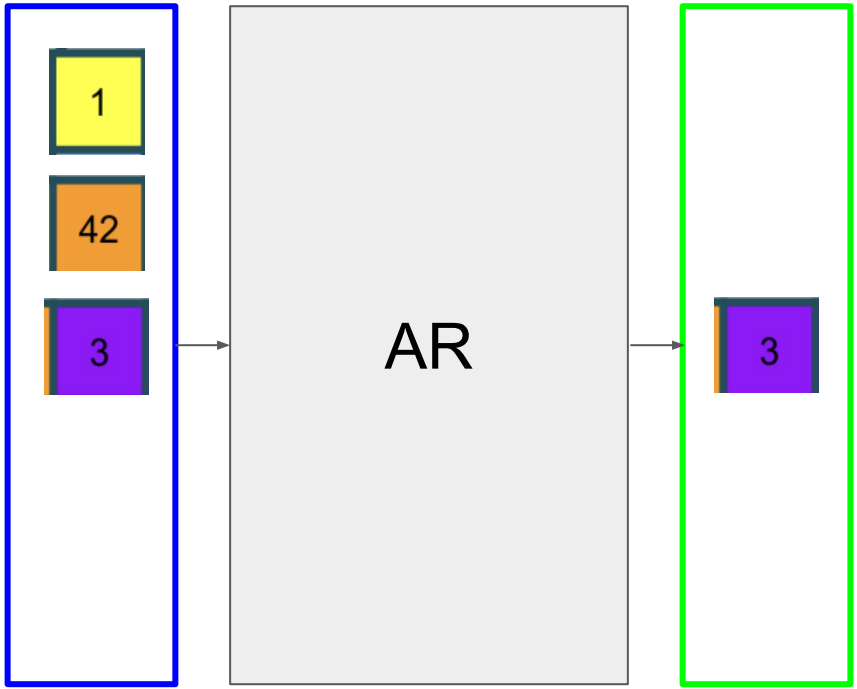
Generation with IAR



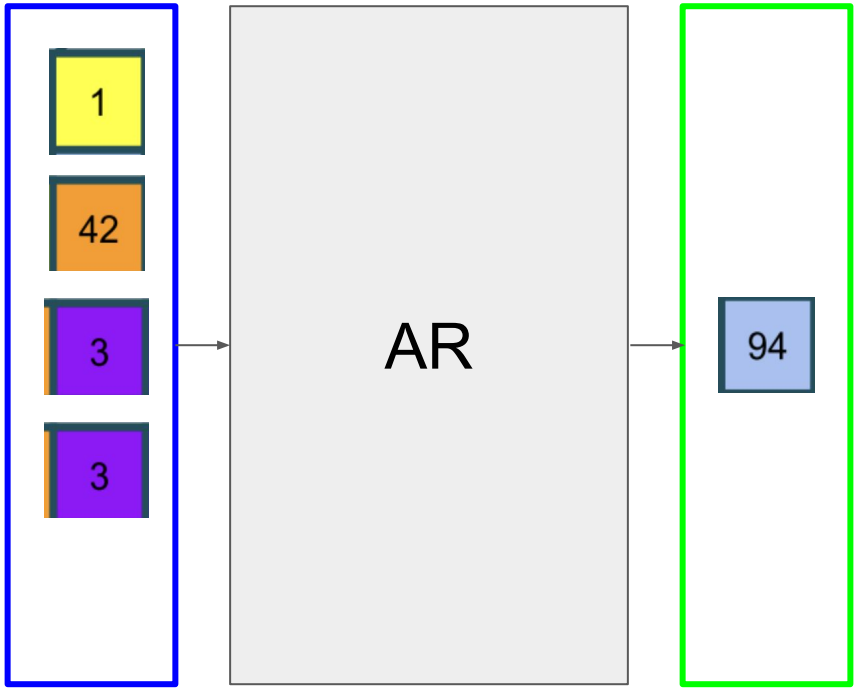
Generation with IAR



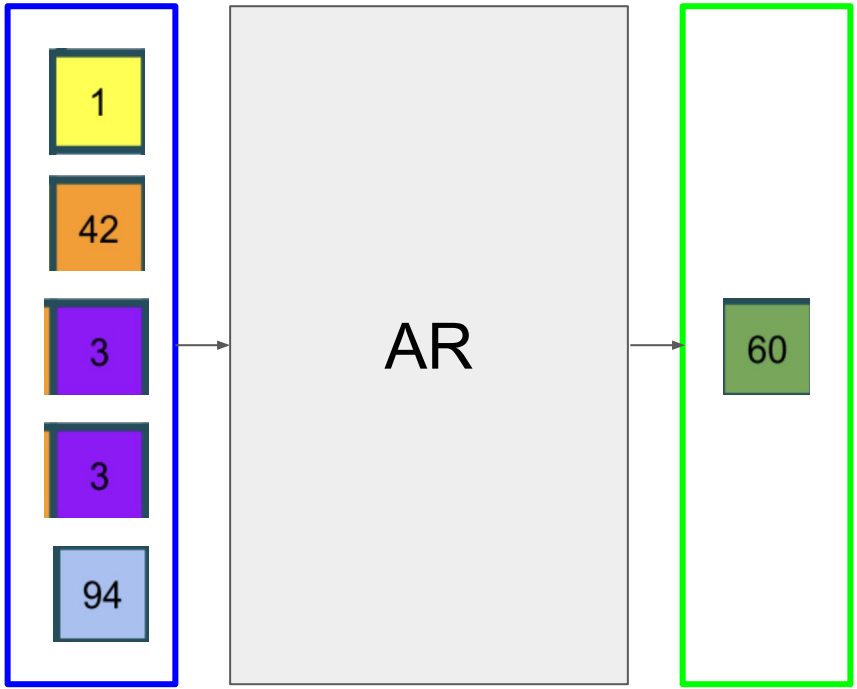
Generation with IAR



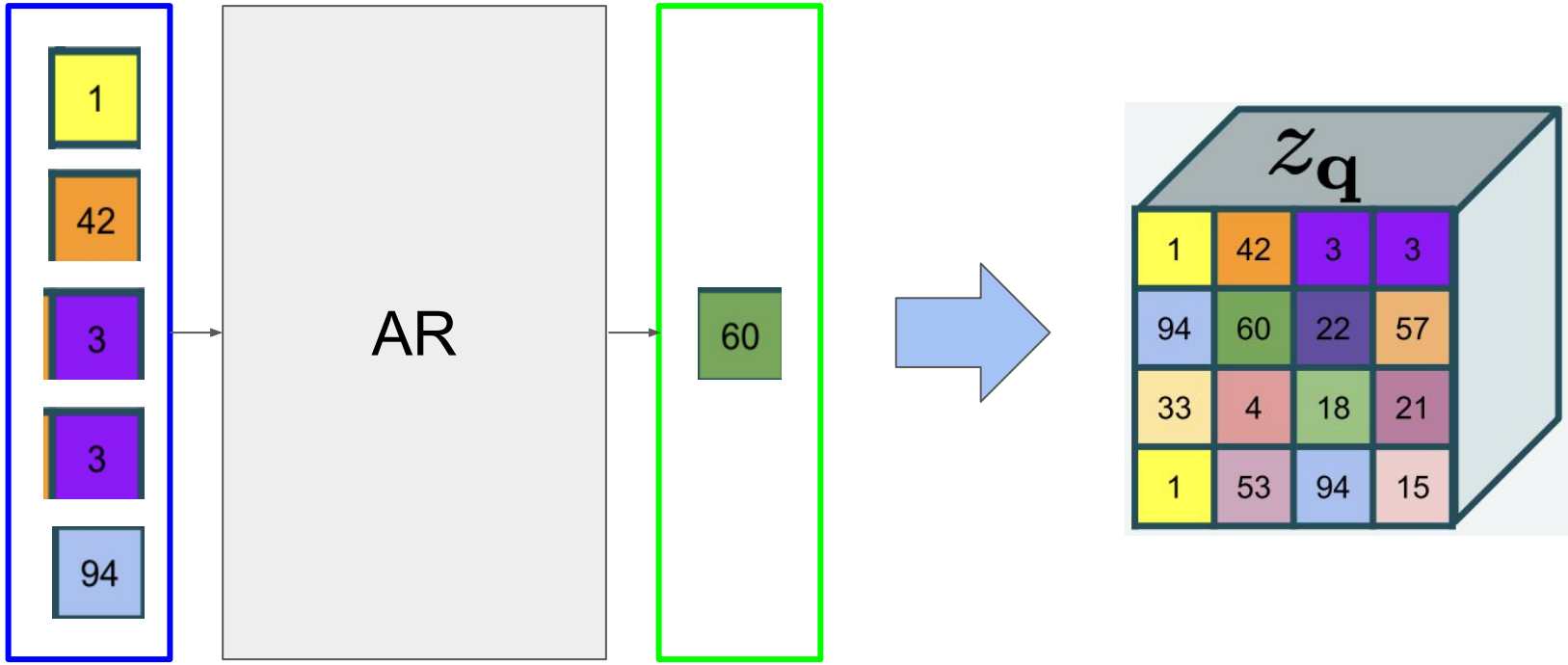
Generation with IAR



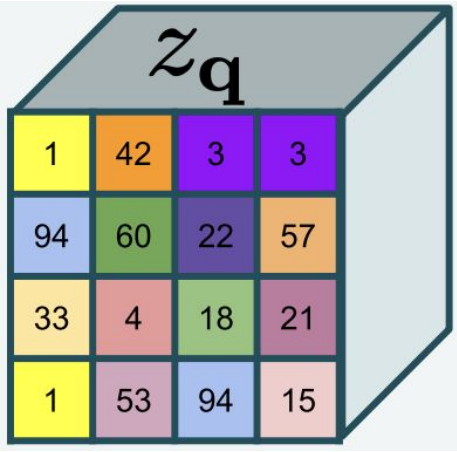
Generation with IAR



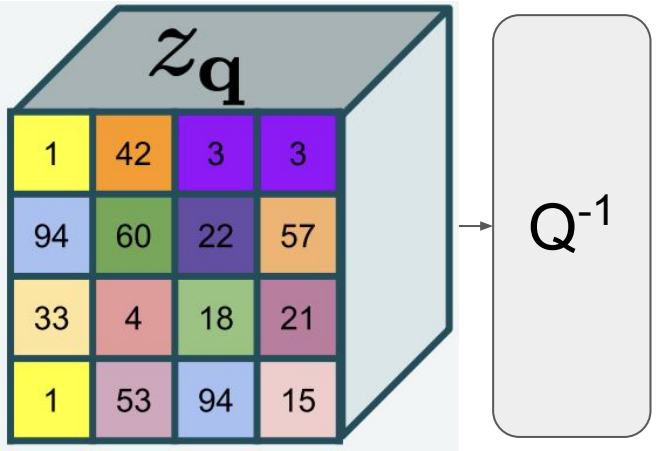
Generation with IAR



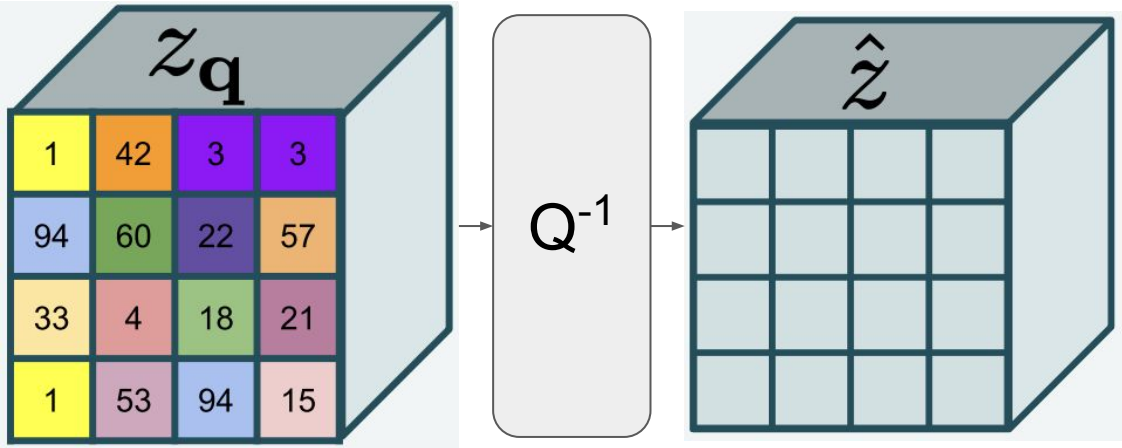
Decoding



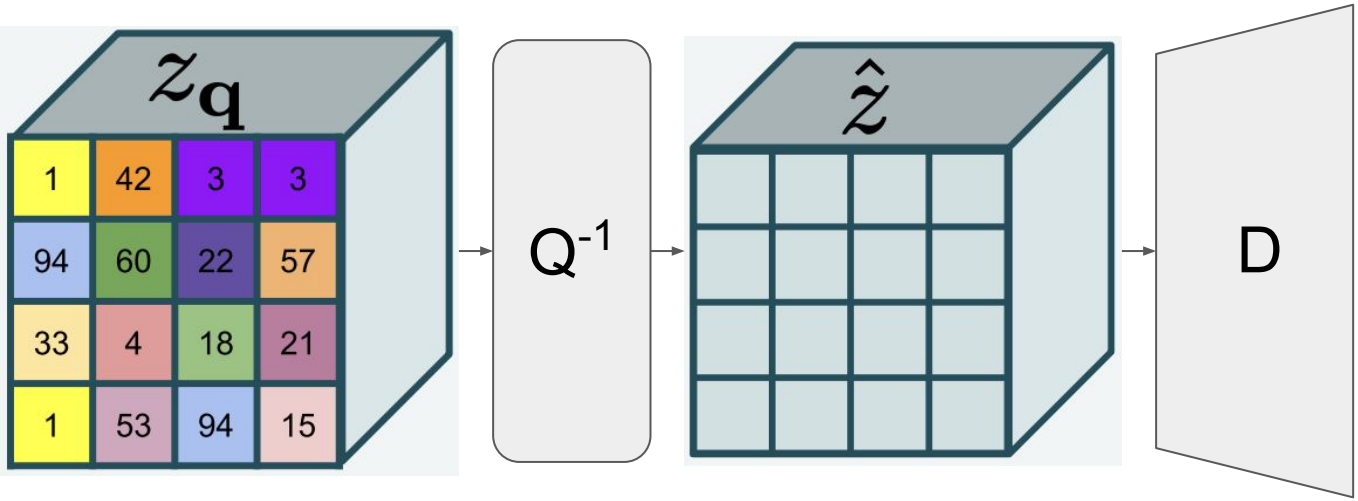
Decoding



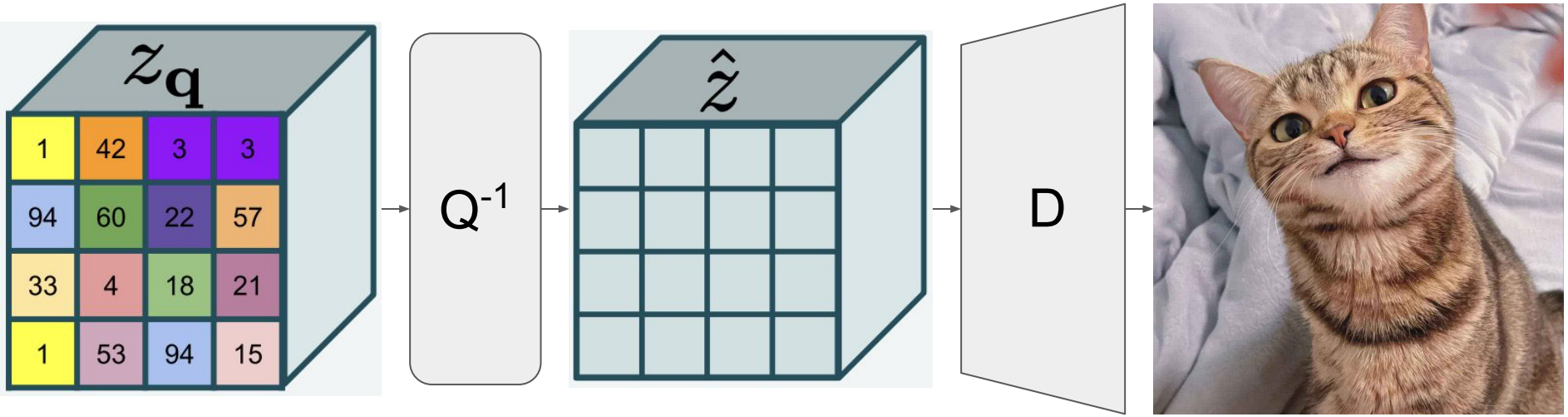
Decoding



Decoding



Decoding



Visual AutoRegressive Model (VAR)

NEURAL INFORMATION PROCESSING SYSTEMS

2024

(Oral presentation)

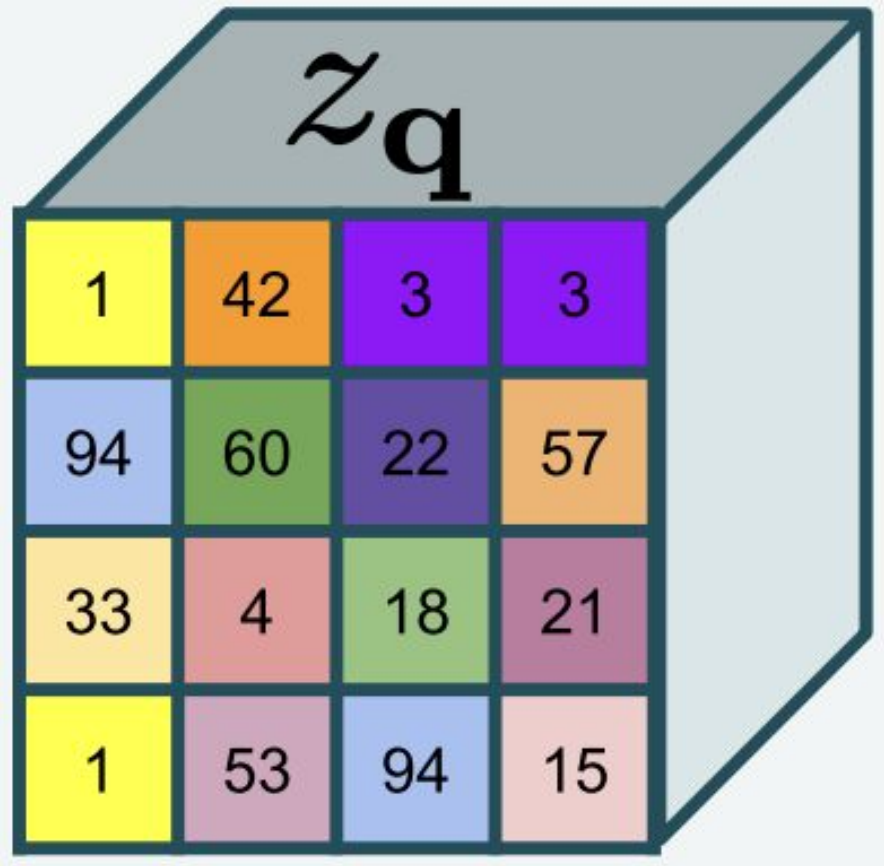
[Tian K., et al. 2024]

Next-token prediction is costly

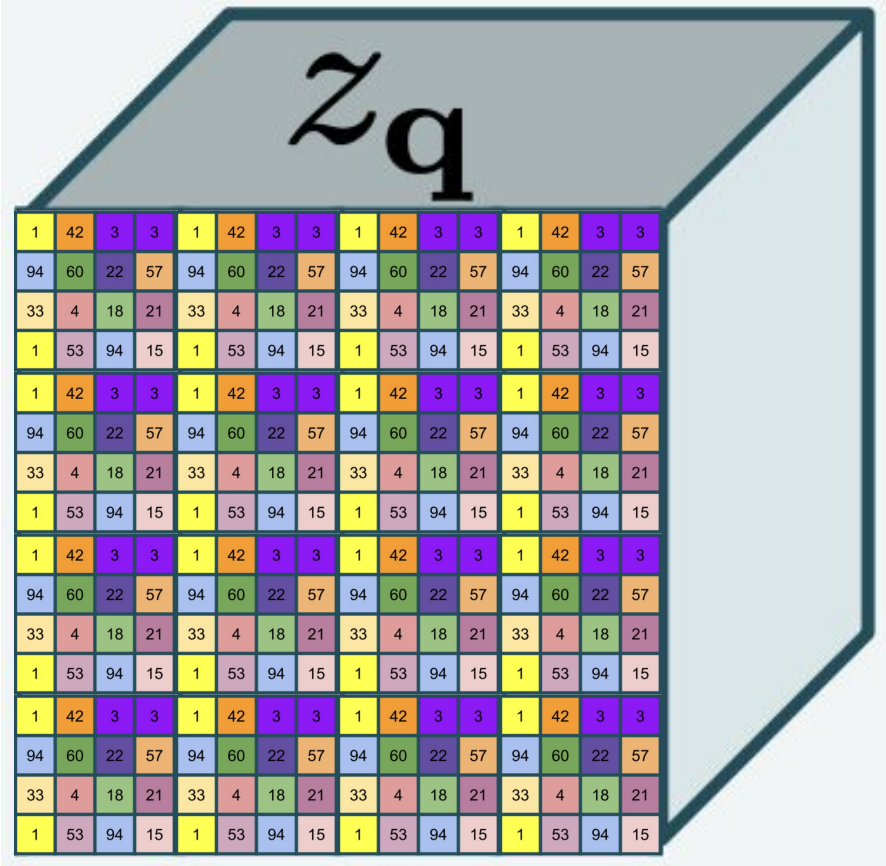
(b) AR: Image generation by **next-image-token** prediction

[Tian K., et al. 2024]

Next-token prediction is costly



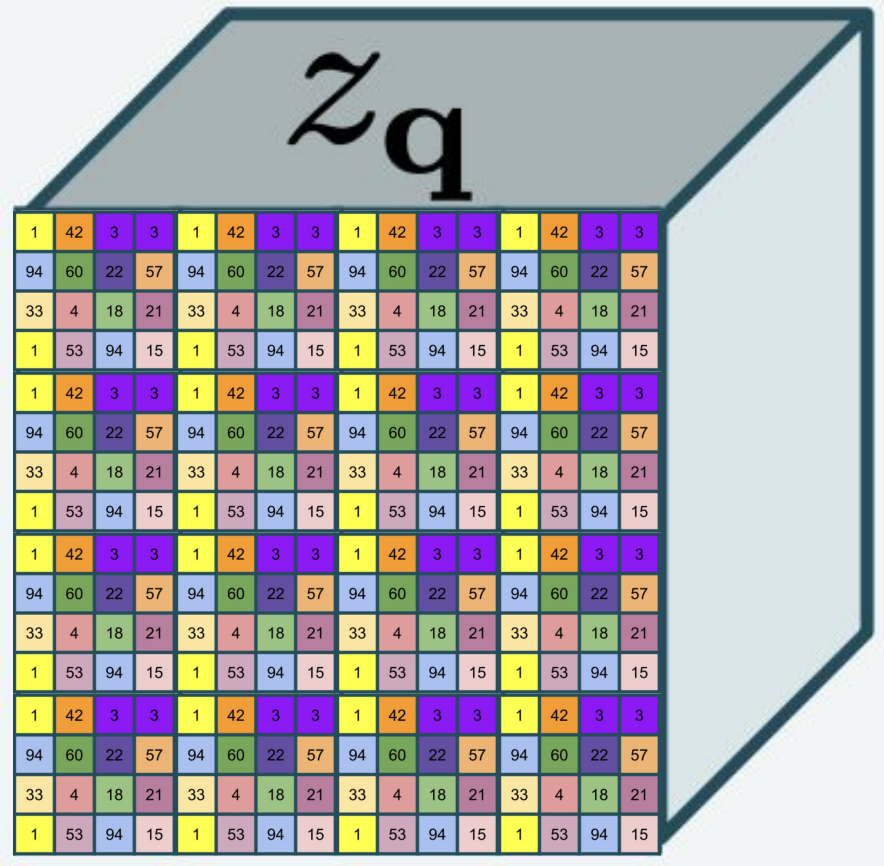
Next-token prediction is costly



1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15

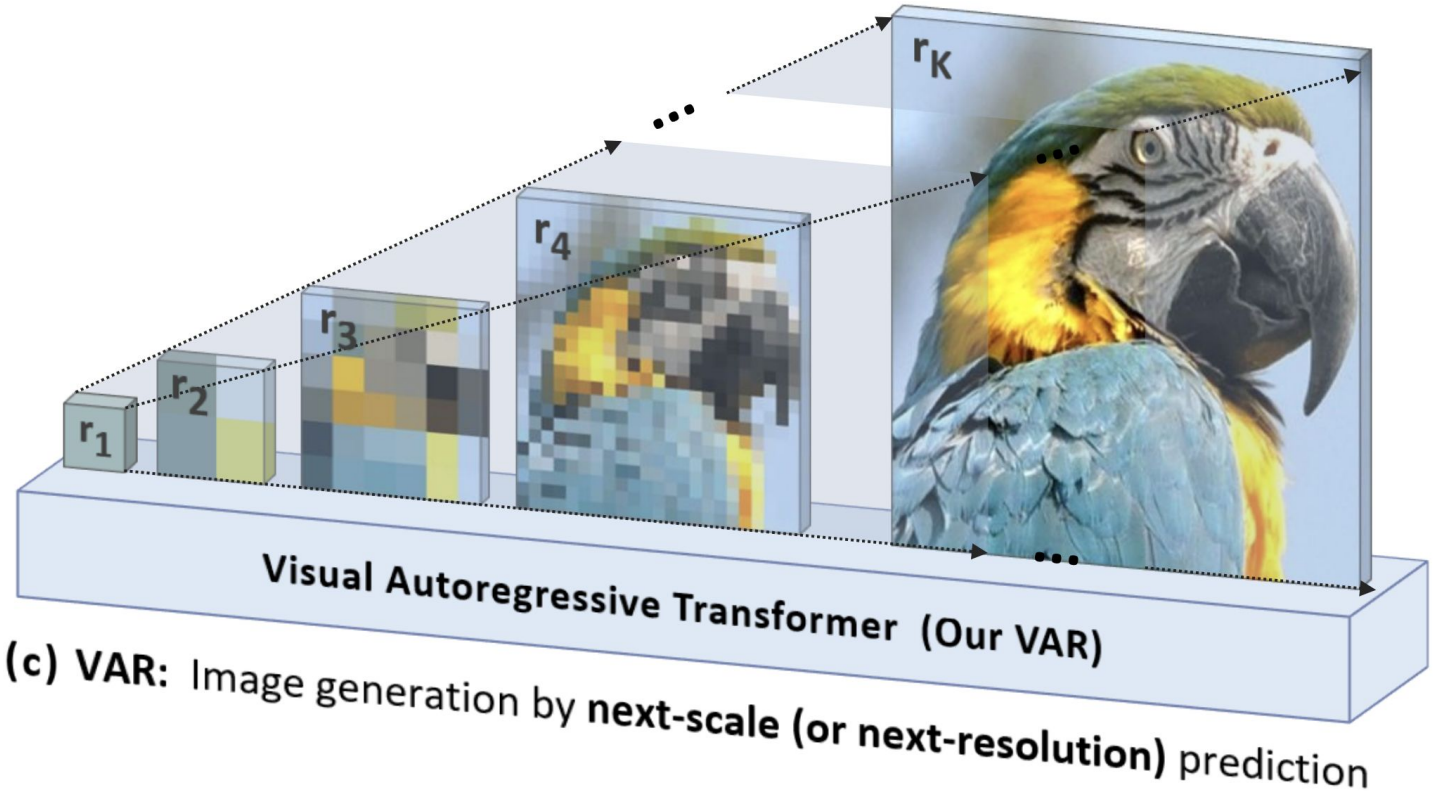
Next-token prediction is costly

16x16 = 256



1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15

VAR: next-scale prediction

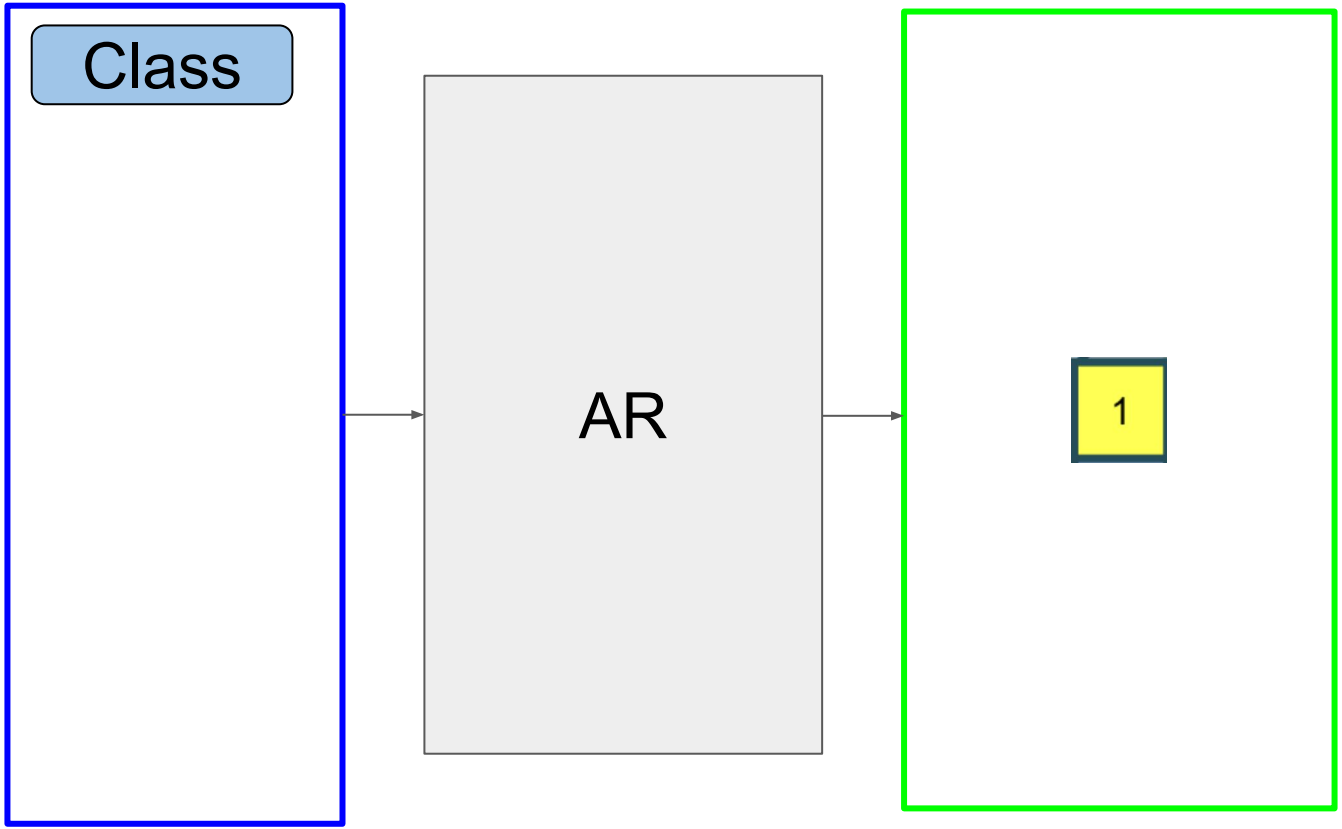


(c) VAR: Image generation by **next-scale (or next-resolution)** prediction

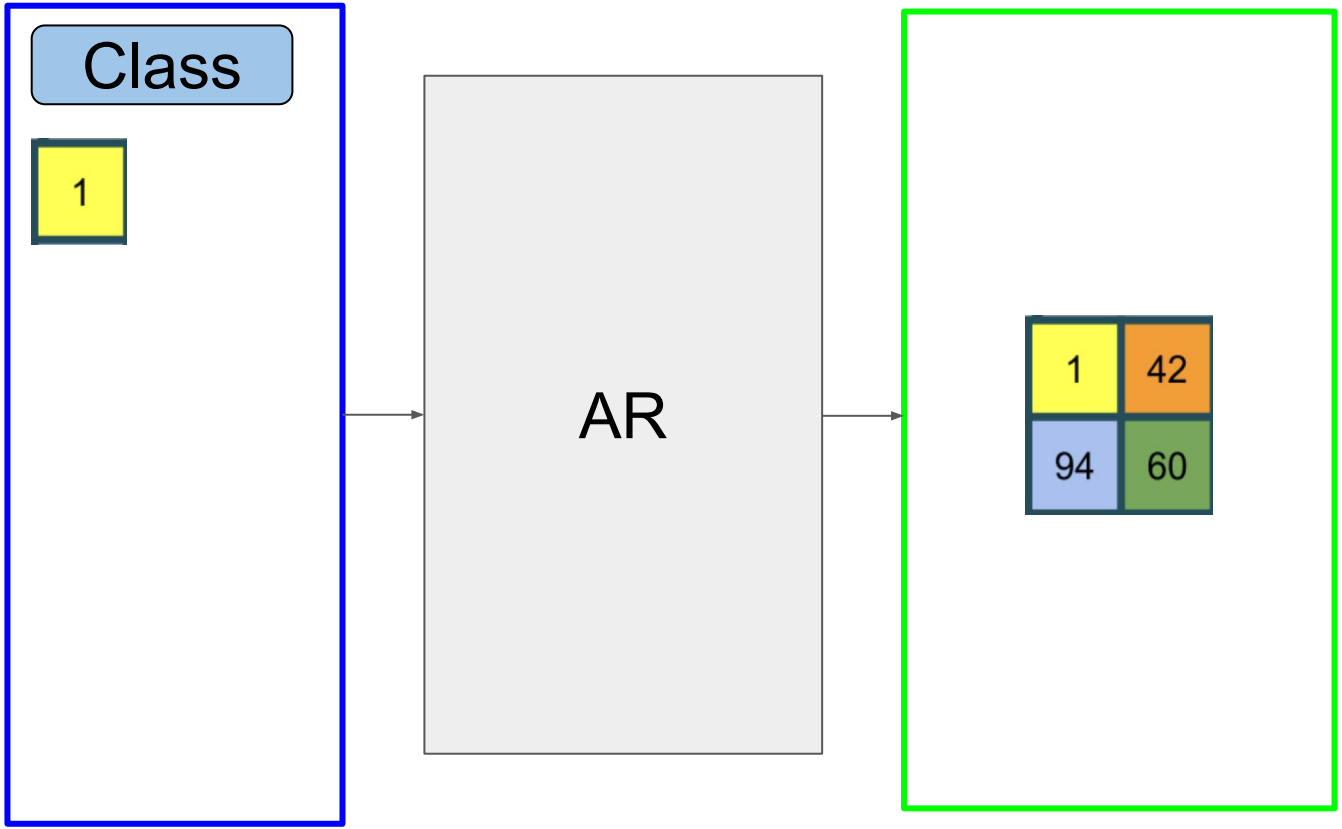
[Tian K., et al. 2024]

VAR: next-scale encoding

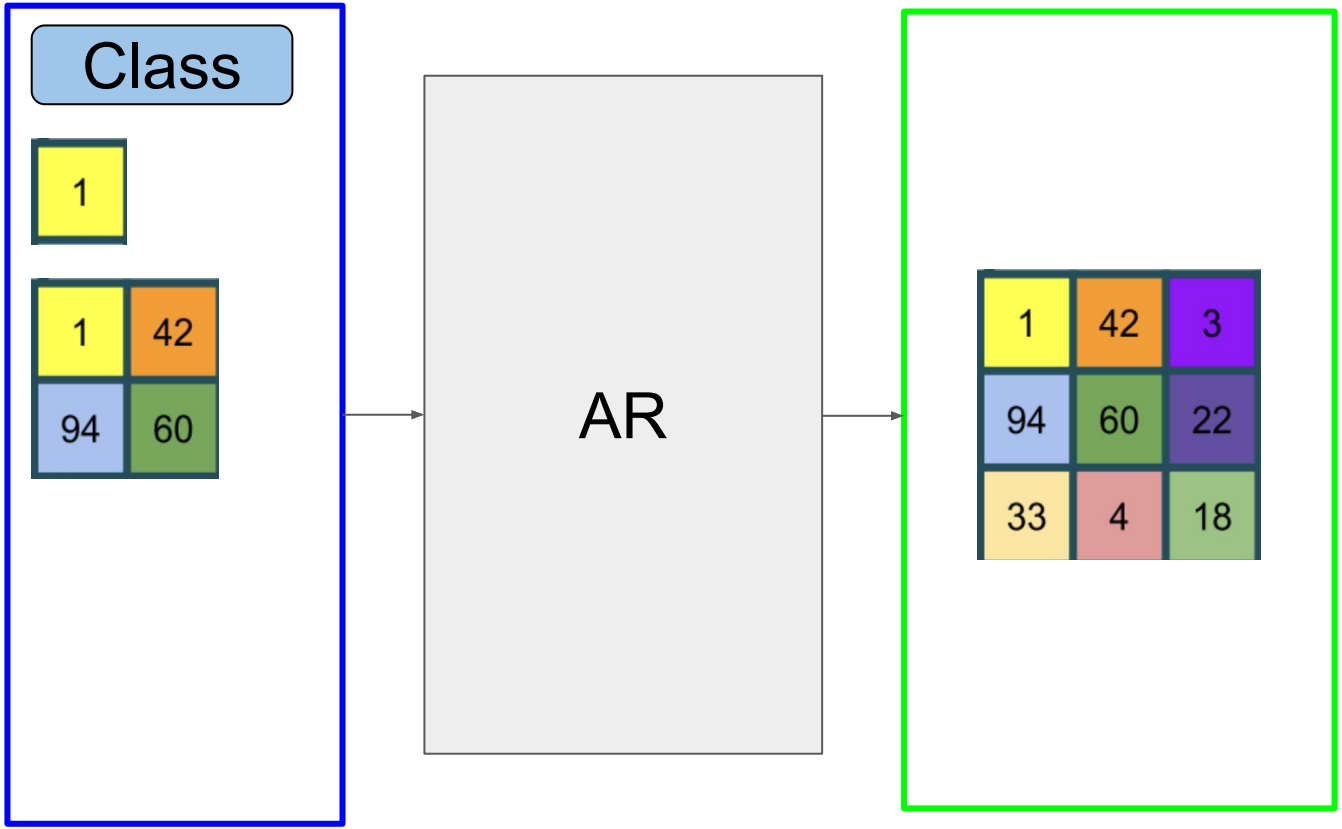
VAR: generation



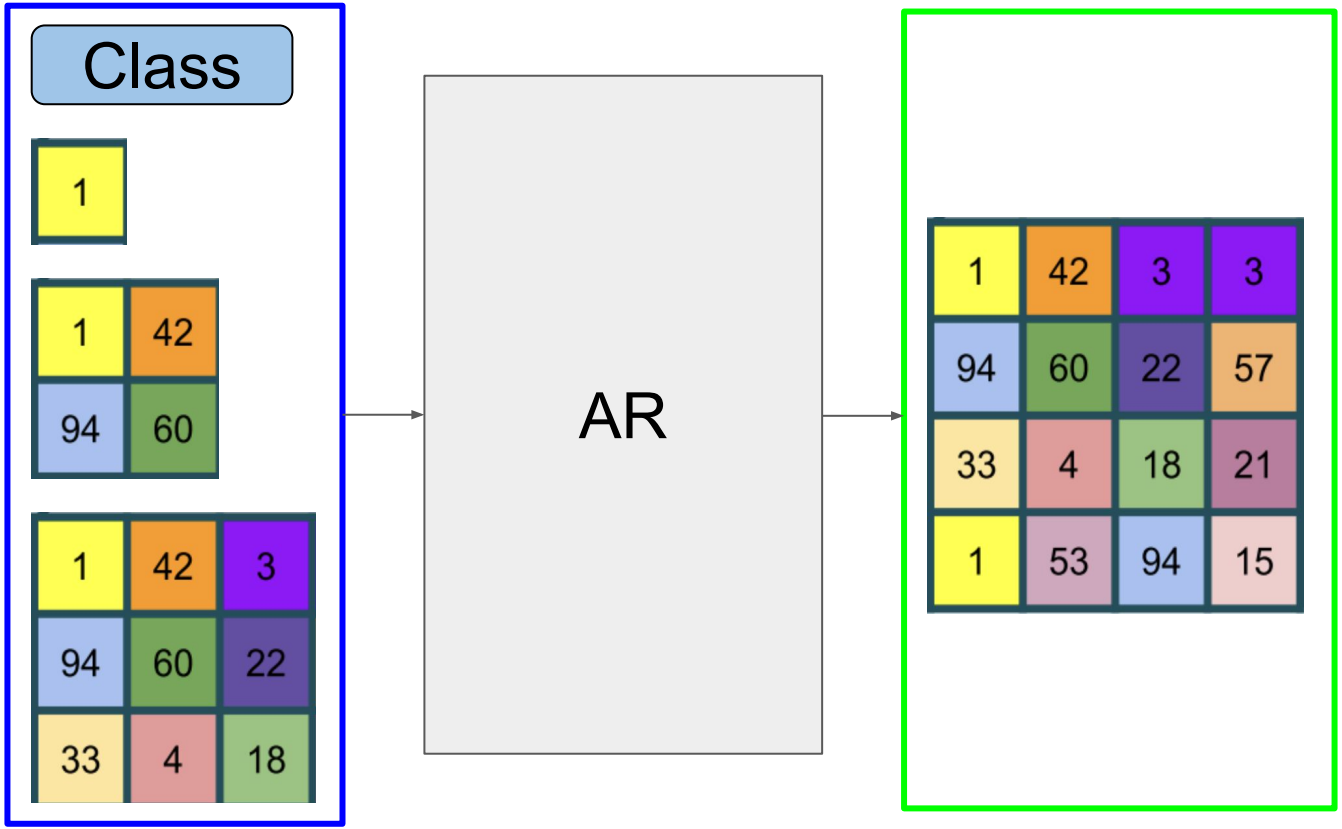
VAR: generation



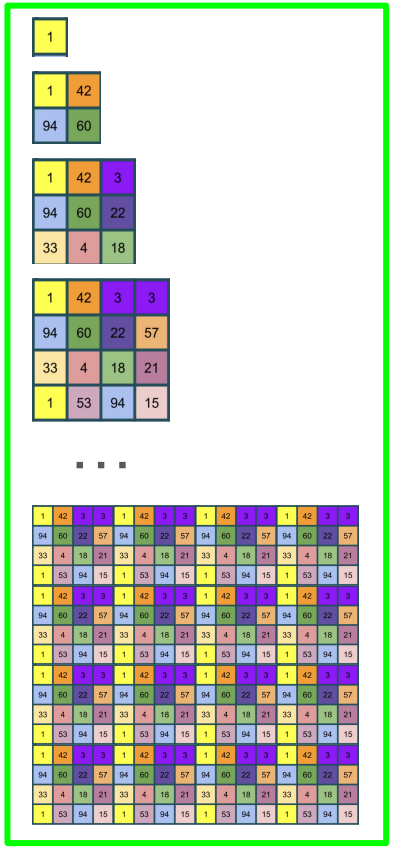
VAR: generation



VAR: generation



VAR: Decoding



1

1	42
94	60

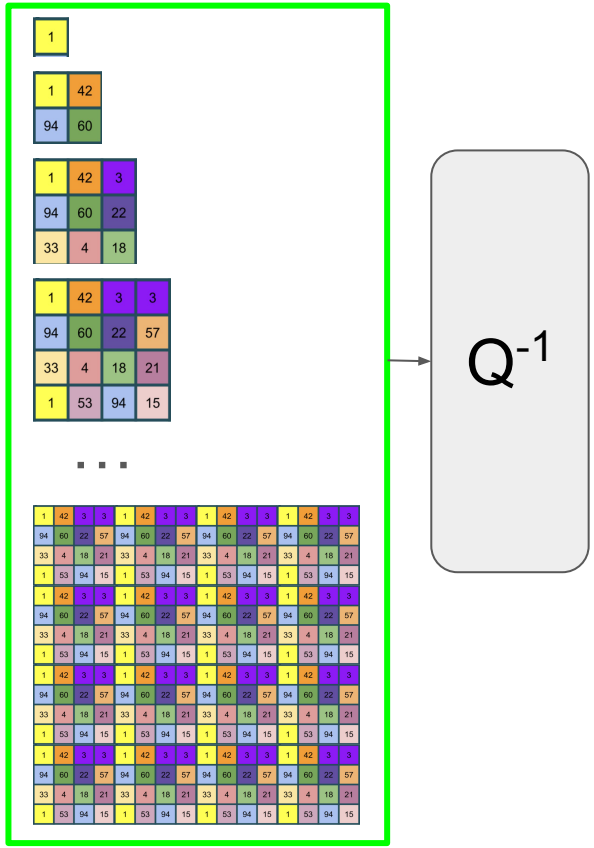
1	42	3
94	60	22
33	4	18

1	42	3	3
94	60	22	57
33	4	18	21
1	53	94	15

...

1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15

VAR: Decoding



1

1	42
94	60

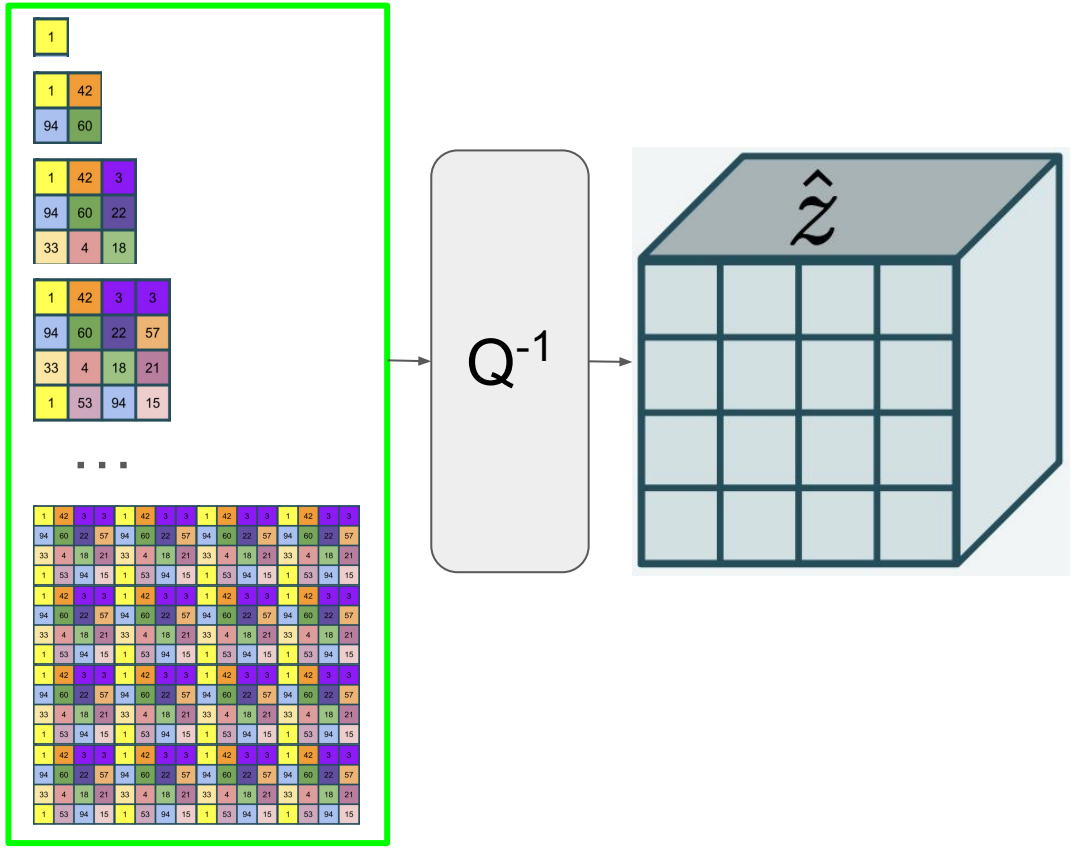
1	42	3
94	60	22
33	4	18

1	42	3	3
94	60	22	57
33	4	18	21
1	53	94	15

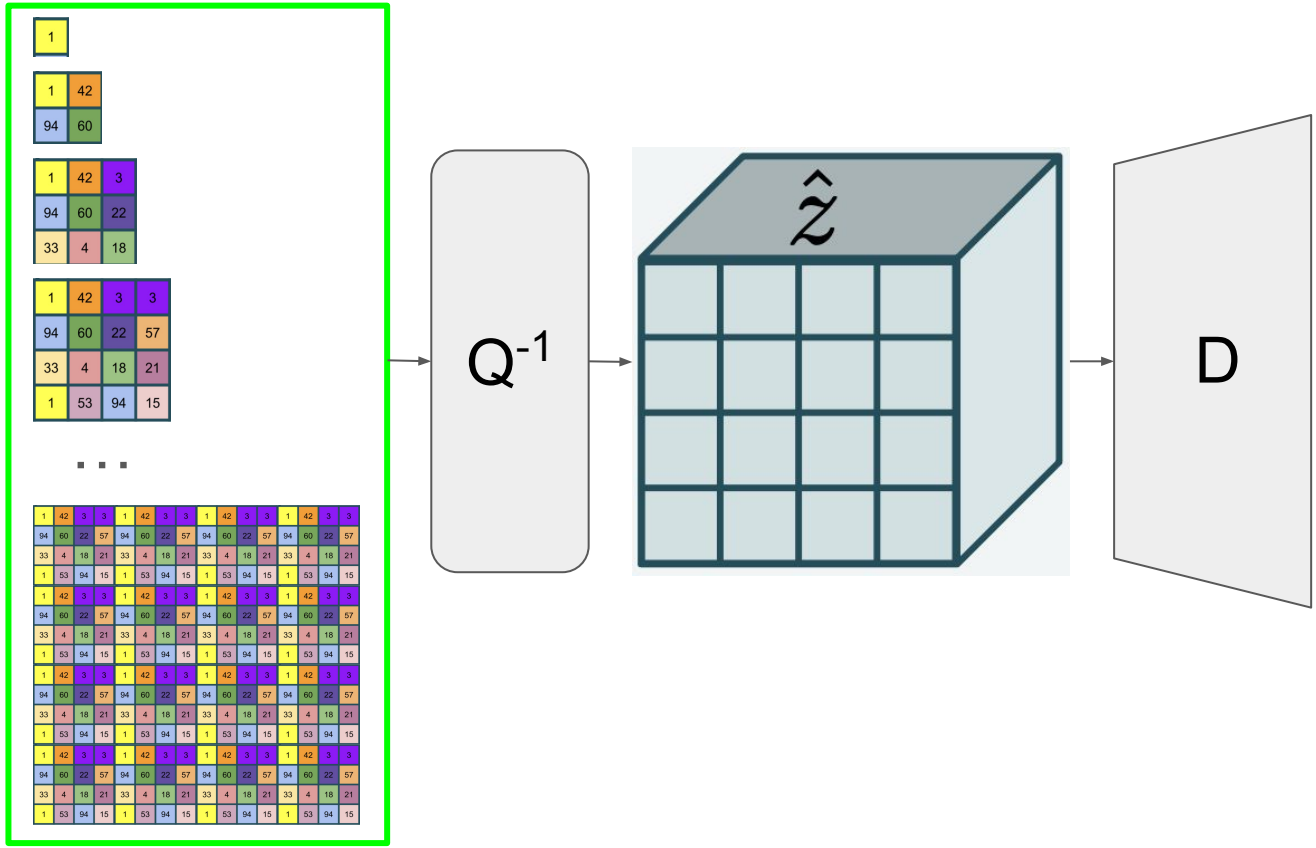
. . .

1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15
1	42	3	3	1	42	3	3	1	42	3	3	1	42	3	3
94	60	22	57	94	60	22	57	94	60	22	57	94	60	22	57
33	4	18	21	33	4	18	21	33	4	18	21	33	4	18	21
1	53	94	15	1	53	94	15	1	53	94	15	1	53	94	15

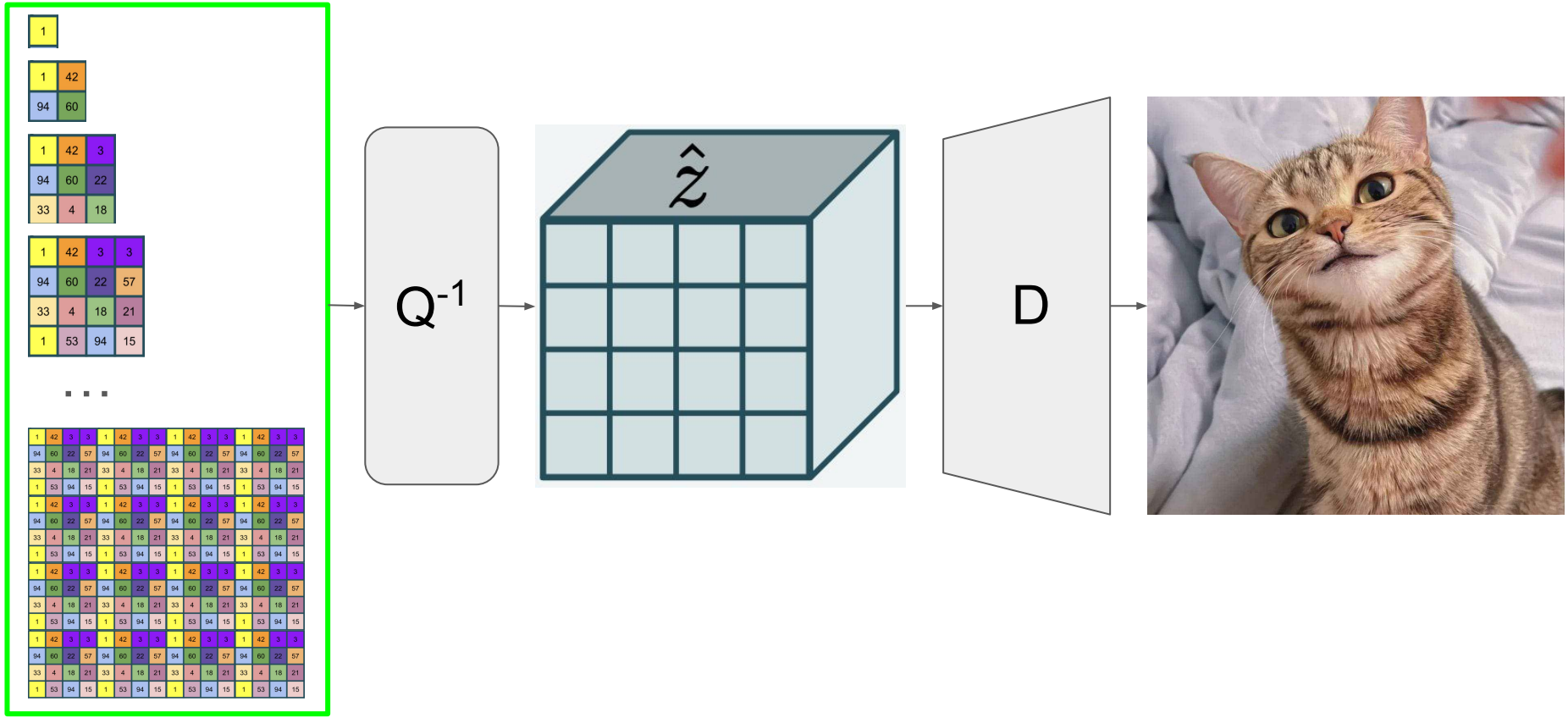
VAR: Decoding



VAR: Decoding

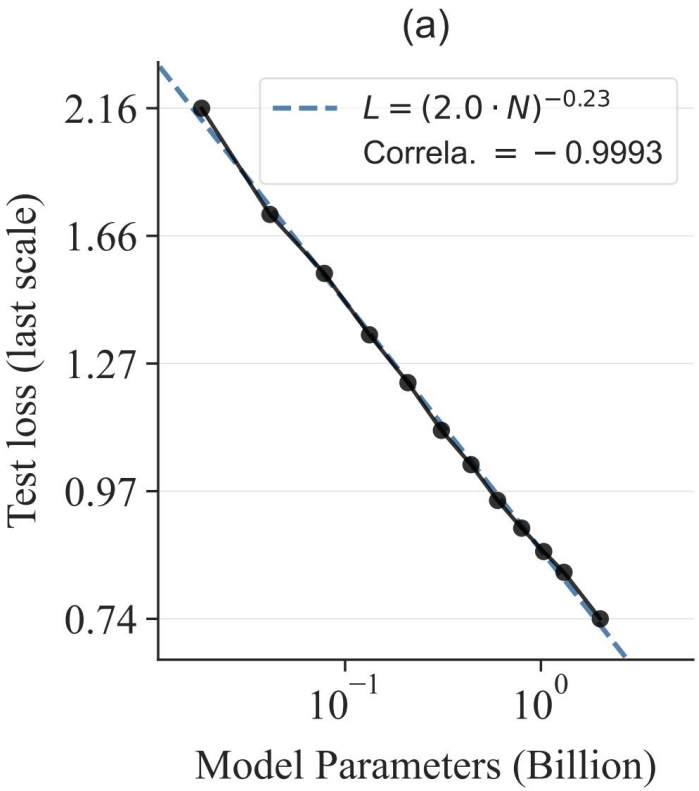


VAR: Decoding



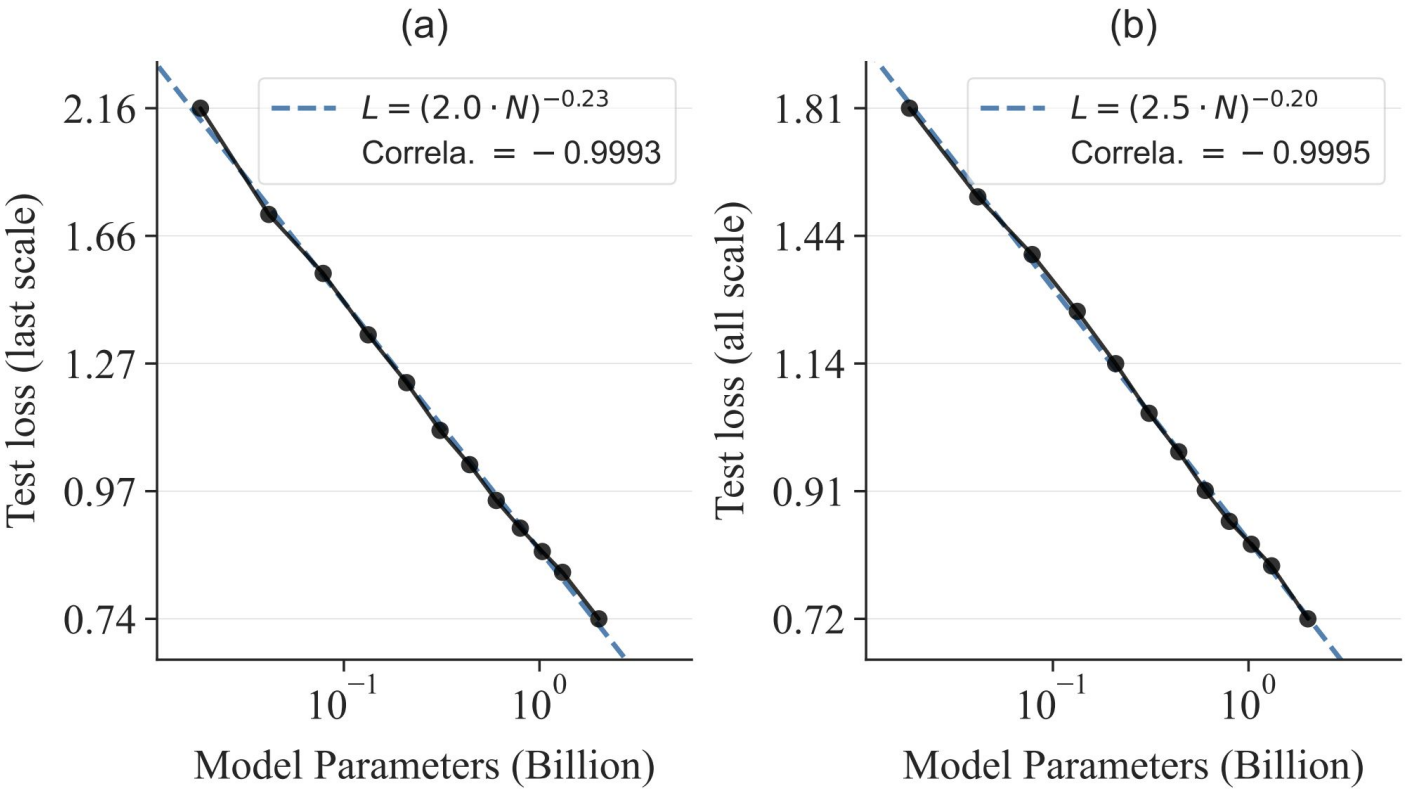
VAR: generation

VAR scales well



[Tian K., et al. 2024]

VAR scales well



[Tian K., et al. 2024]

VAR has a great performance

Model	Model Type	FID

[Tian K., et al. 2024]

VAR has a great performance

Model	Model Type	FID
Previous SOTA	DM	2.10
Previous SOTA	IAR	3.80

[Tian K., et al. 2024]

VAR has a great performance

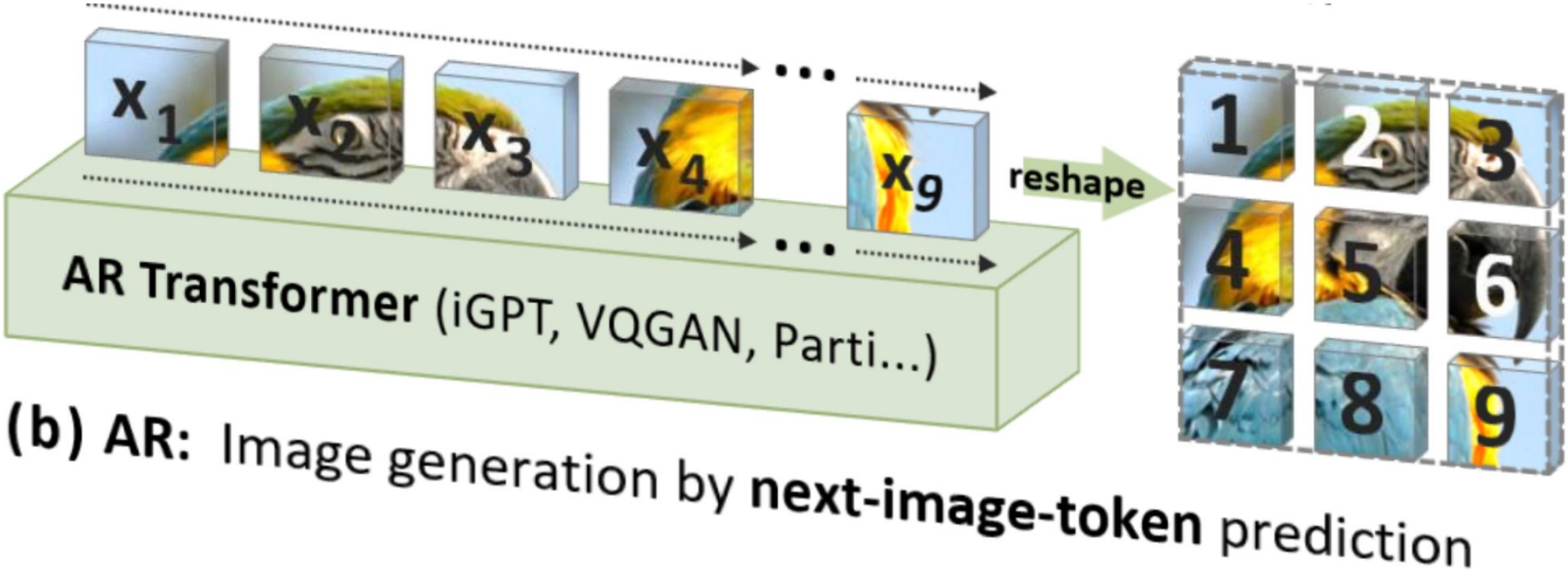
Model	Model Type	FID
Previous SOTA	DM	2.10
Previous SOTA	IAR	3.80
VAR-d16	IAR	3.30
VAR-d20	IAR	2.57
VAR-d24	IAR	2.09
VAR-d30	IAR	**1.92**

[Tian K., et al. 2024]

Randomized AutoRegressive Model (RAR)

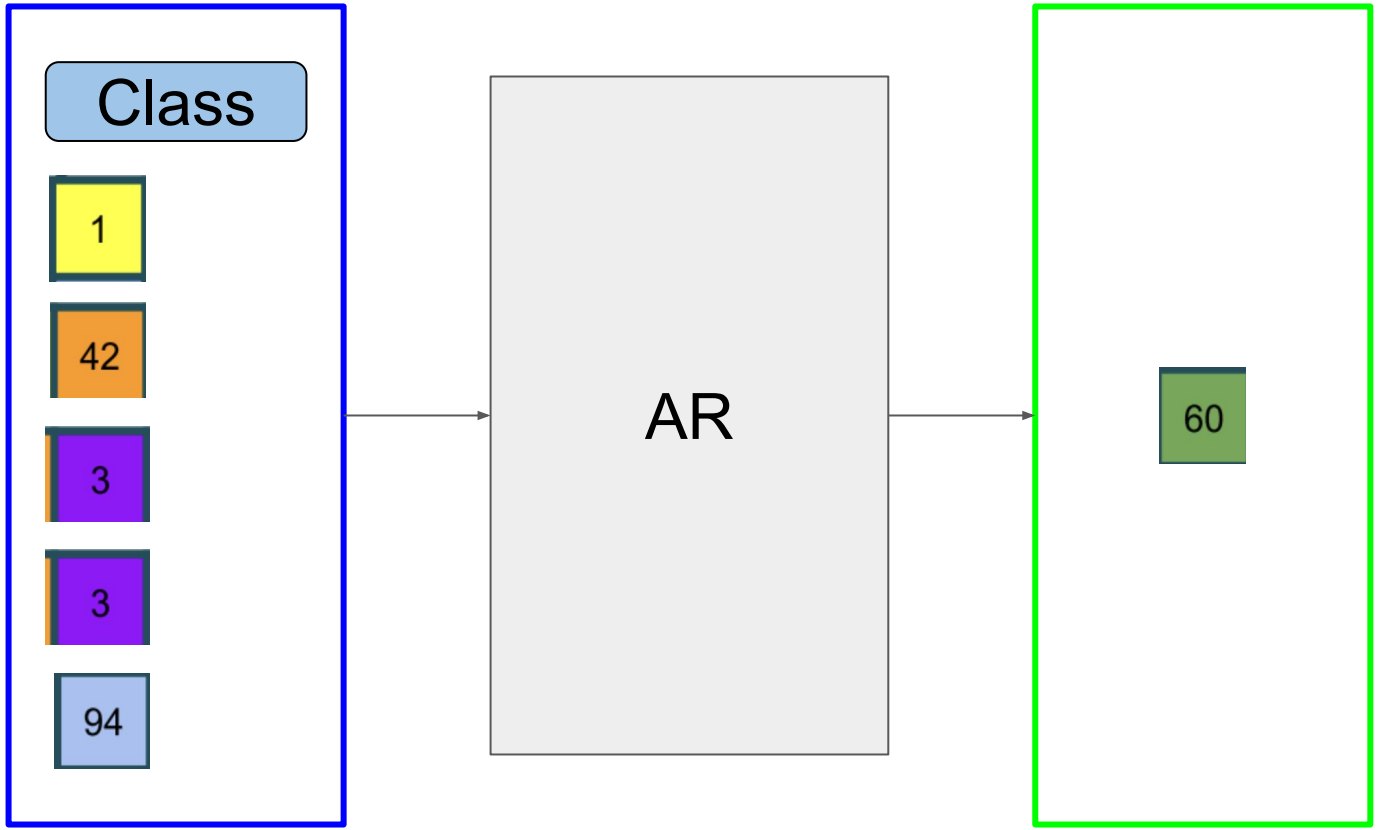
[Yu Q., et al. 2024]

Randomized AutoRegressive Model (RAR)



[Tian K., et al. 2024]

RAR: generation



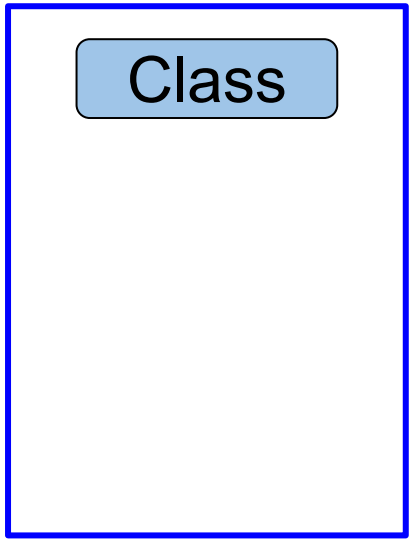
RAR scales well & has a great performance

Model	FID
VAR-d30	1.92

RAR scales well & has a great performance

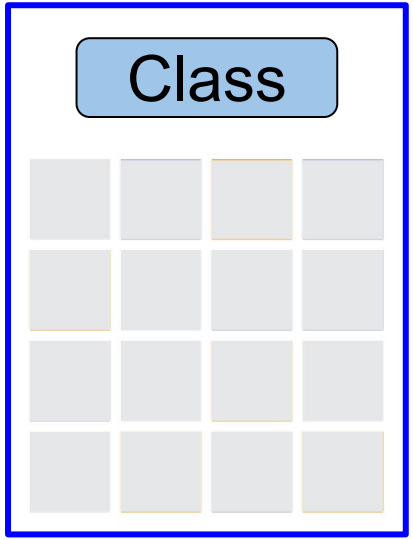
Model	FID
VAR-d30	1.92
RAR-B	1.95
RAR-L	1.70
RAR-XL	1.50
RAR-XXL	**1.48**

Masked AutoRegressive Modeling (MAR)

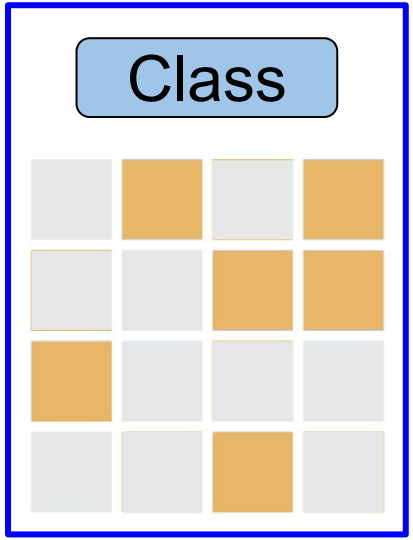


[Li T., et al. 2024]

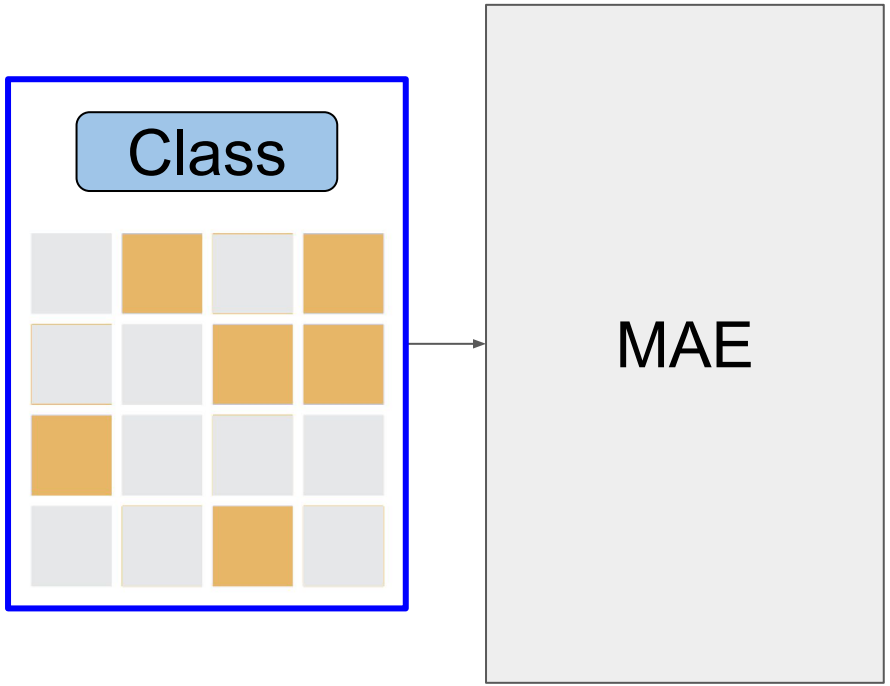
Masked AutoRegressive Modeling (MAR)



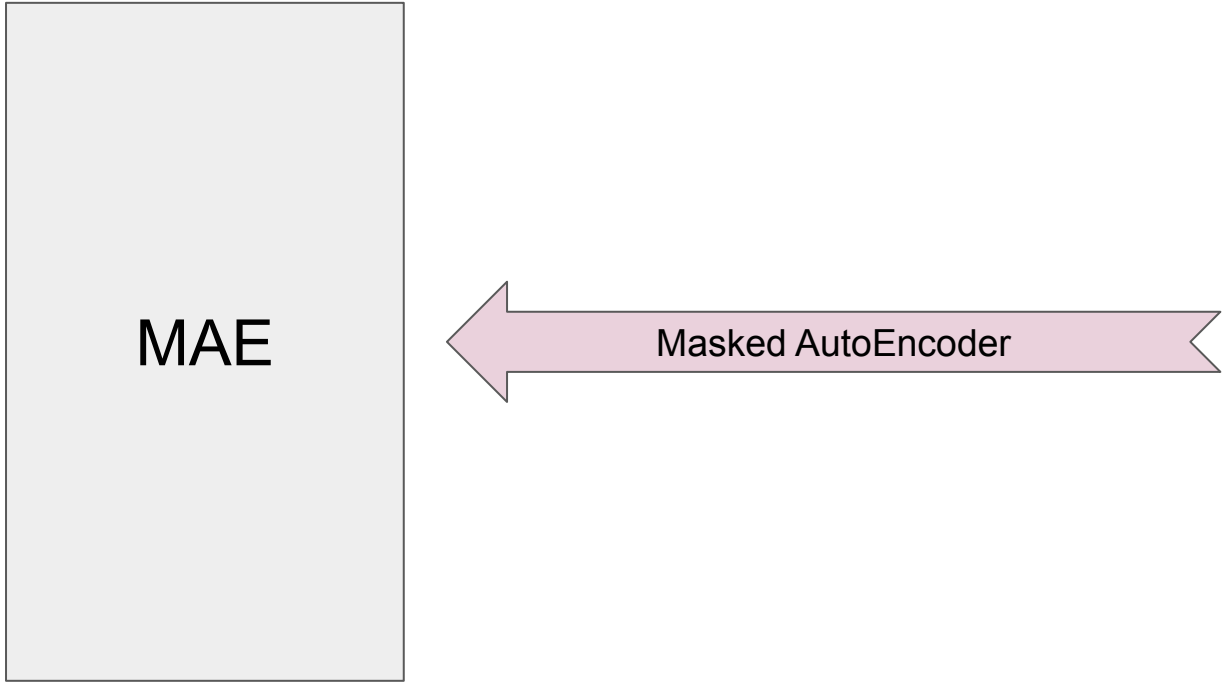
Masked AutoRegressive Modeling (MAR)



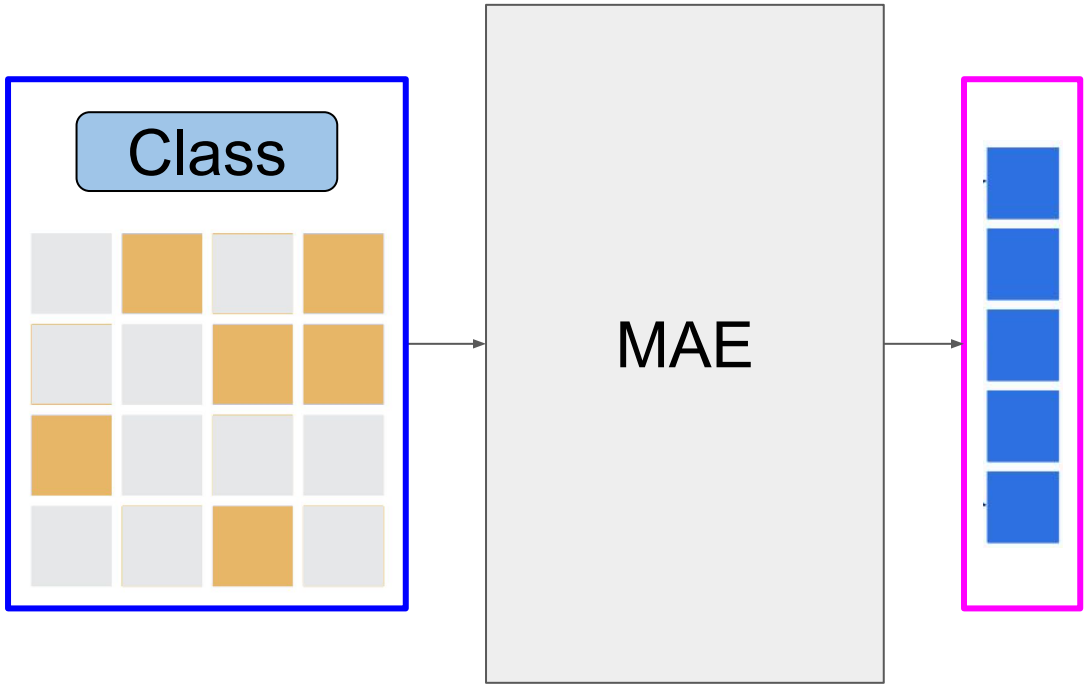
Masked AutoRegressive Modeling (MAR)



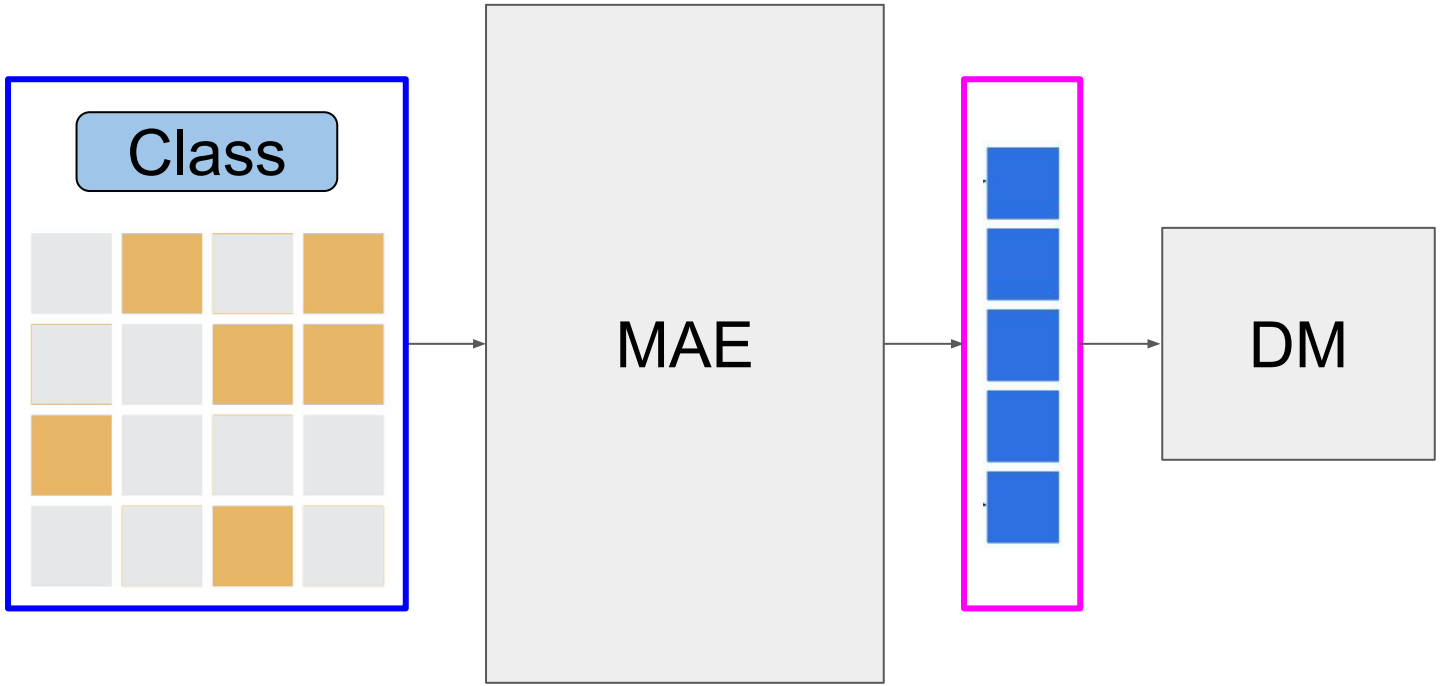
Masked AutoRegressive Modeling (MAR)



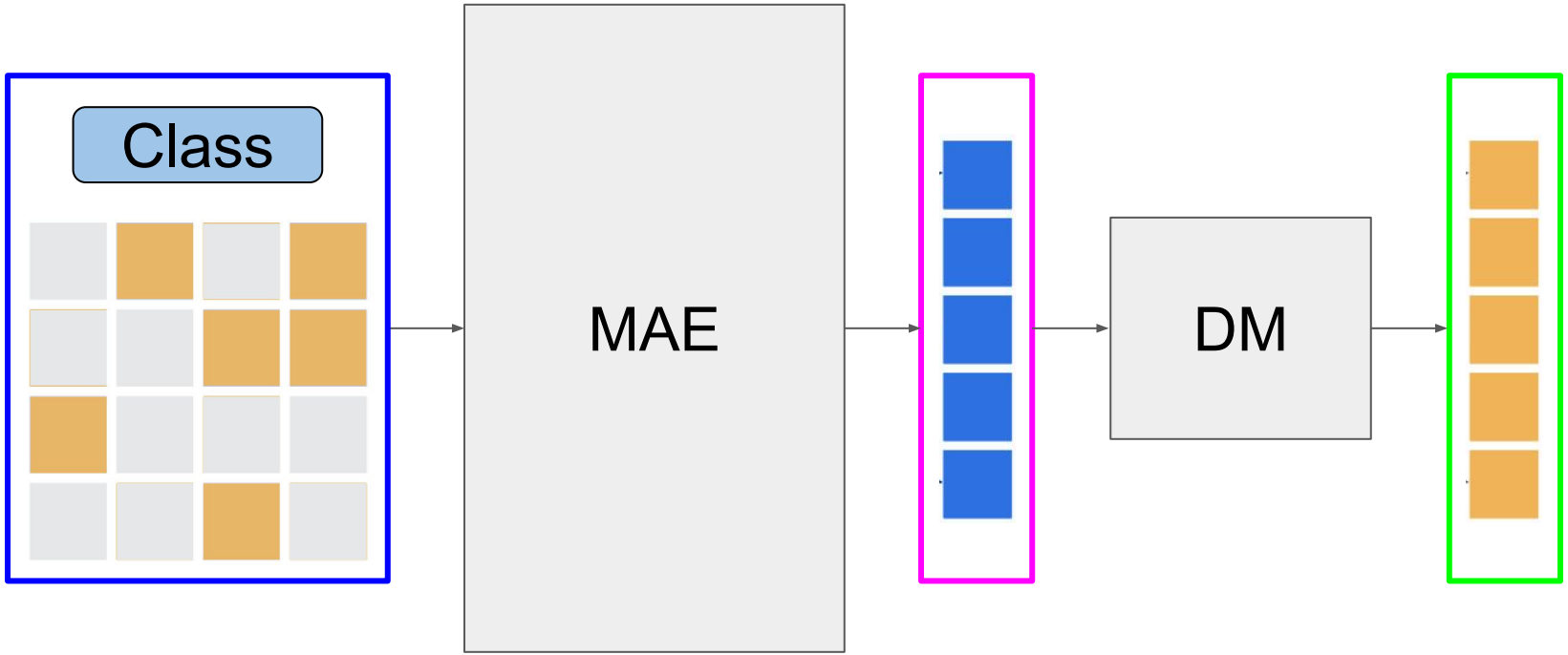
Masked AutoRegressive Modeling (MAR)



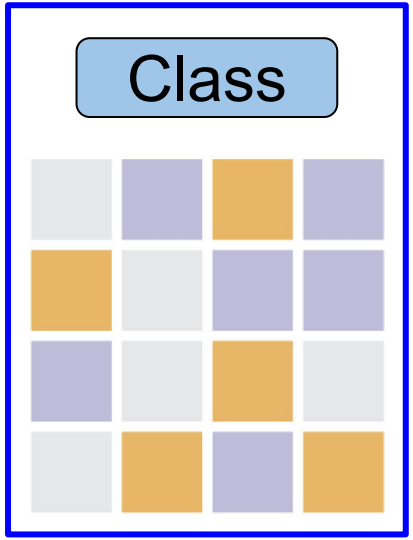
Masked AutoRegressive Modeling (MAR)



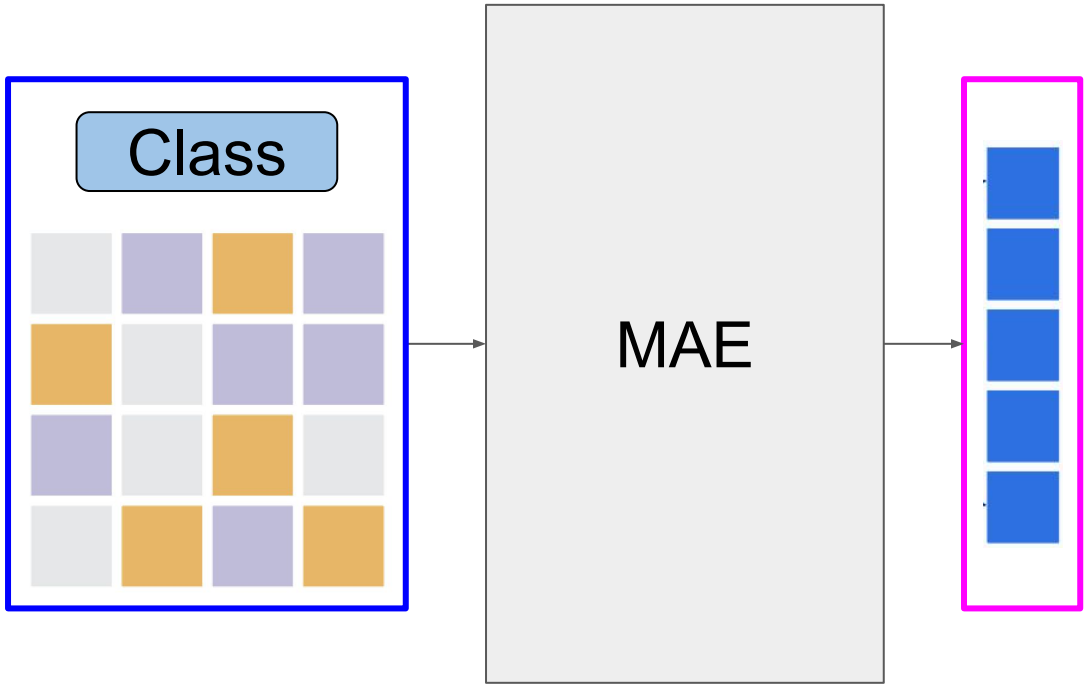
Masked AutoRegressive Modeling (MAR)



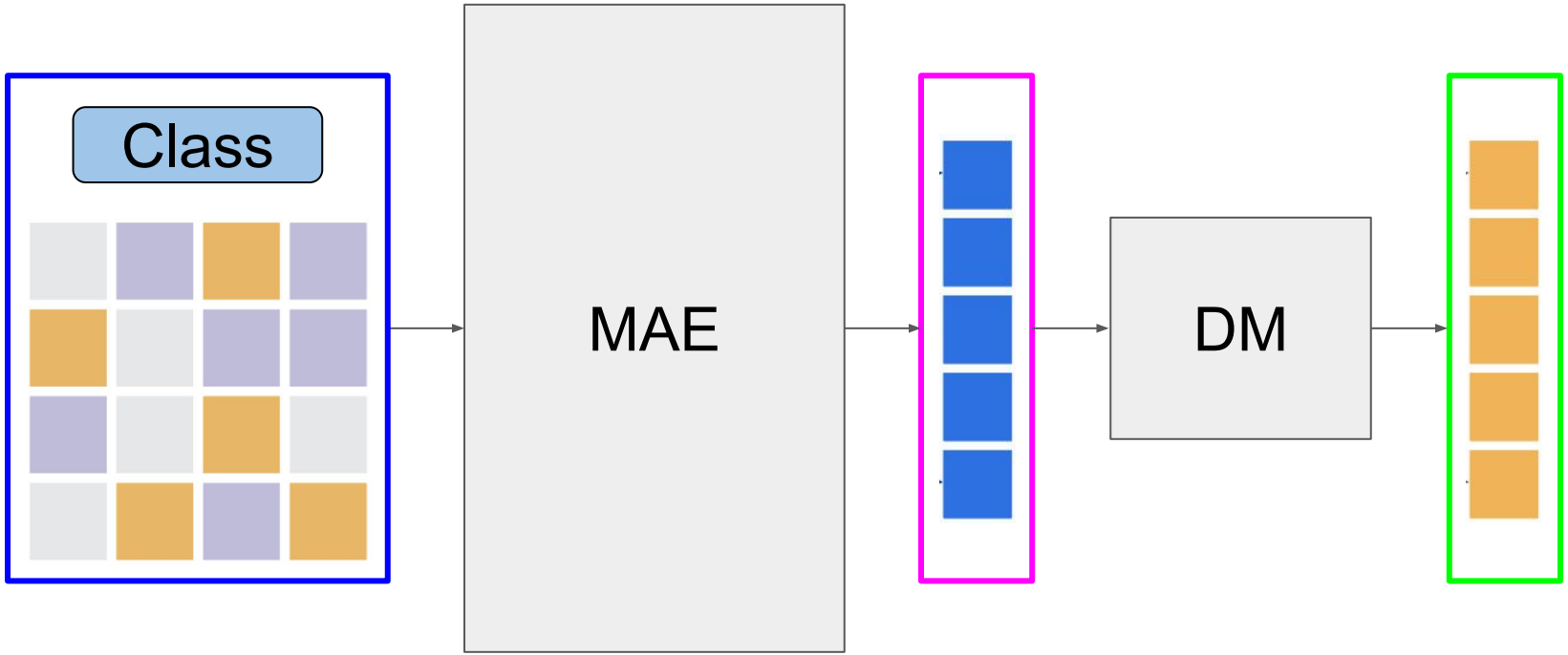
Masked AutoRegressive Modeling (MAR)



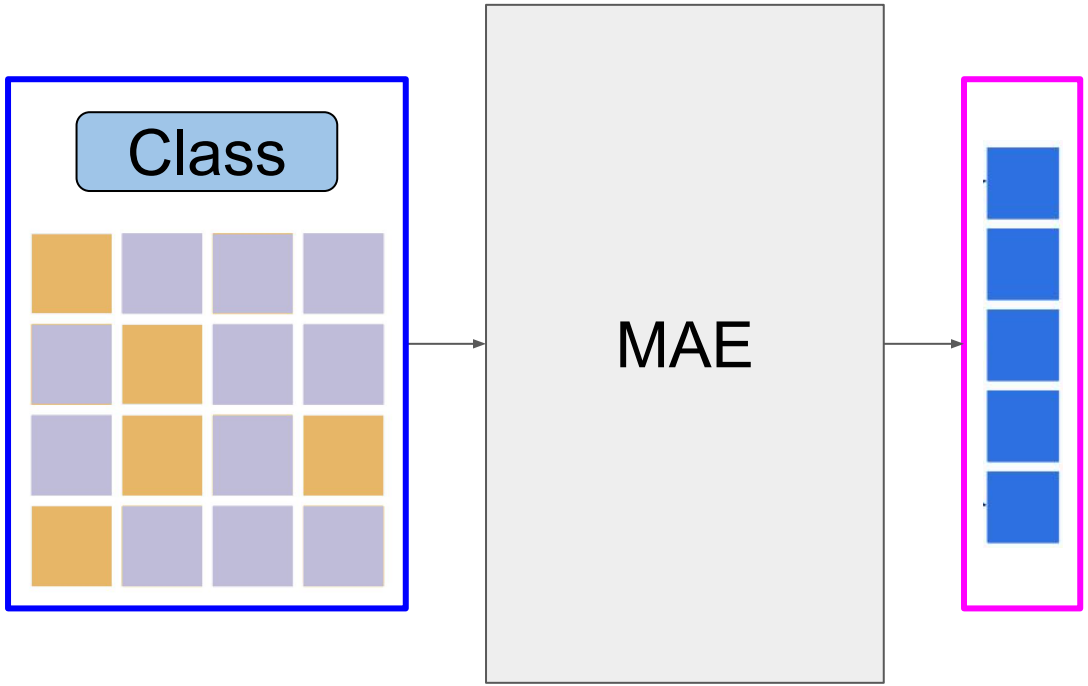
Masked AutoRegressive Modeling (MAR)



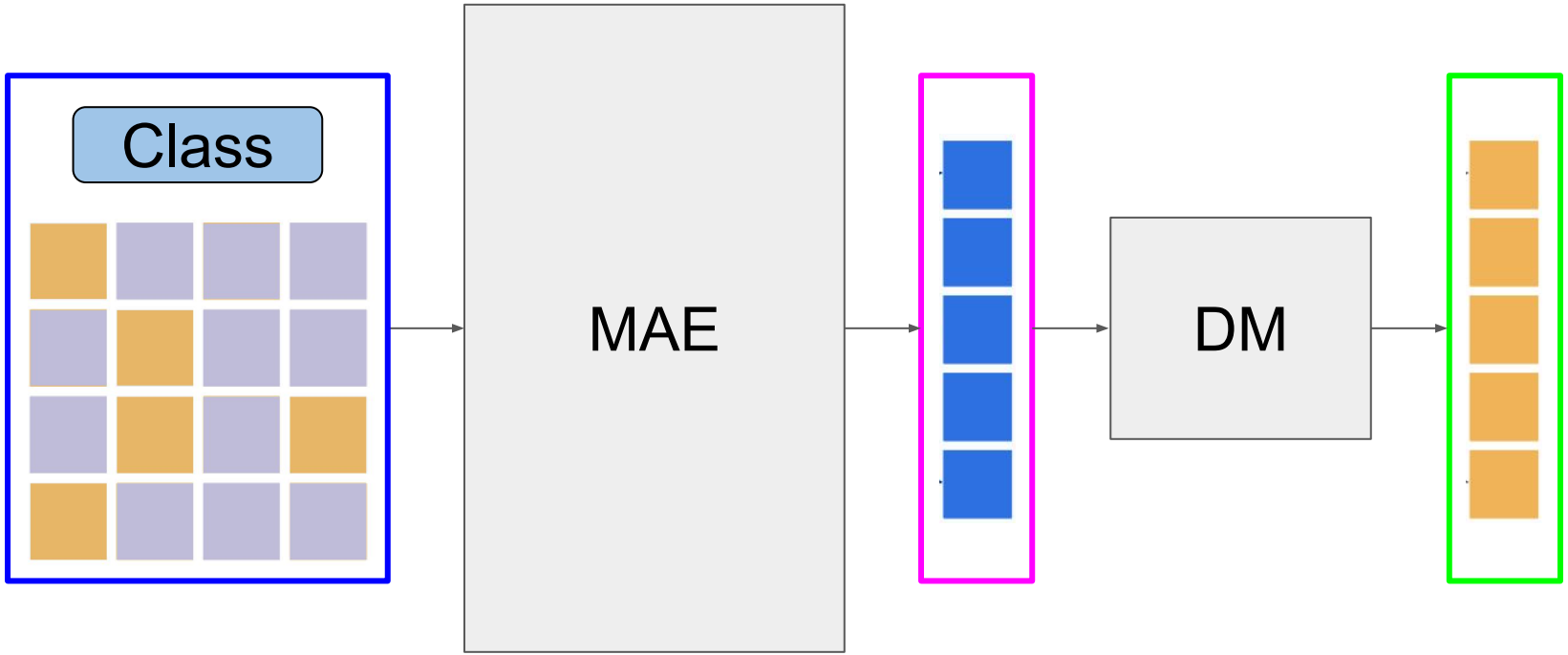
Masked AutoRegressive Modeling (MAR)



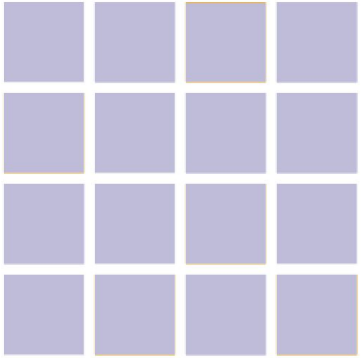
Masked AutoRegressive Modeling (MAR)



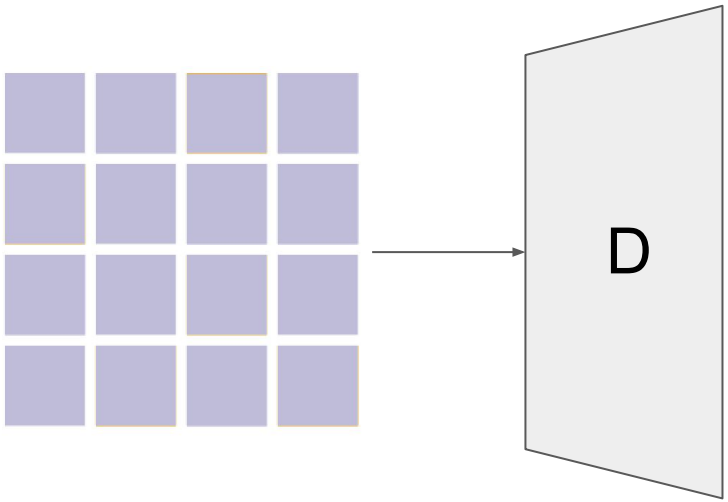
Masked AutoRegressive Modeling (MAR)



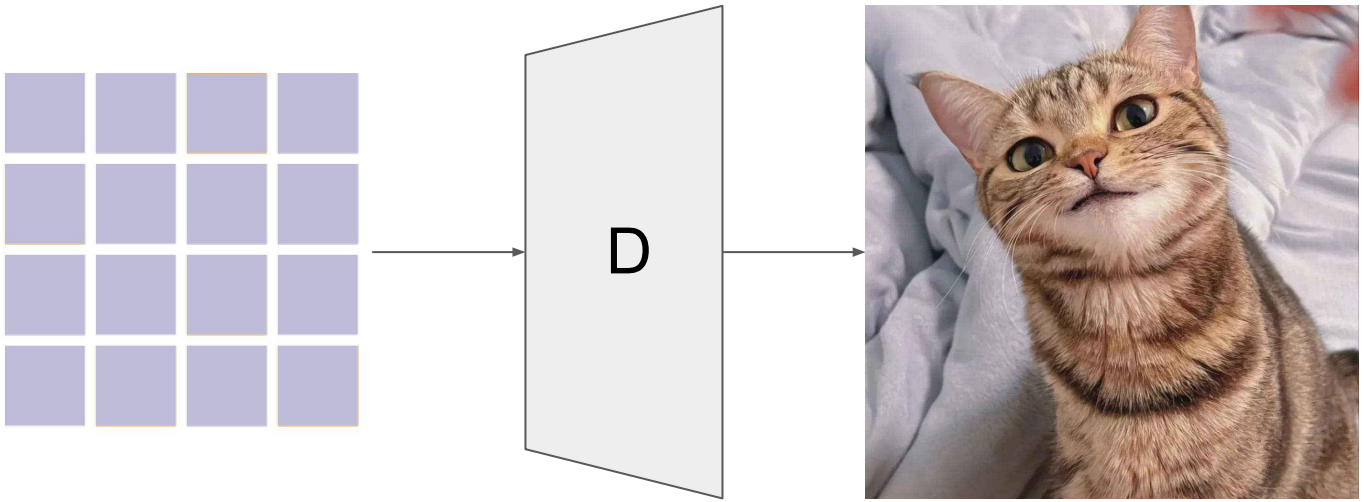
Masked AutoRegressive Modeling (MAR)



Masked AutoRegressive Modeling (MAR)



Masked AutoRegressive Modeling (MAR)



MAR has a great performance

Model	FID
VAR-d30	1.92
RAR-XXL	1.48

[Li T., et al. 2024]

MAR has a great performance

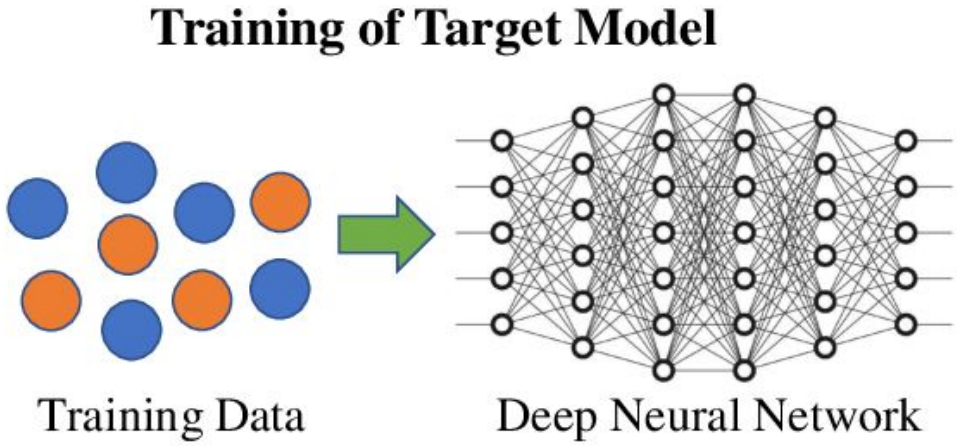
Model	FID
VAR-d30	1.92
RAR-XXL	**1.48**
MAR-B	2.31
MAR-L	1.78
MAR-H	1.55

[Li T., et al. 2024]

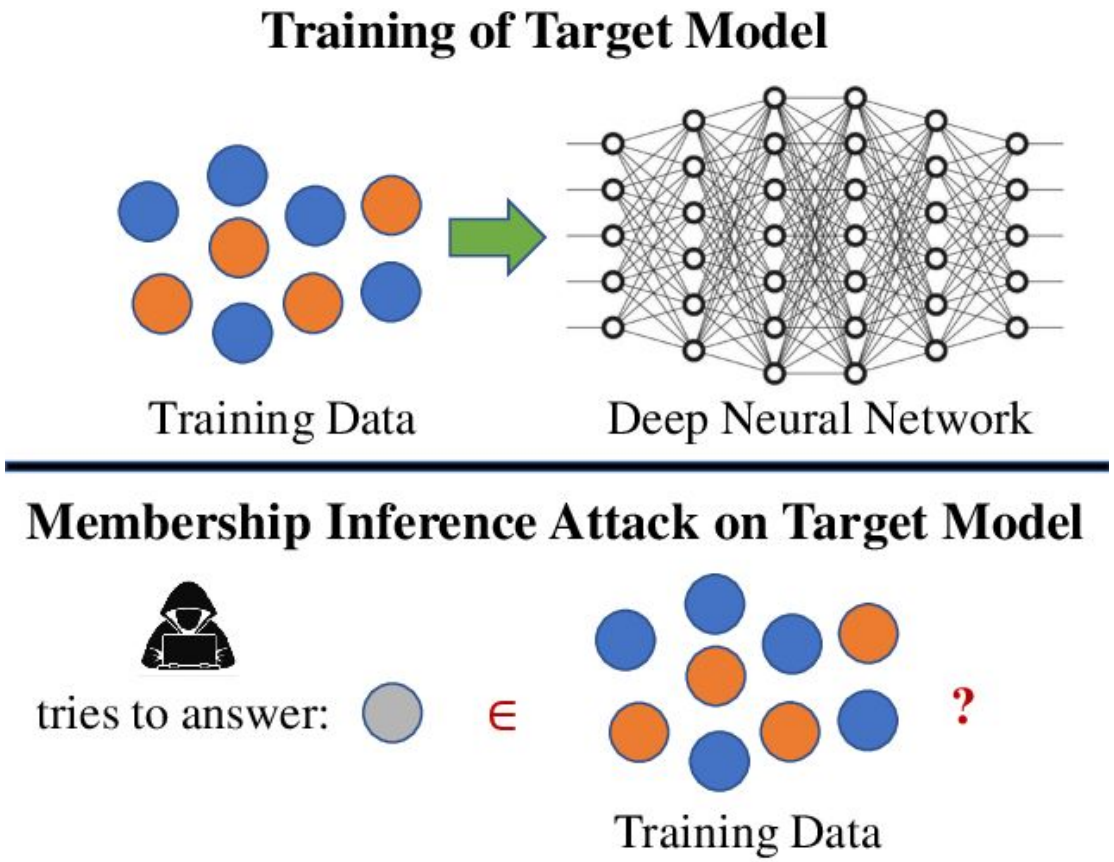
Next up: attacks!

1. MIA
2. DI
3. Memorization

Membership Inference Attacks (MIAs)

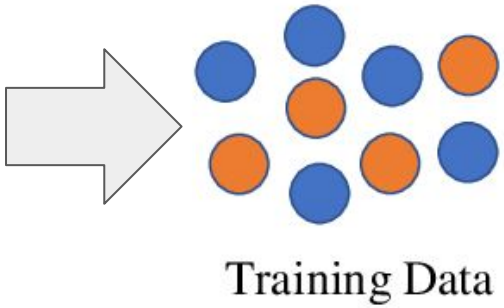


Membership Inference Attacks (MIAs)

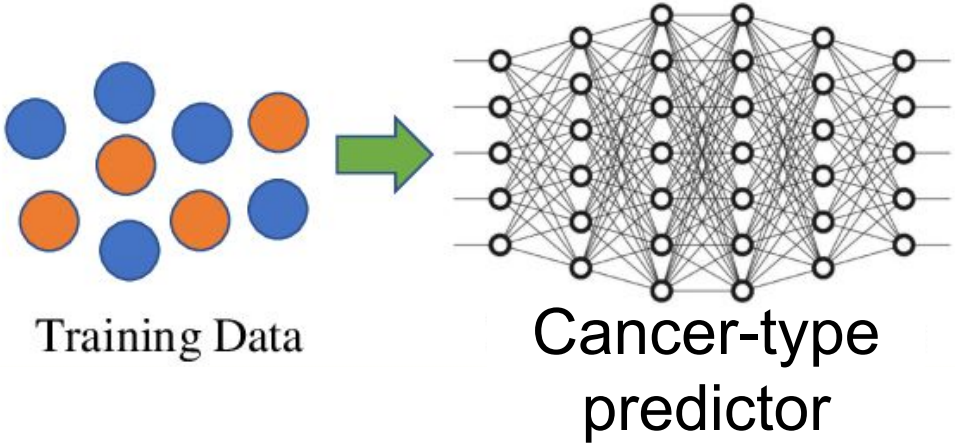


Privacy risk of MIA

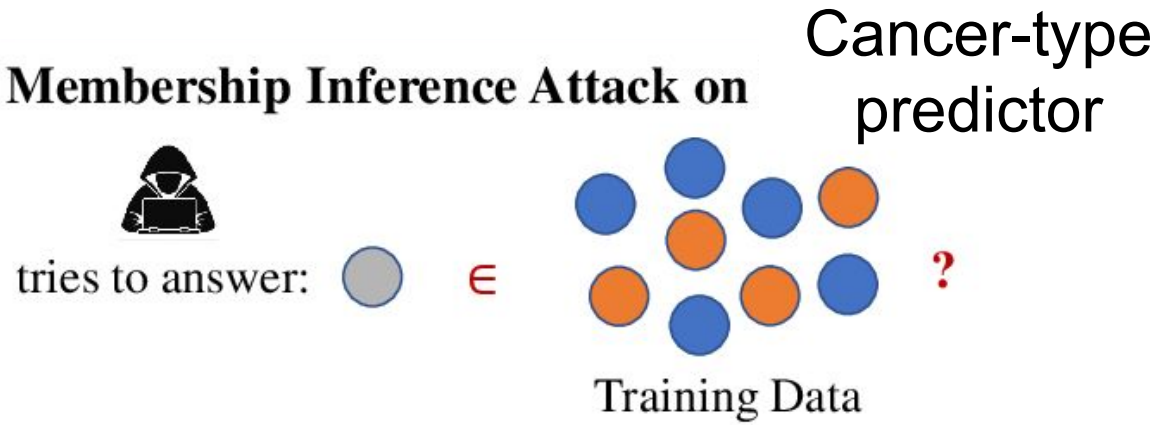
Patients with cancer



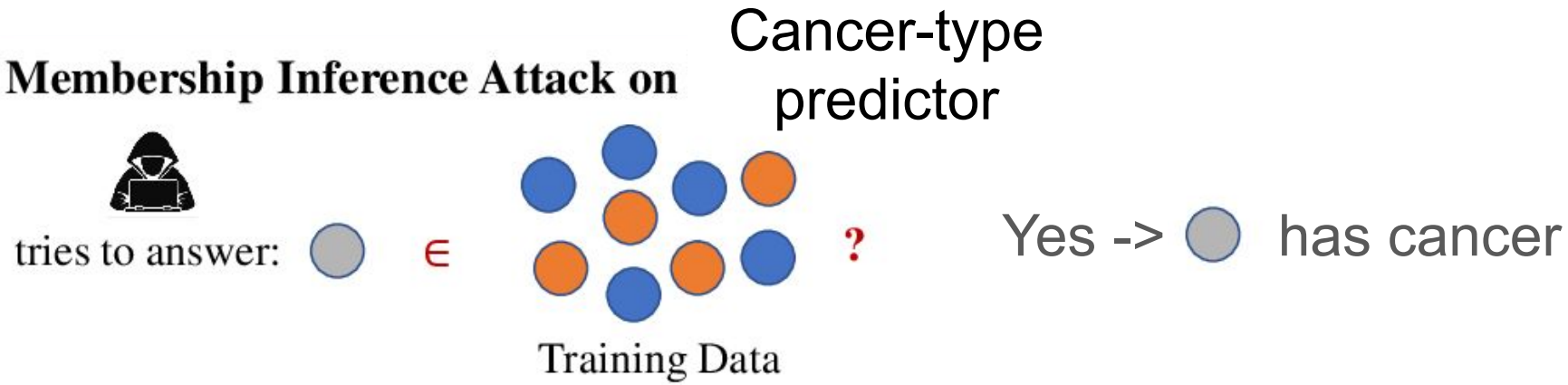
Privacy risk of MIA



Privacy risk of MIA



Privacy risk of MIA



Use case: lawsuits!

Reuters

World Business Markets Sustainability Legal Breakingviews Technology In

Getty Images lawsuit says Stability AI misused photos to train AI

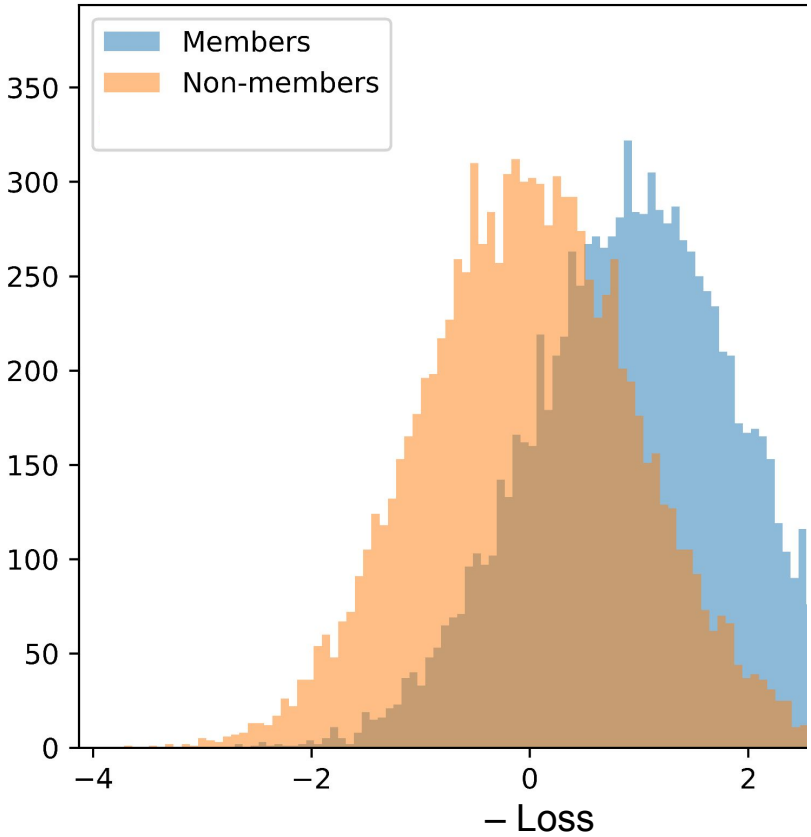
By **Blake Brittain**

February 6, 2023 6:32 PM GMT+1 · Updated 2 years ago

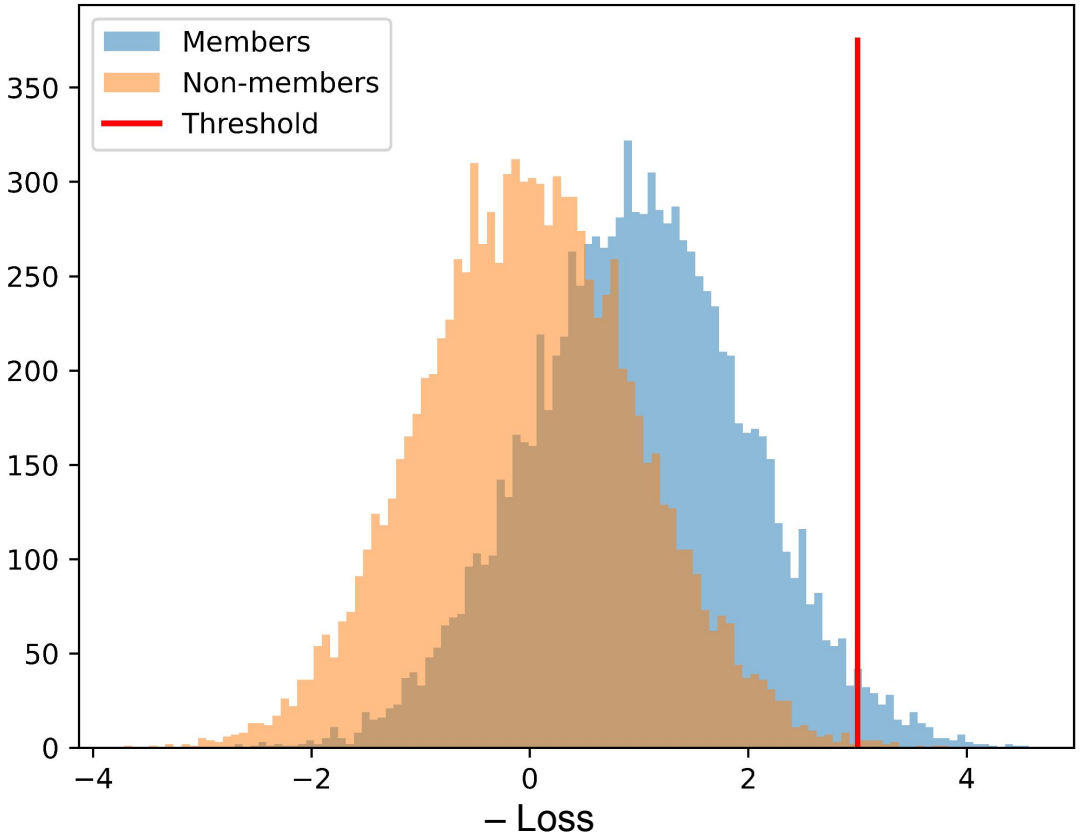
Aa

MIA: general intuition

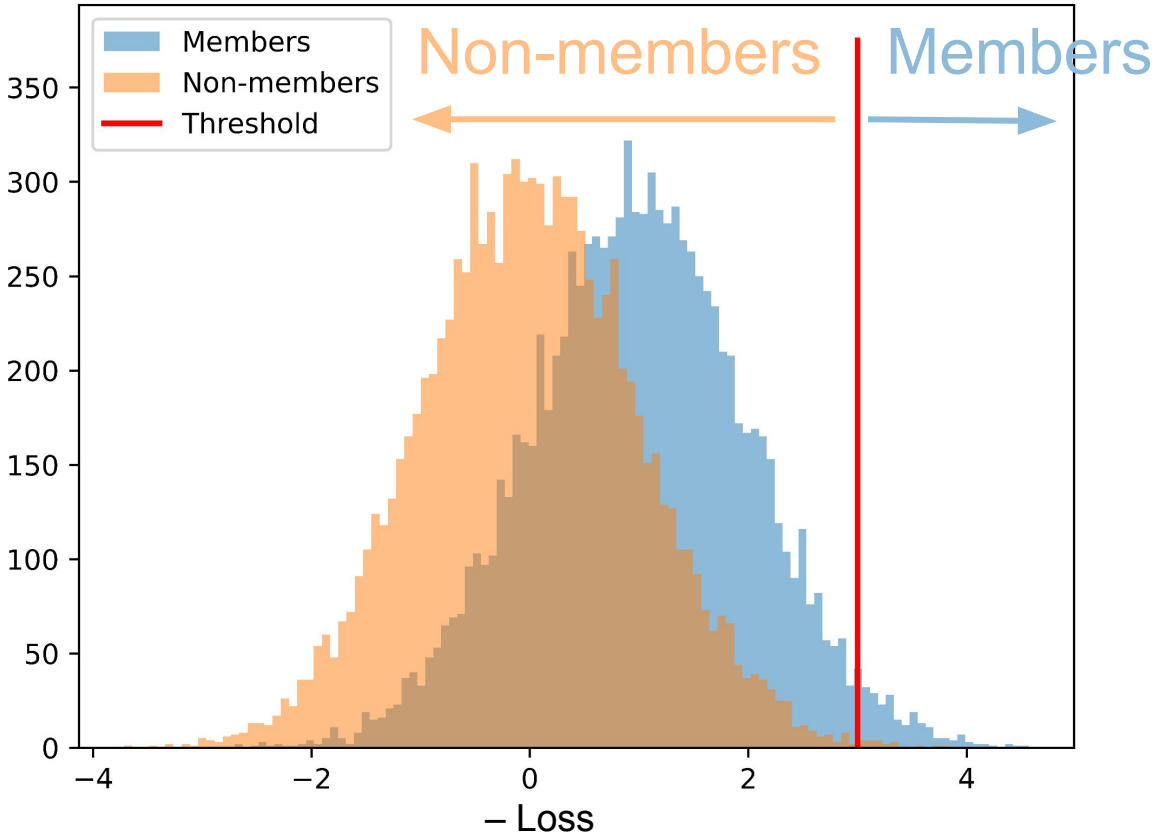
MIA: general intuition



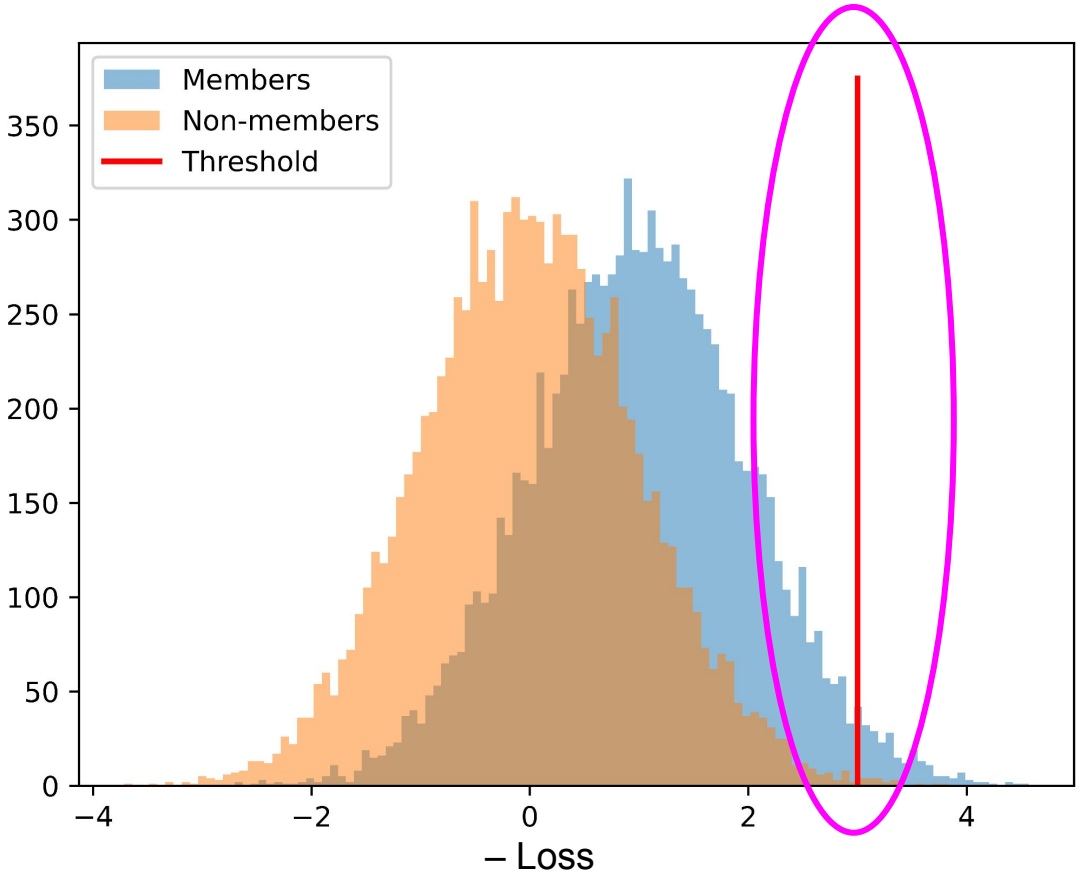
Loss Threshold Attack



Loss Threshold Attack



Why such a threshold?

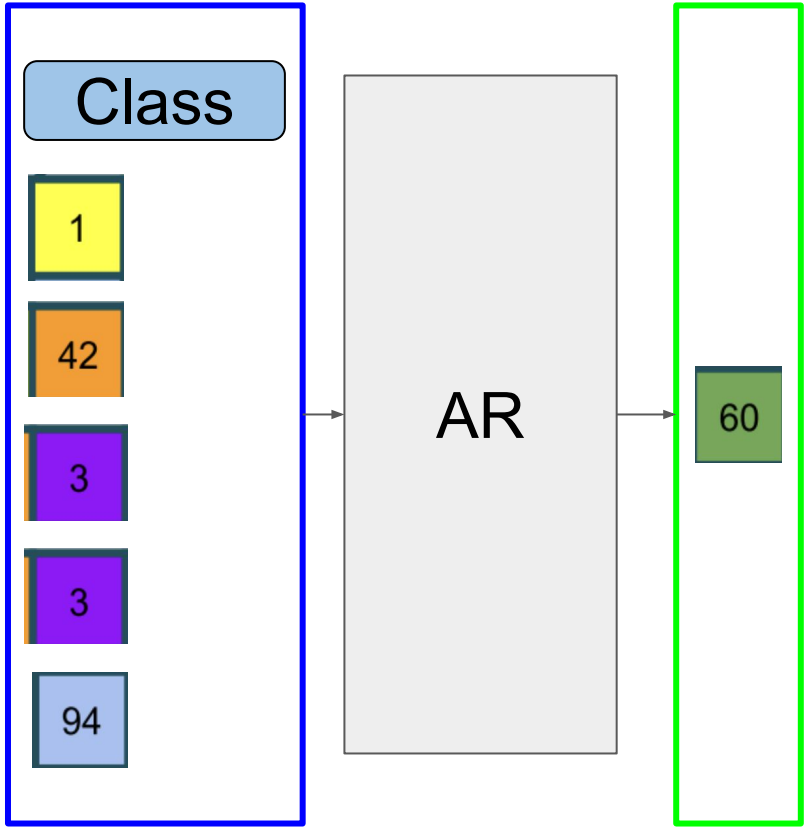


MIA metric

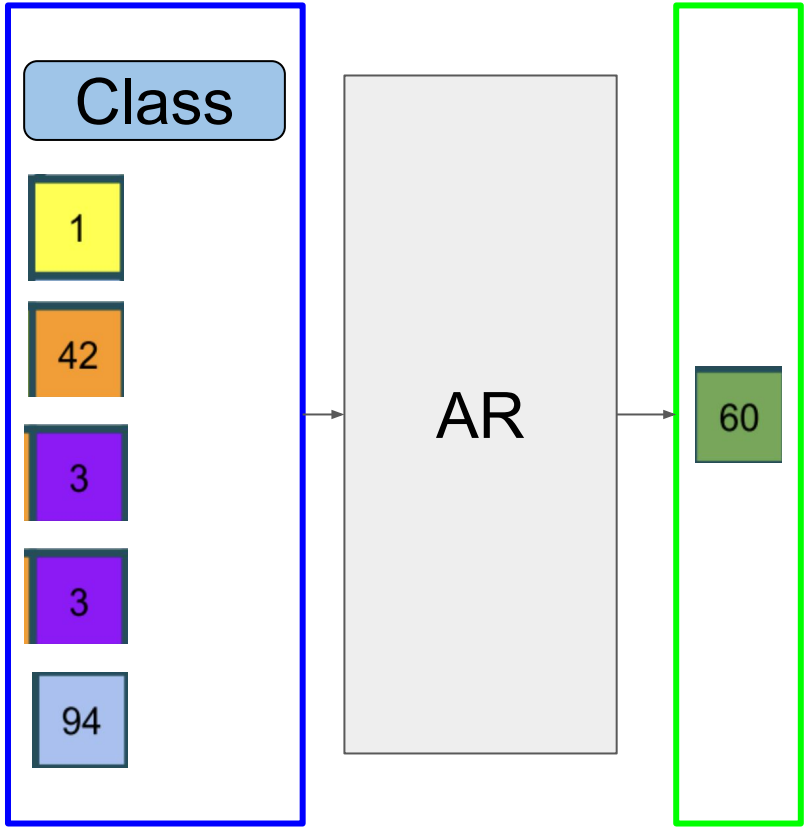
TPR@FPR=1%

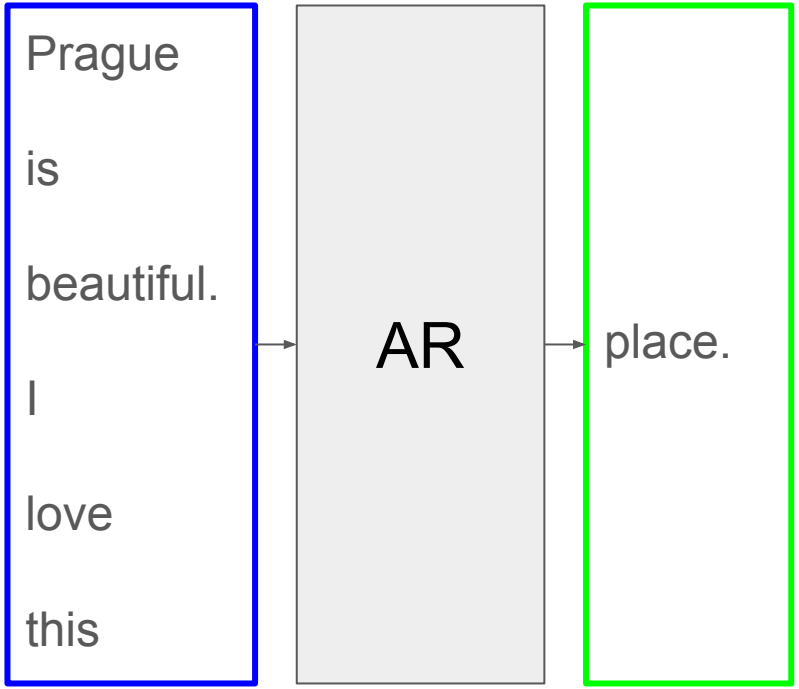
More successful MIA
=> less private model

Starting point: MIAs for LLMs

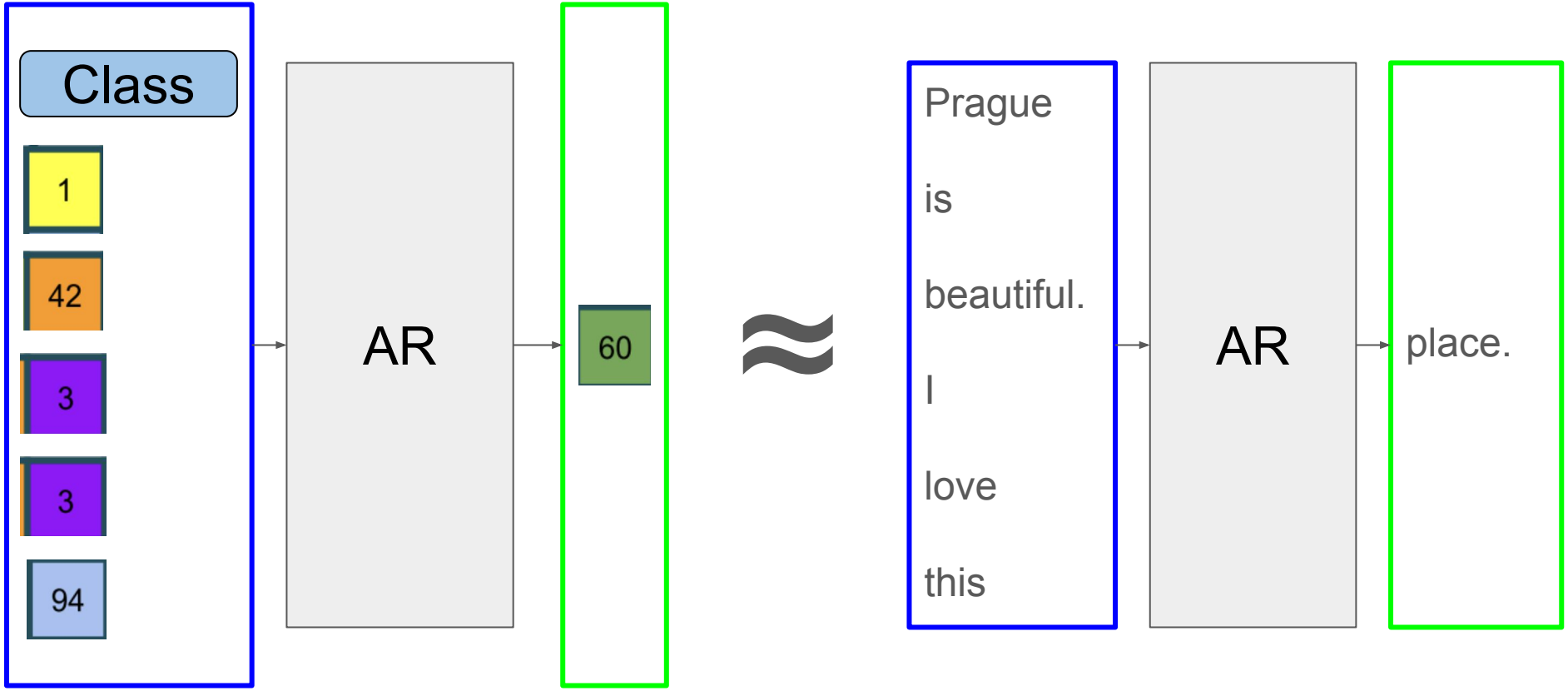


Starting point: MIAs for LLMs

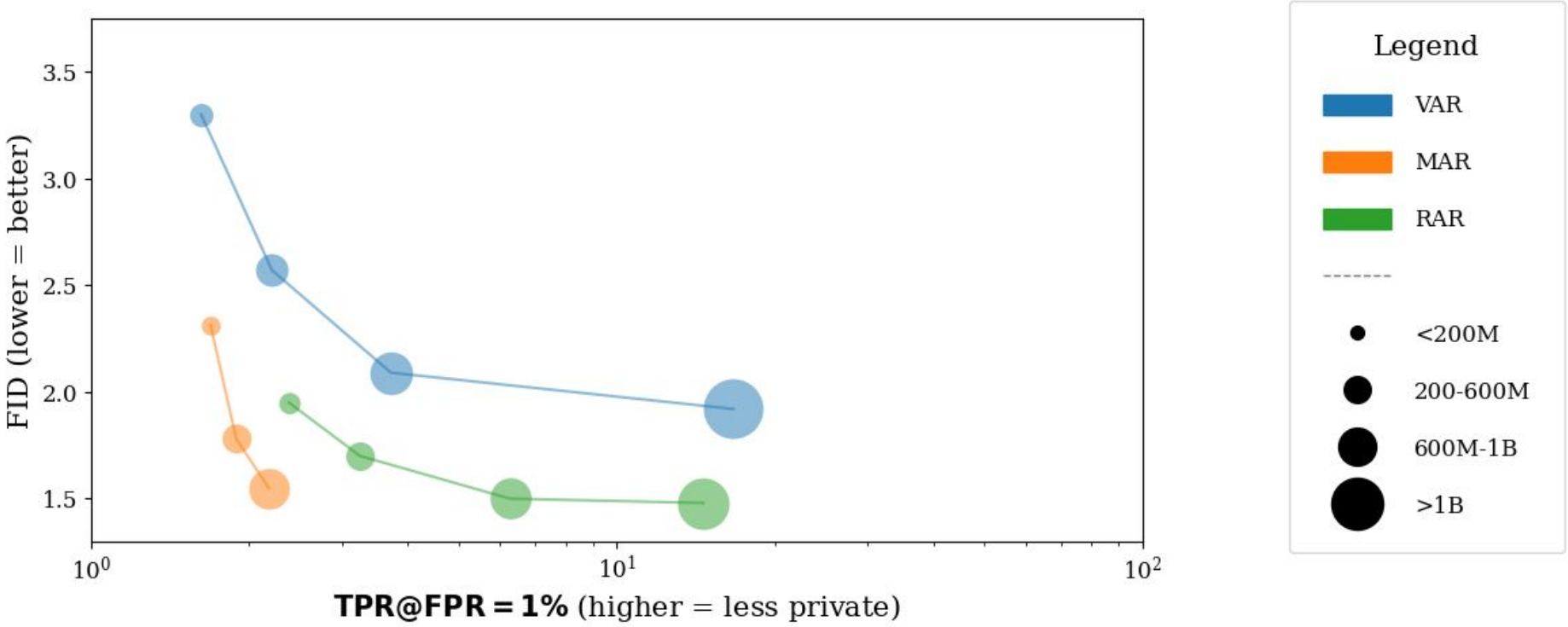




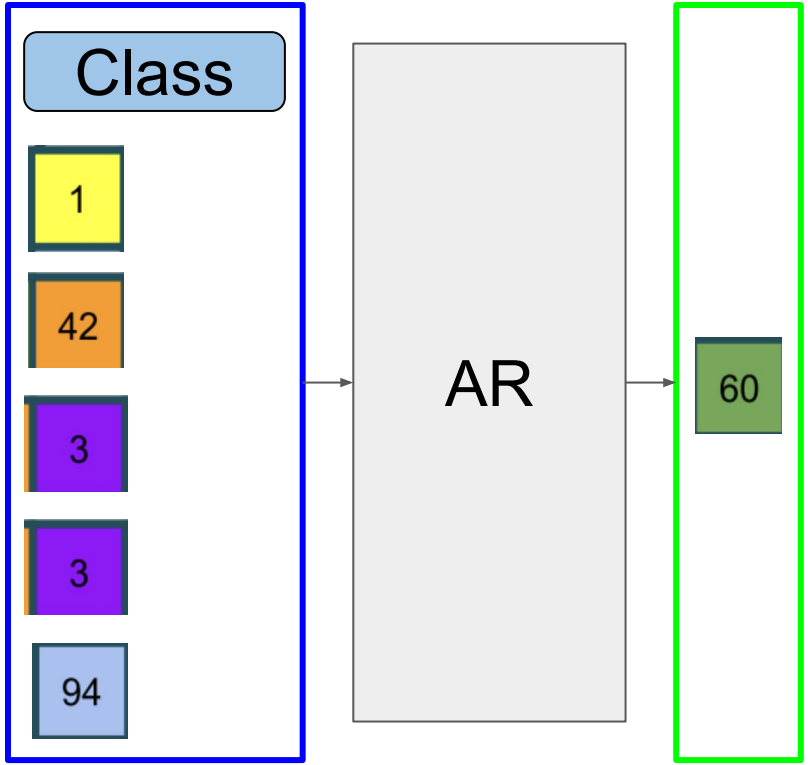
Starting point: MIAs for LLMs



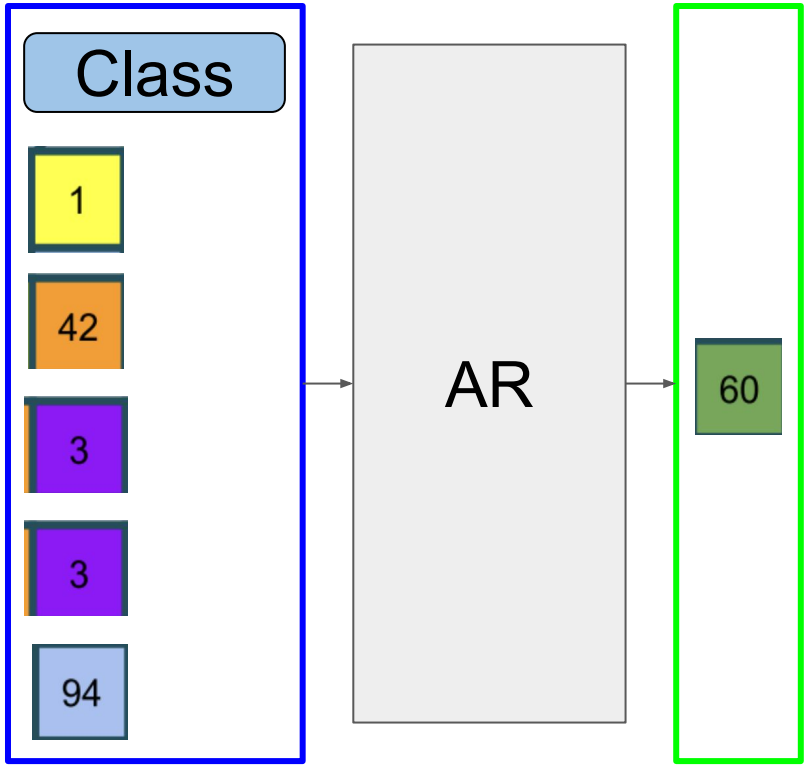
Starting point: MIAs for LLMs

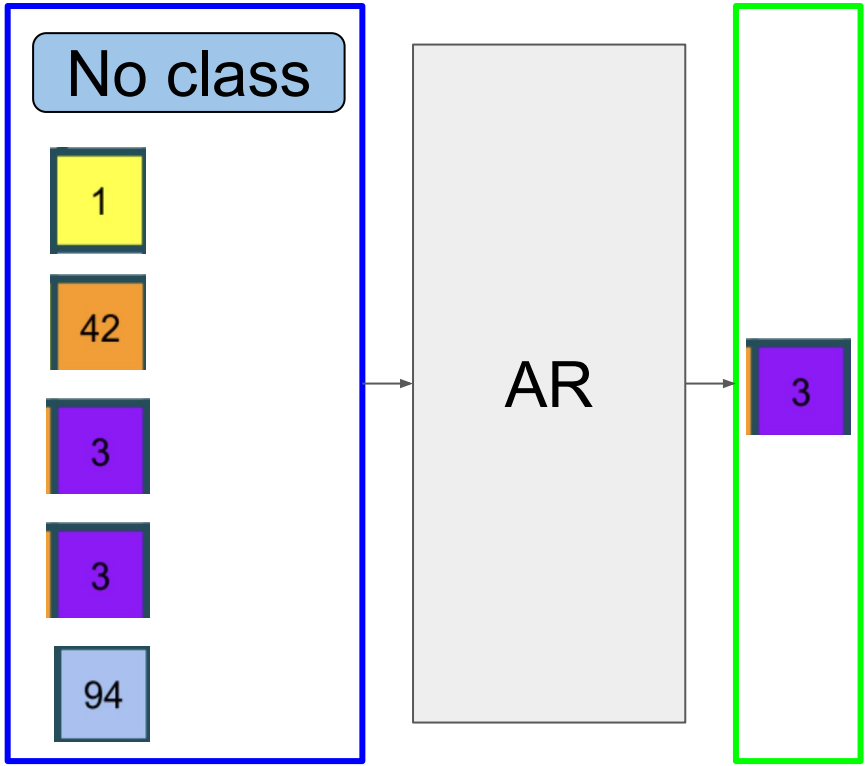


Our improvements: VAR and RAR



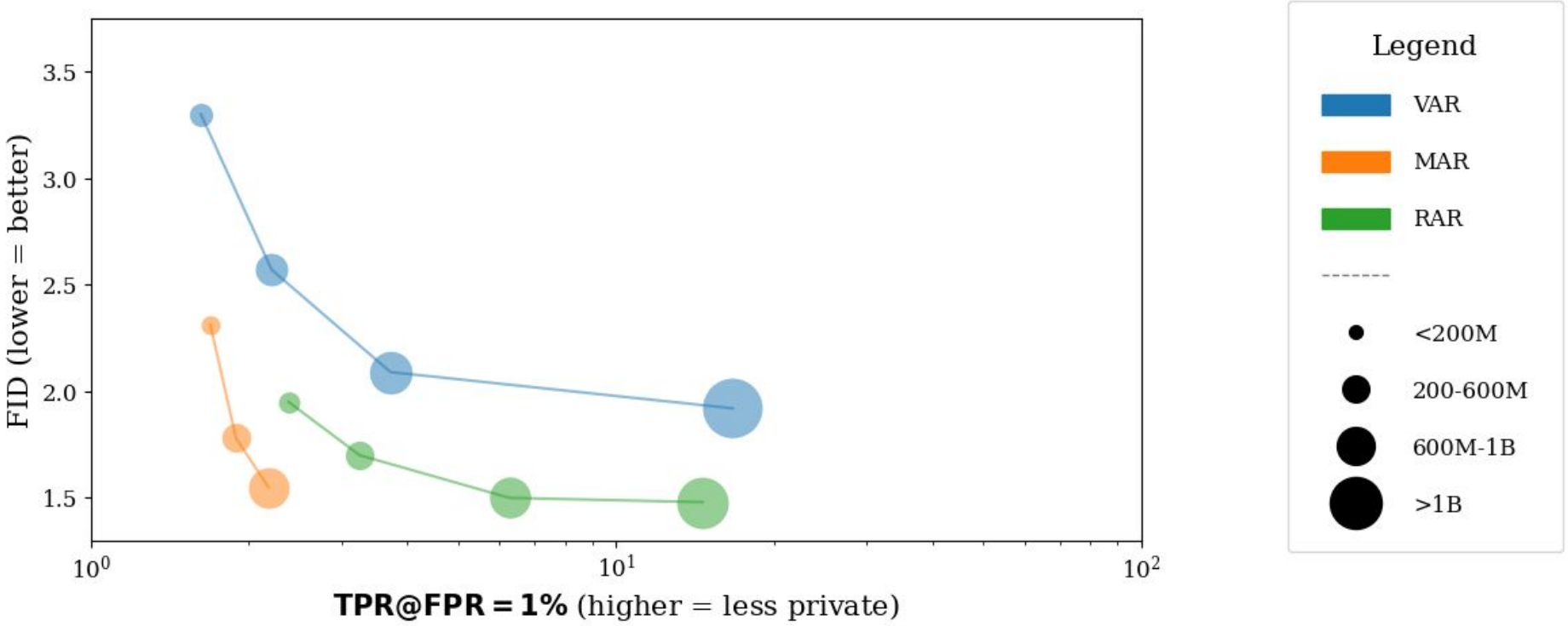
Our improvements: VAR and RAR



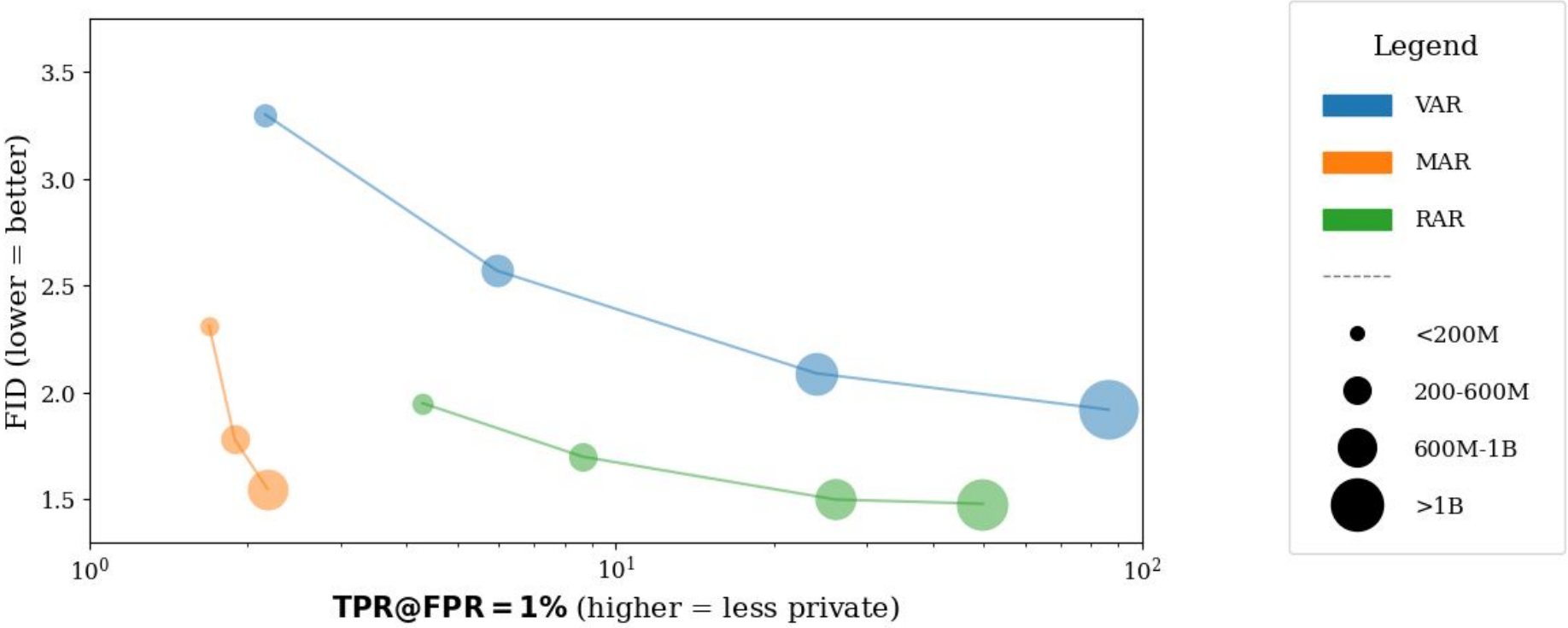


Our improvements: VAR and RAR

Before


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

After

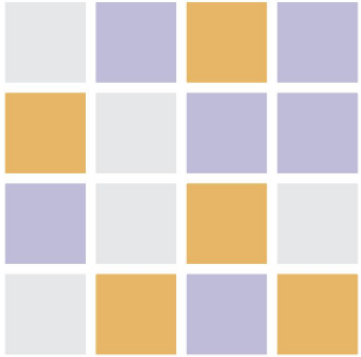

FID (lower = better)
3.5
3.0
2.5
2.0
1.5
10^0
10^1
10^2
TPR@FPR = 1% (higher = less private)
Legend
VAR
MAR
RAR
<200M
200-600M
600M-1B
>1B

Biggest improvements

Model	Before	After
VAR-d30	16.68	**86.38**
RAR-XXL	14.62	**49.80**

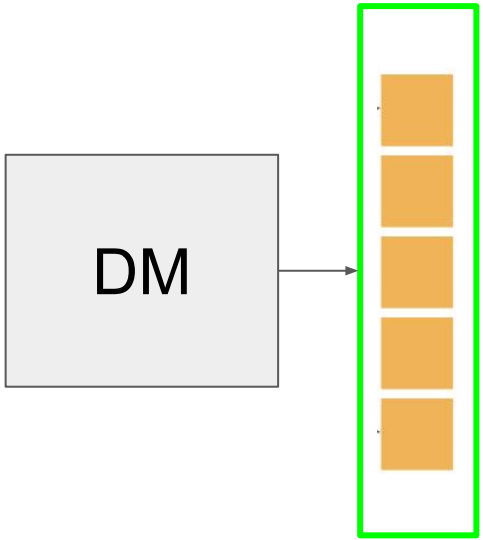
Our improvements: MAR

1. Adjust batch size



Our improvements: MAR

2. Manipulate the Diffusion Model



Our improvements: MAR

Model	Baseline
MAR-B	1.69
MAR-L	1.89
MAR-H	2.18

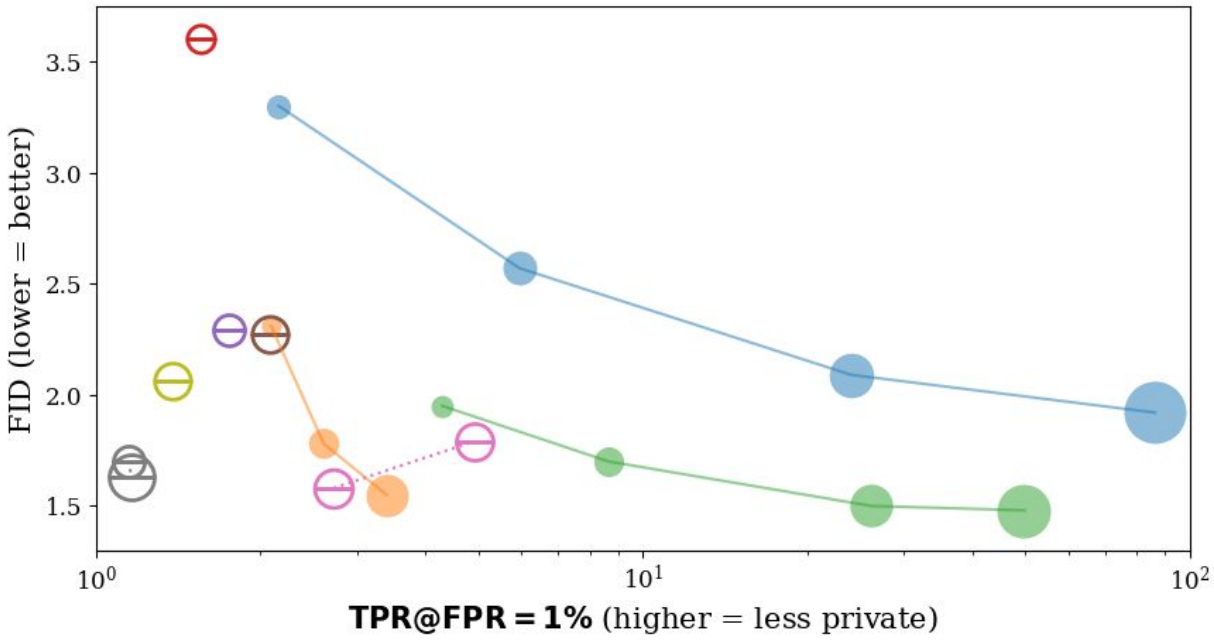
Our improvements: MAR

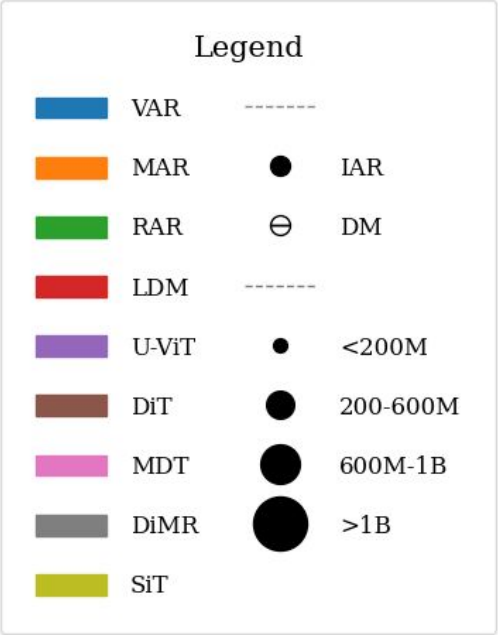
Model	Baseline	+Adjusted Batch Size
MAR-B	1.69	1.88
MAR-L	1.89	2.25
MAR-H	2.18	2.88

Our improvements: MAR

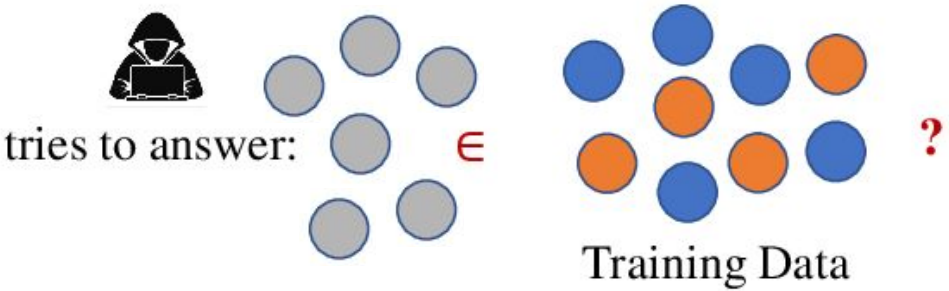
Model	Baseline	+Adjusted Batch Size	+Manipulations of Diffusion Model
MAR-B	1.69	1.88	**2.09**
MAR-L	1.89	2.25	**2.61**
MAR-H	2.18	2.88	**3.40**

IARs are more prone to MIAs than DMs





Dataset Inference (DI)



Use case: lawsuits!

Reuters

World Business Markets Sustainability Legal Breakingviews Technology Investigati

Artists take new shot at Stability, Midjourney in updated copyright lawsuit

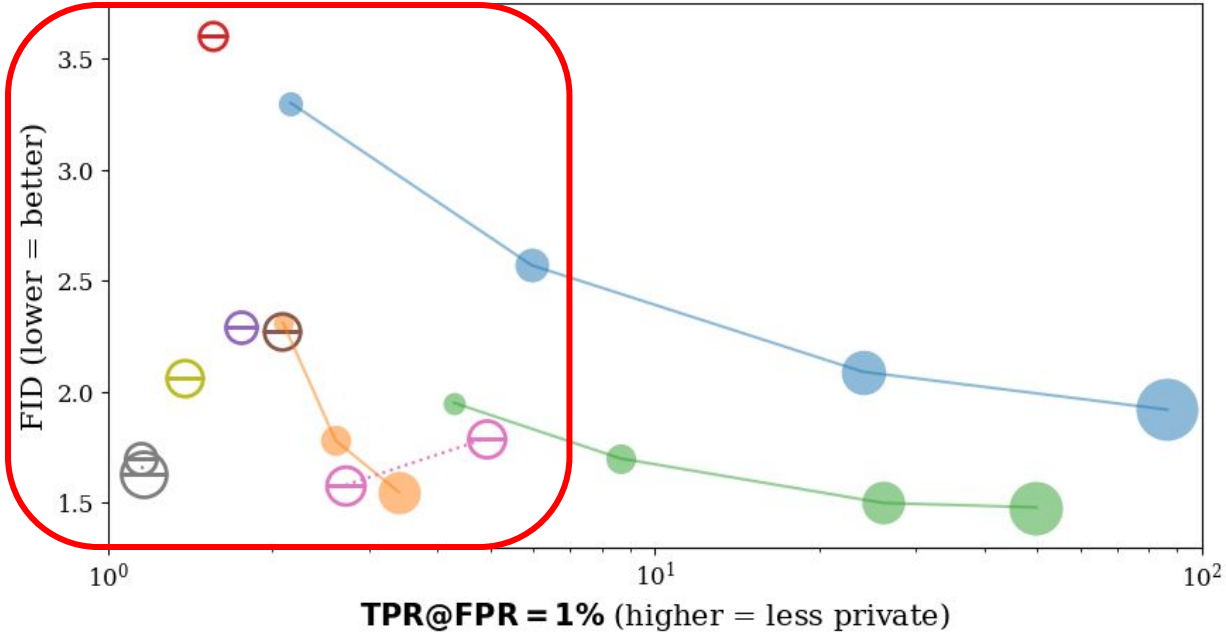
By **Blake Brittain**

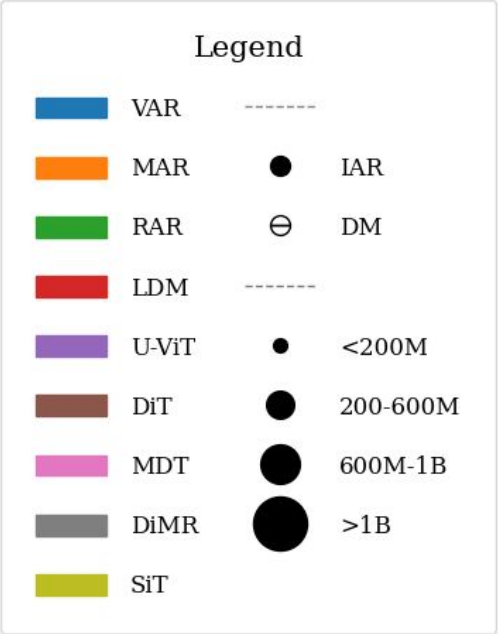
November 30, 2023 8:47 PM GMT+1 · Updated a year ago

Aa

Why DI: it's easier than MIA

~random guessing





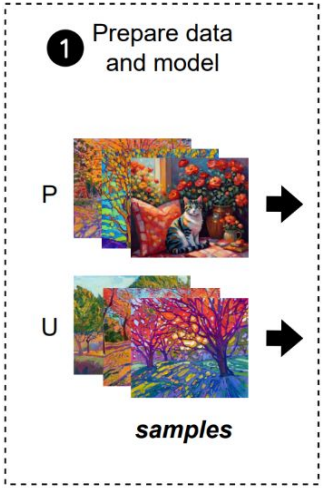
Copyrighted Data Identification (CDI)

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

CVPR

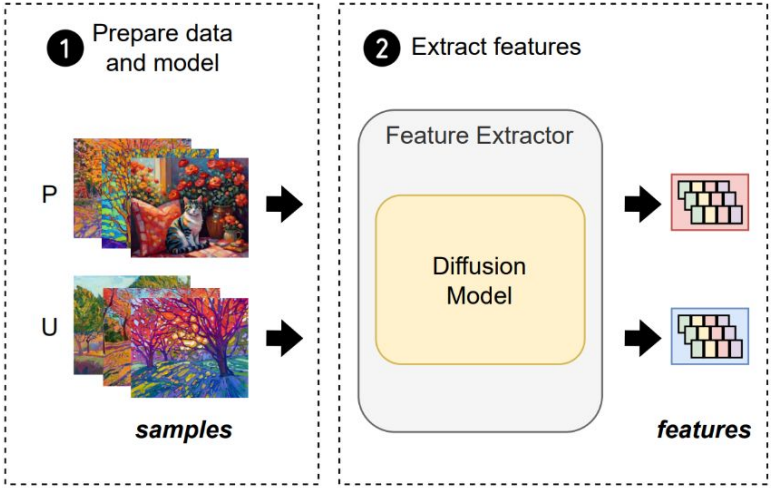
2025

CDI pipeline



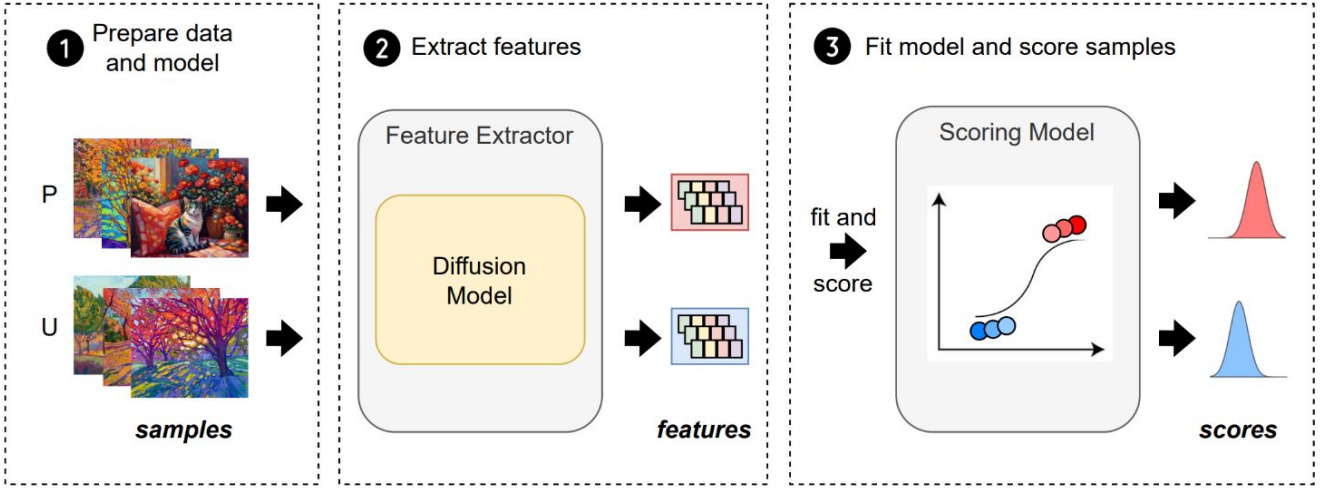
[Dubiński J., et al. 2024]

CDI pipeline



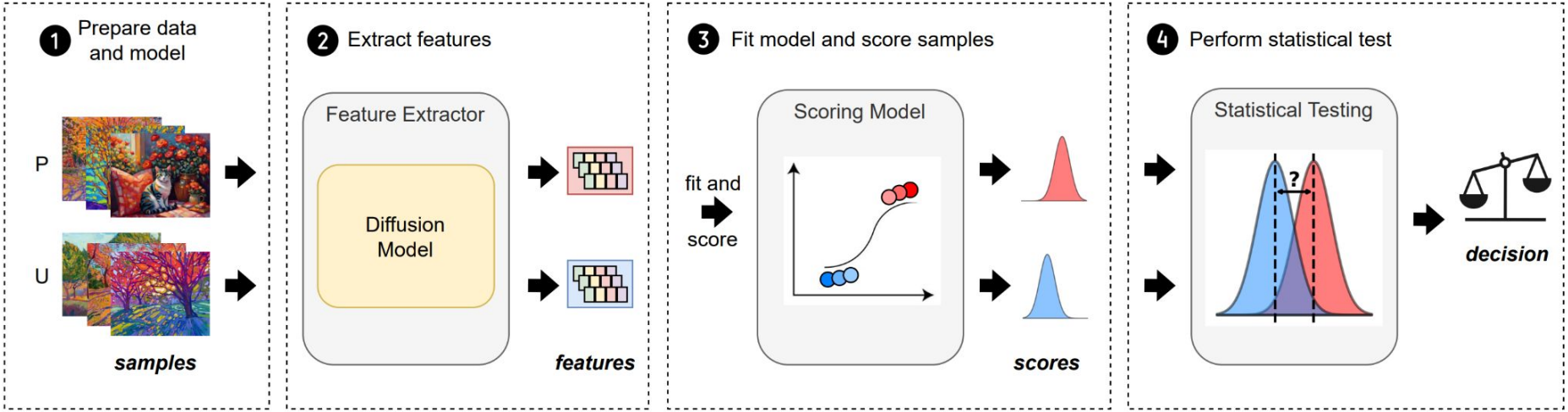
[Dubiński J., et al. 2024]

CDI pipeline



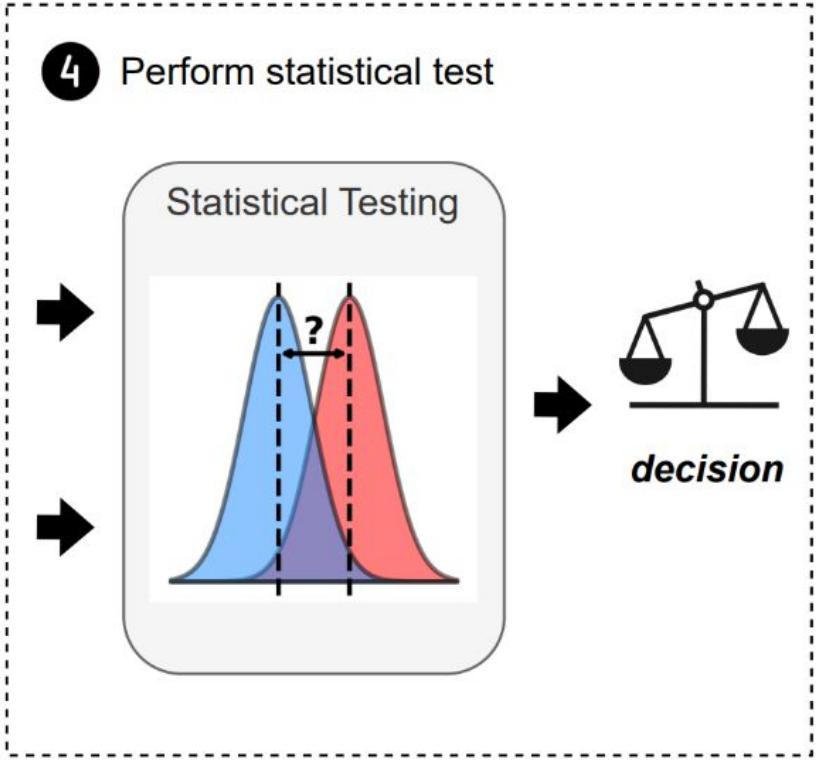
[Dubiński J., et al. 2024]

CDI pipeline



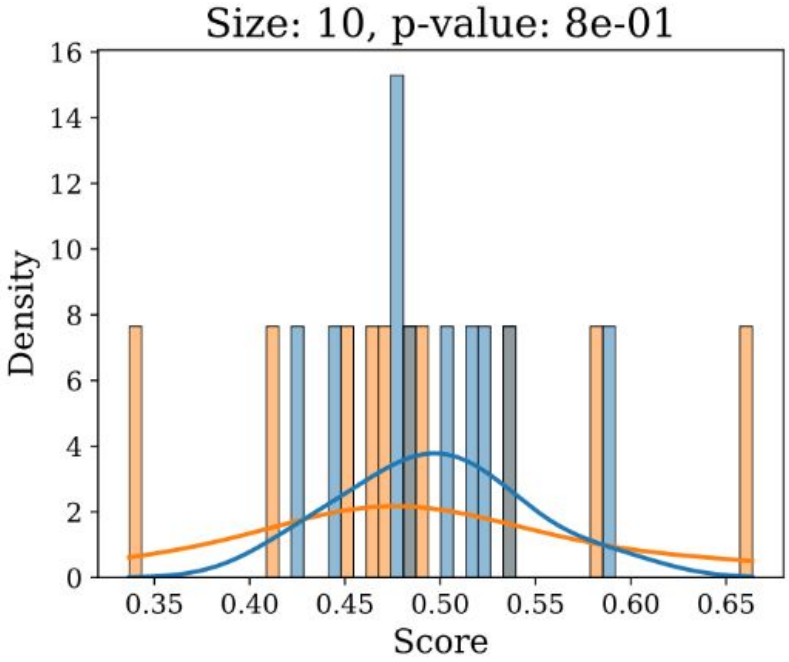
[Dubiński J., et al. 2024]

Key component: difference between distributions



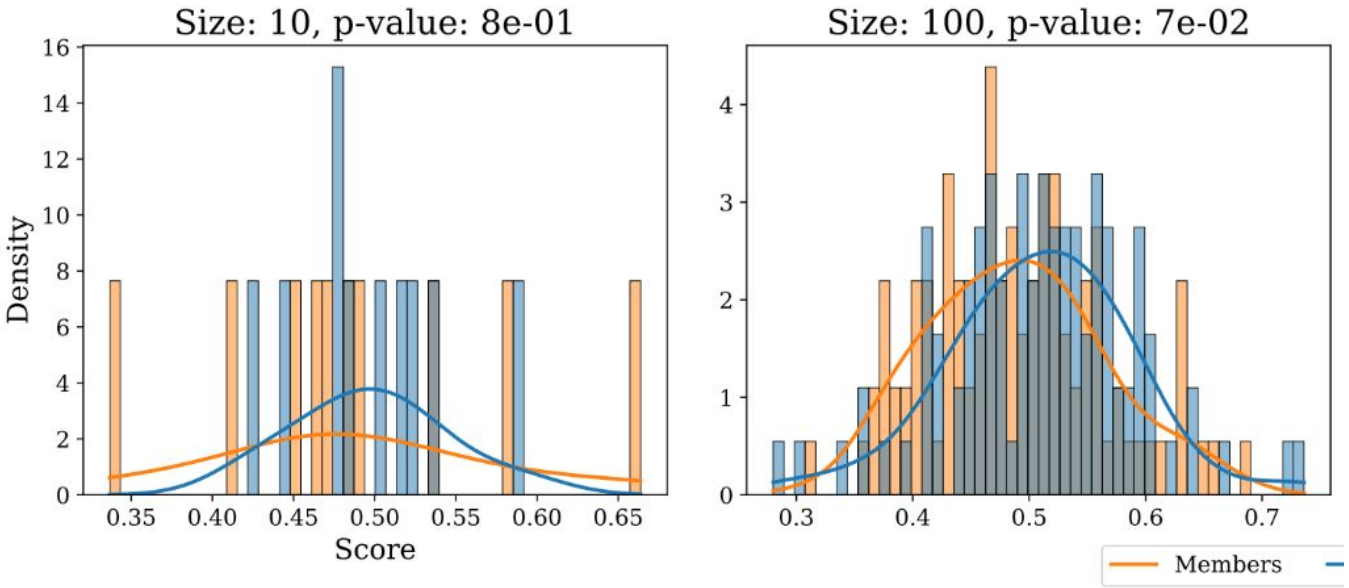
[Dubiński J., et al. 2024]

What matters? Size of P!



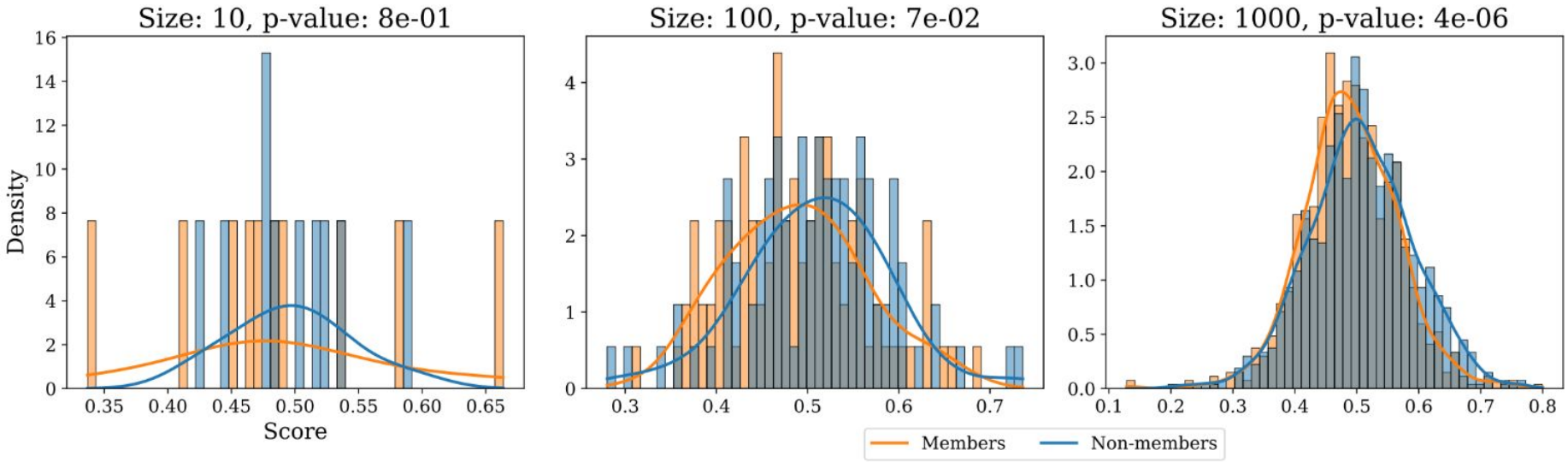
[Dubiński J., et al. 2024]

What matters? Size of P!



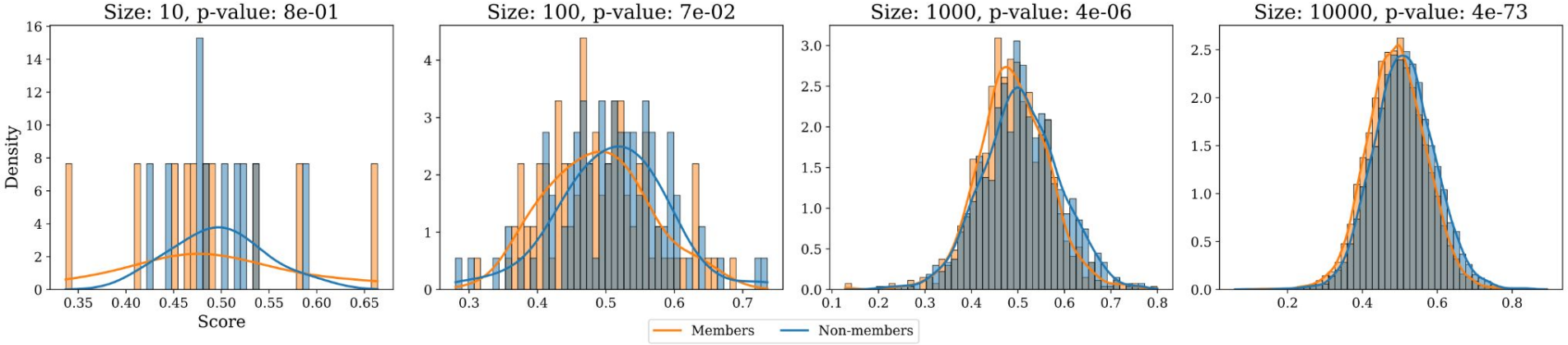
[Dubiński J., et al. 2024]

What matters? Size of P!



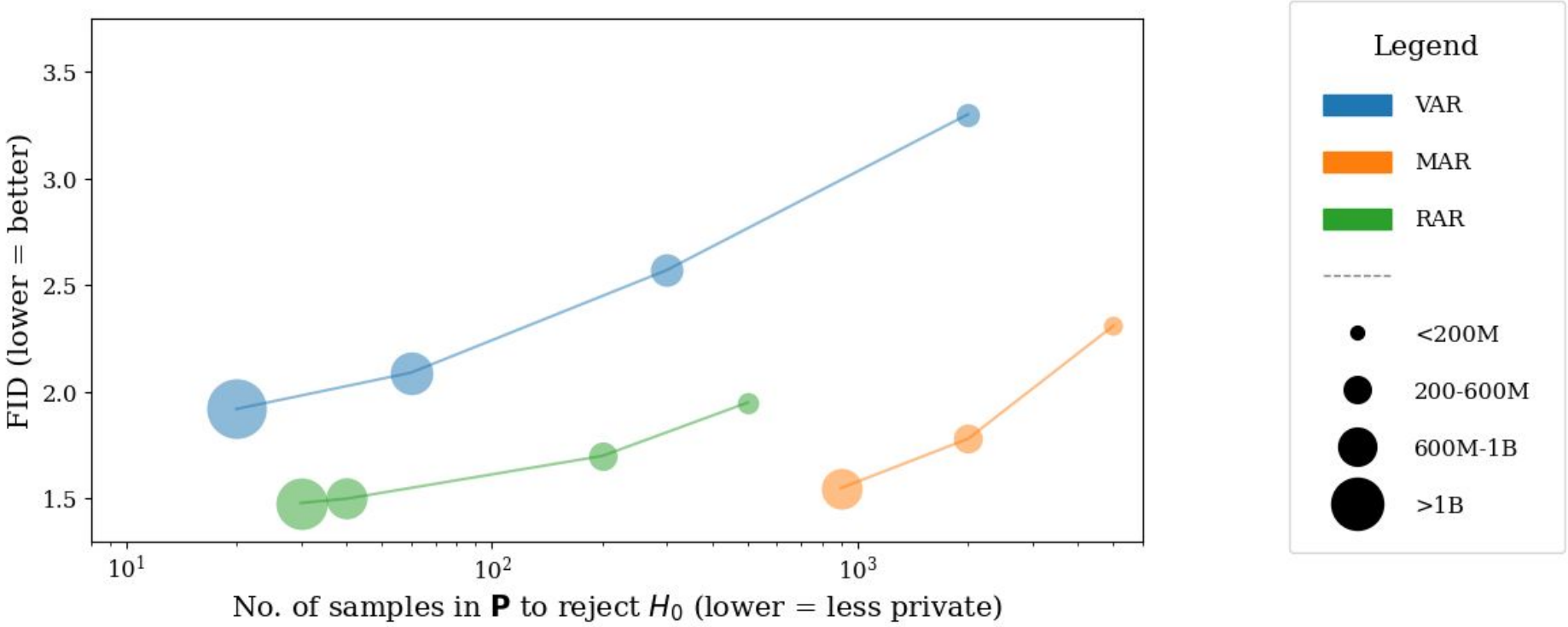
[Dubiński J., et al. 2024]

What matters? Size of P!

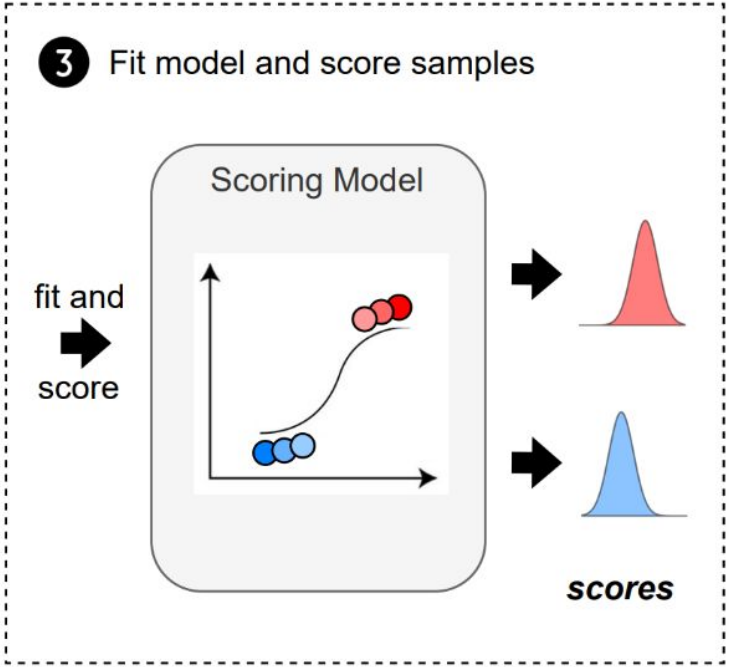


[Dubiński J., et al. 2024]

CDI: initial results

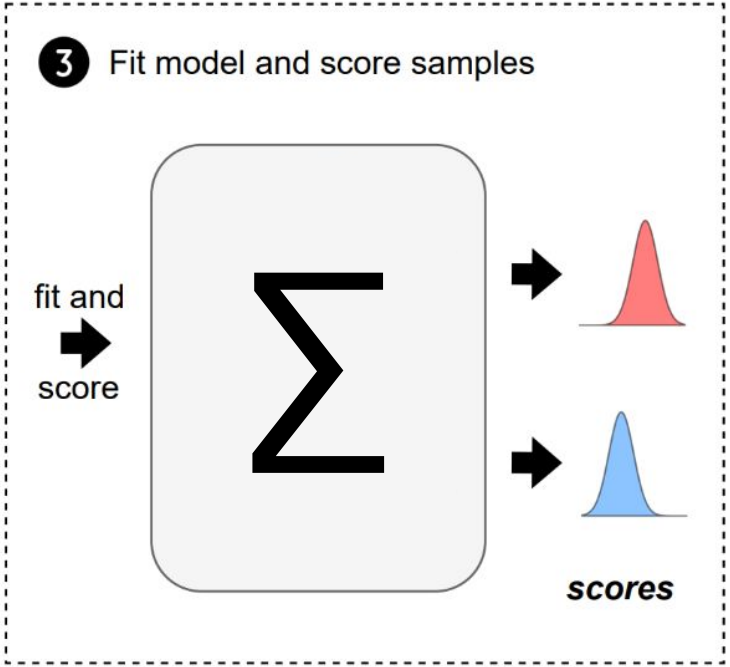


Our improvement: drop the scoring function

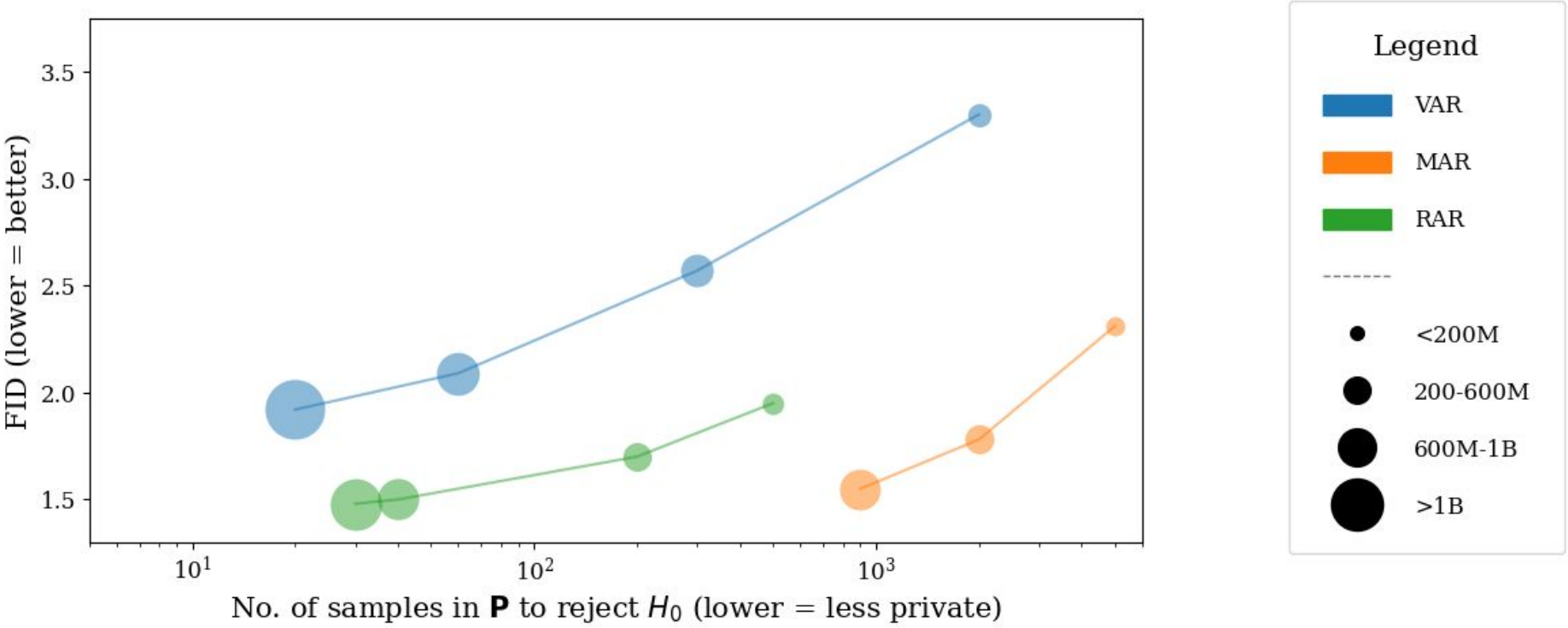


[Dubiński J., et al. 2024]

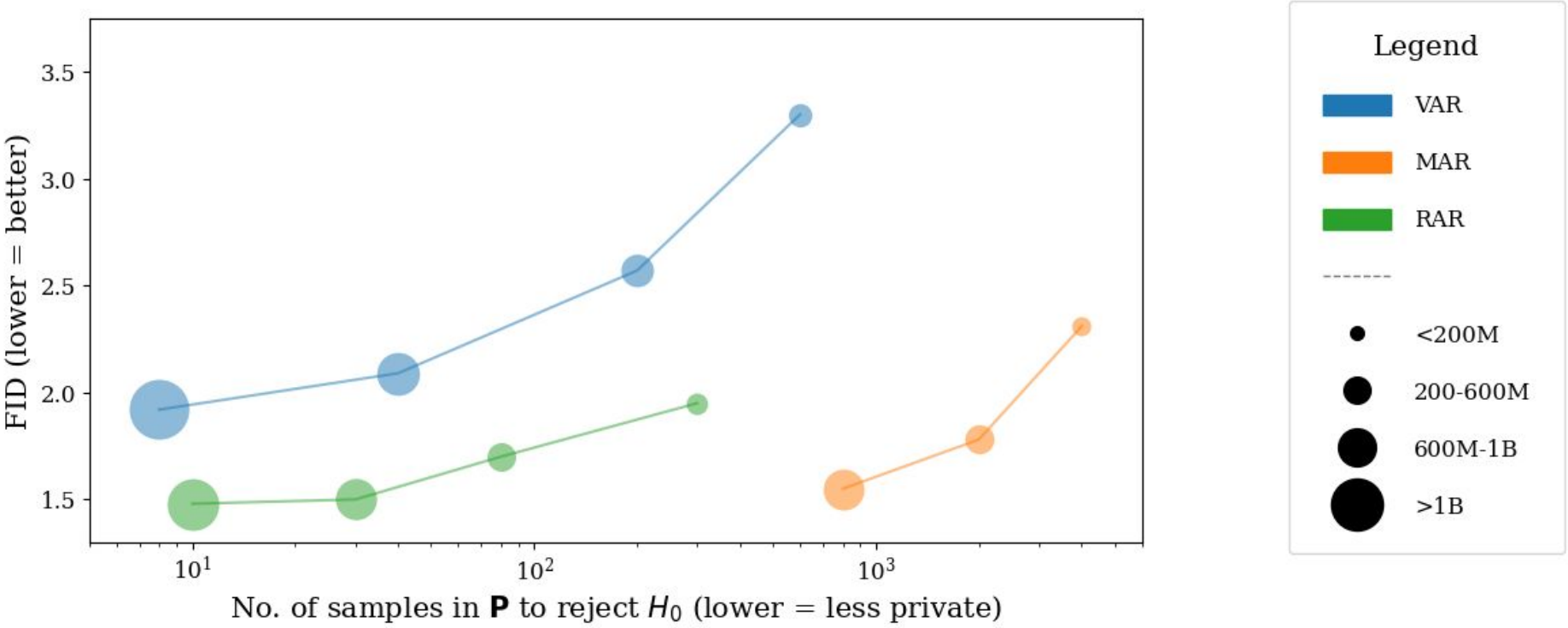
Our improvement: drop the scoring function & replace with sum



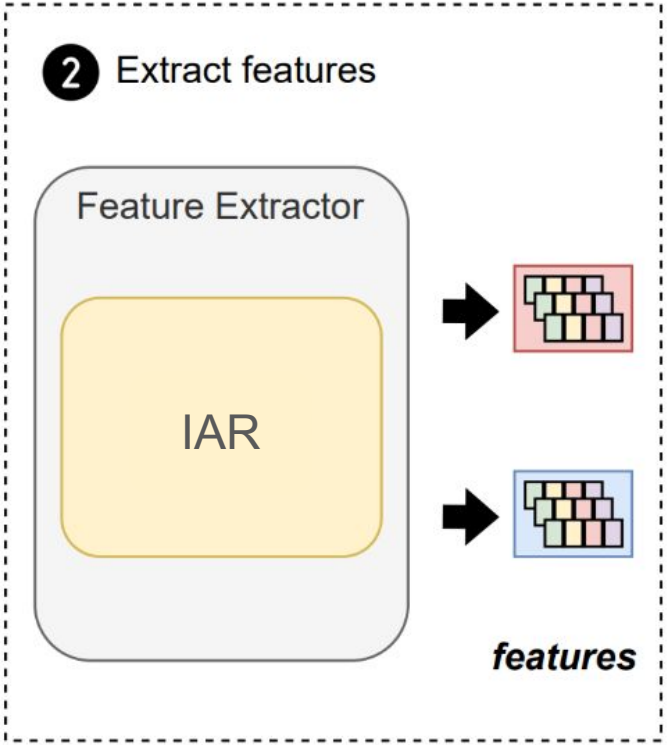
Before


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

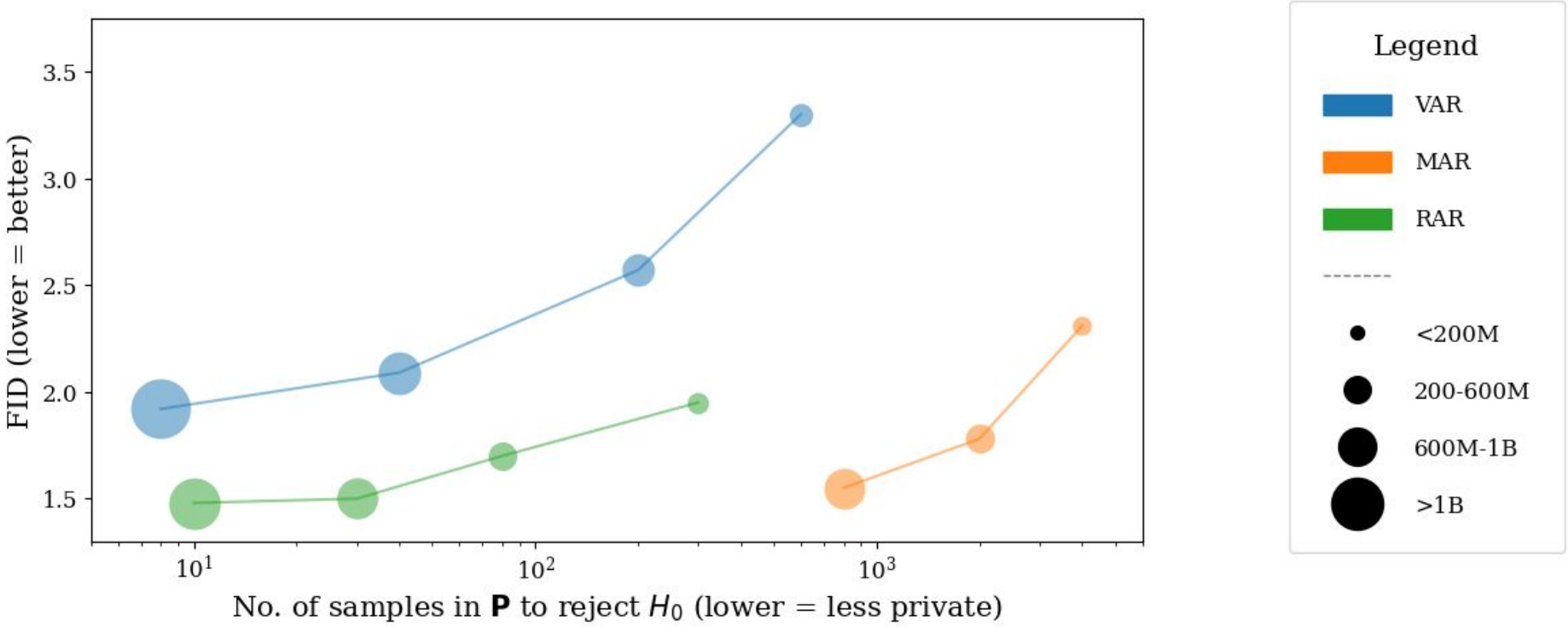
After


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

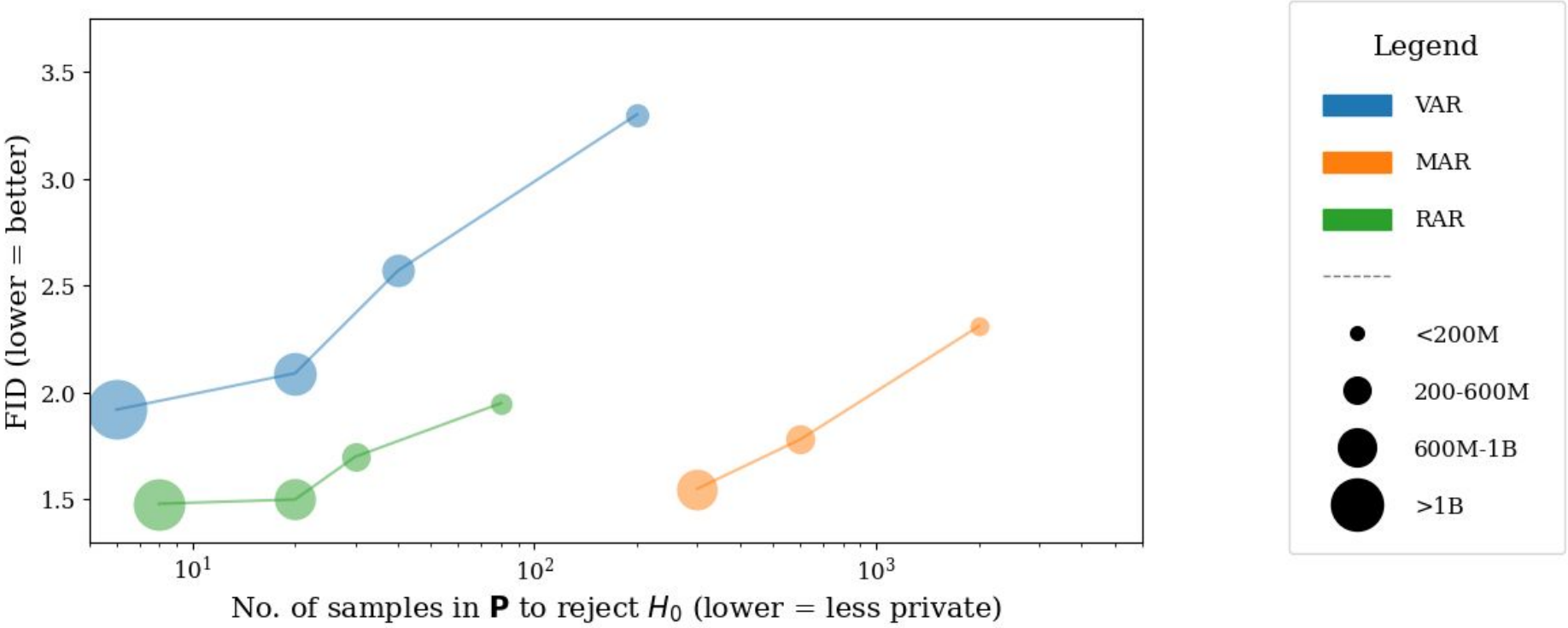
Our improvement: use our MIAs



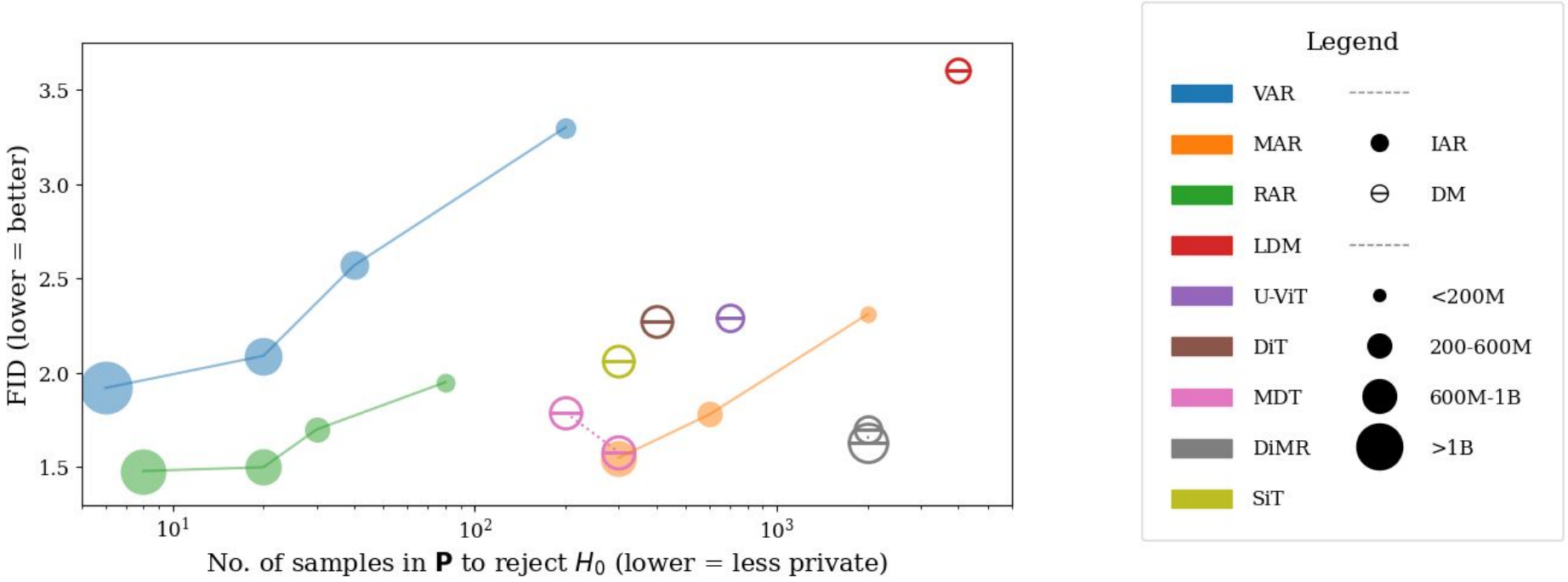
Before


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

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

IARs are more prone to DI than DMs



Memorization

Show that DMs can memorize and generate training data

Say why this is an extreme privacy risk

Show how does it work in DMs (brute-force generation)

Show how does it work in LLMs (find prefix)

Memorization

Training Set

Caption: Living in the light with Ann Graham Lotz

Generated Image

Prompt: Ann Graham Lotz

[Carlini N., et al. 2023]

Memorization in DMs

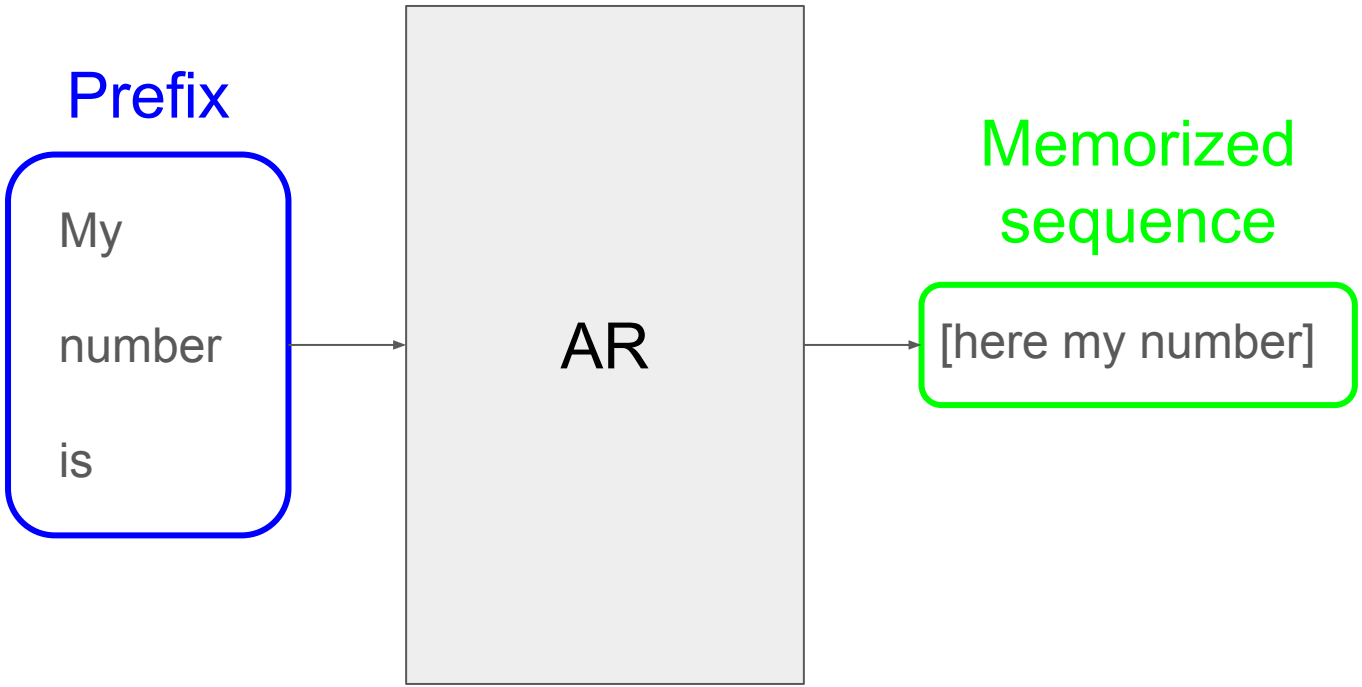
Trigger prompt

Living in the light with Ann Graham Lotz

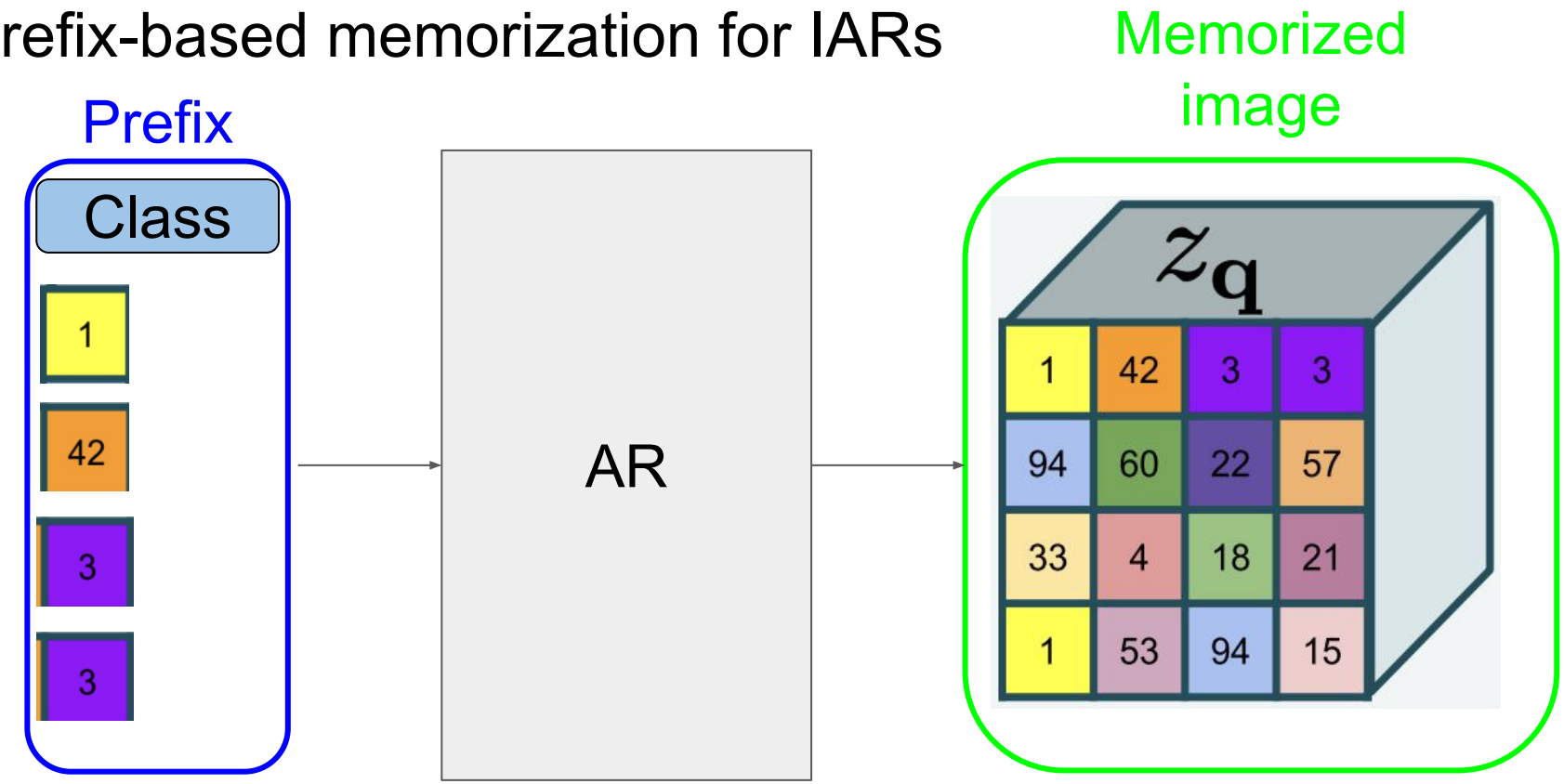
DM

Memorized image

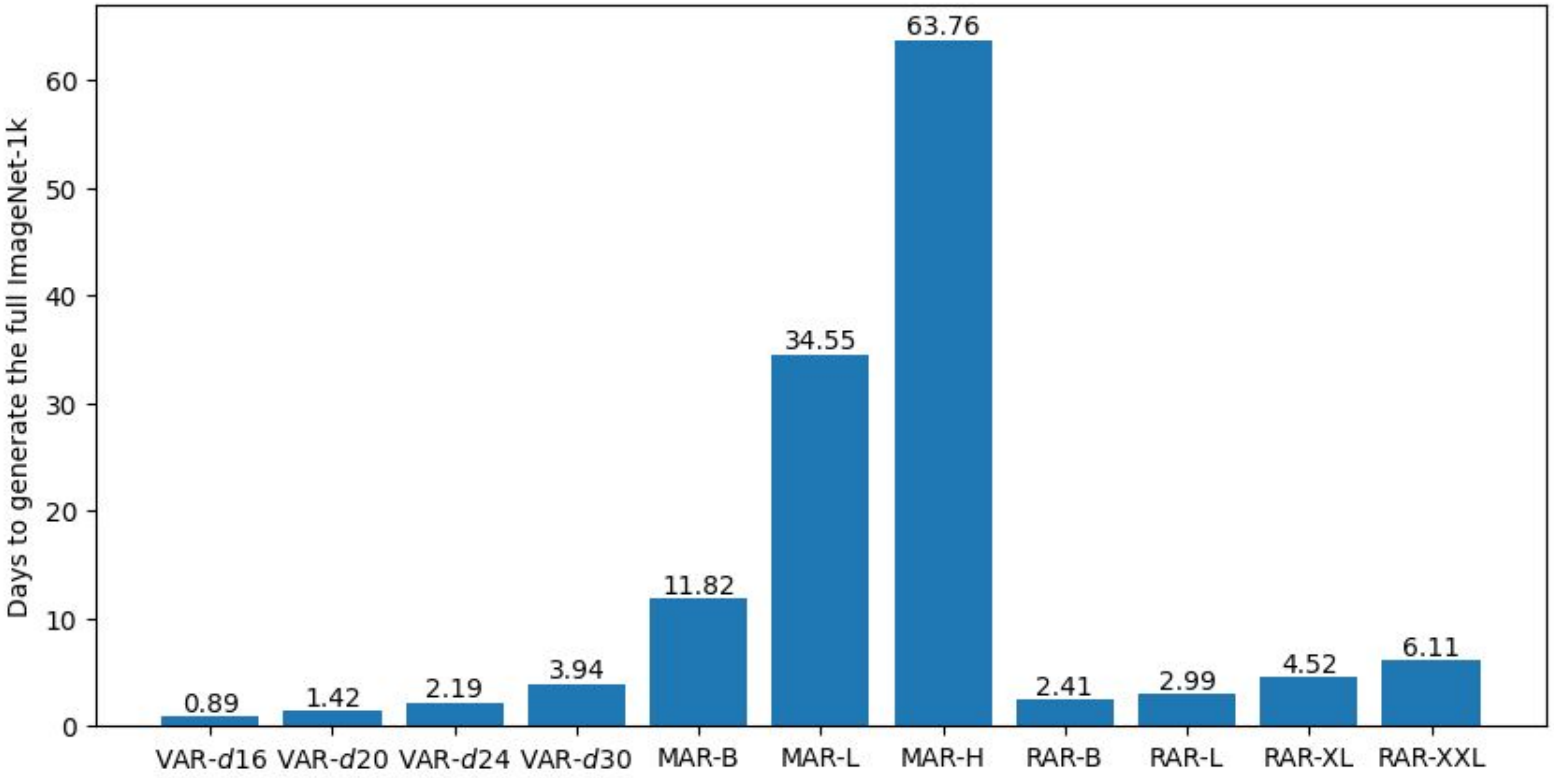
Memorization in LLMs



Prefix-based memorization for IARs



Problem: generation is costly!



Idea: single pass is cheap

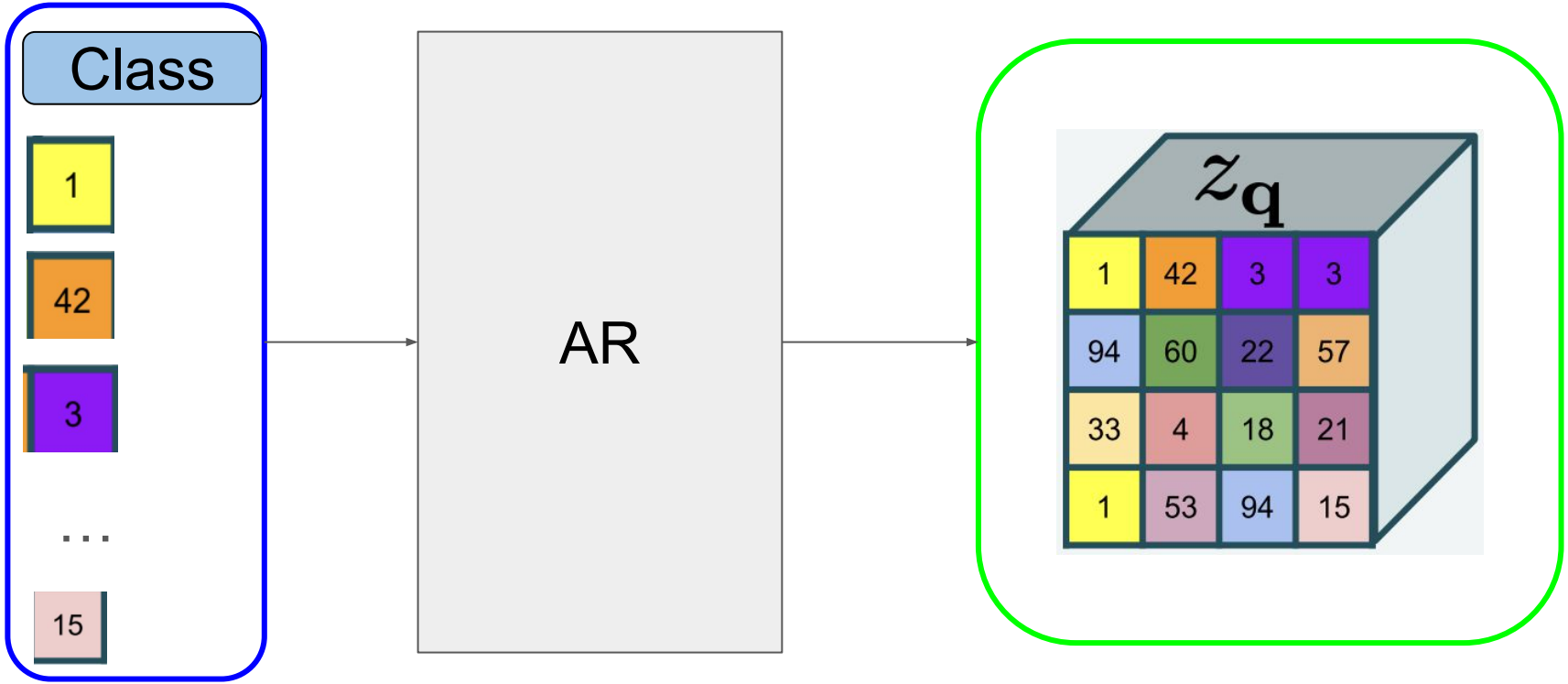
Show how is it done

Highlight the intuition

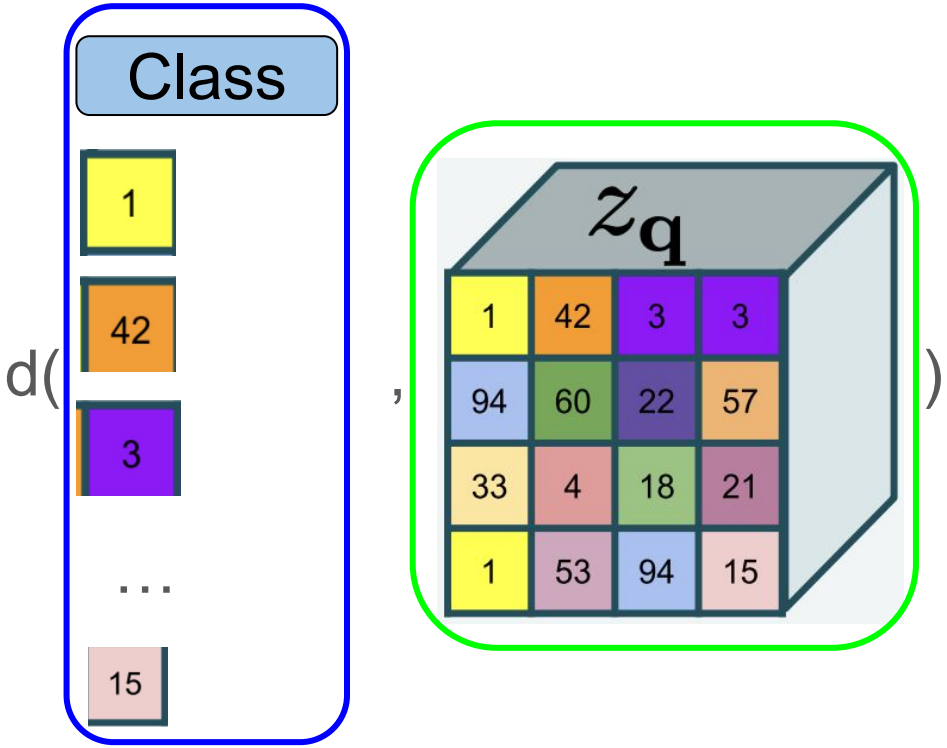
Show the whole procedure of filtering (top-5 per class)

Show relation of the distance to SSCD

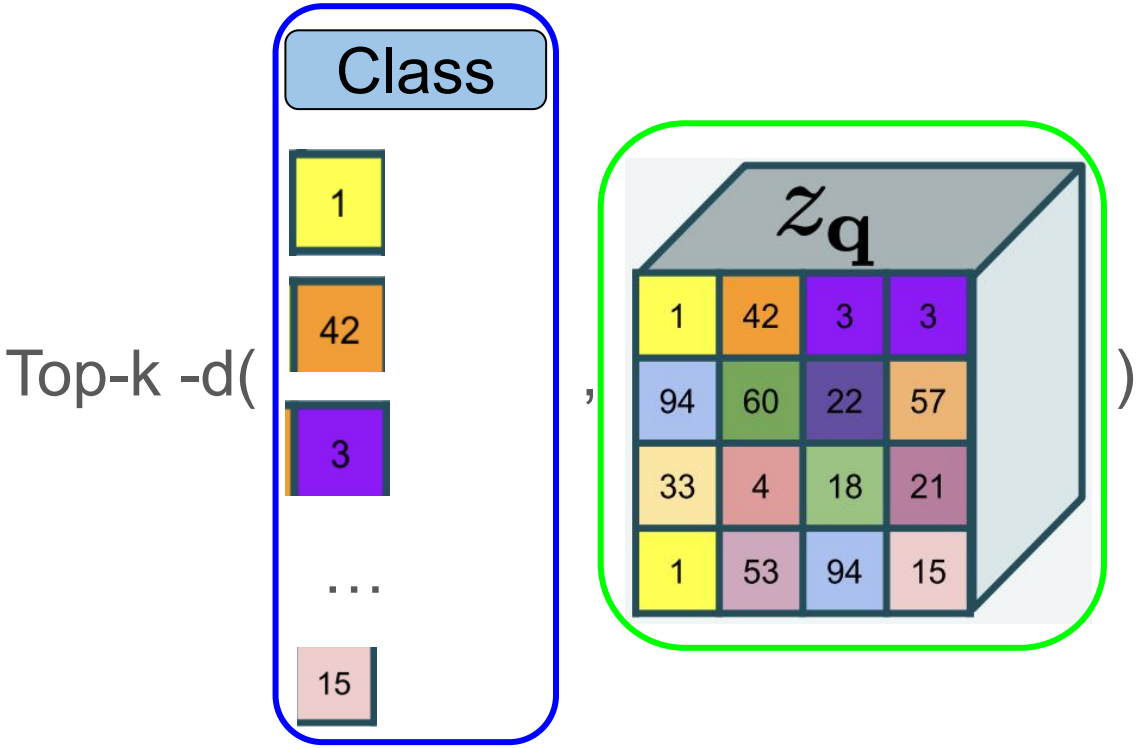
Idea: single pass is cheap



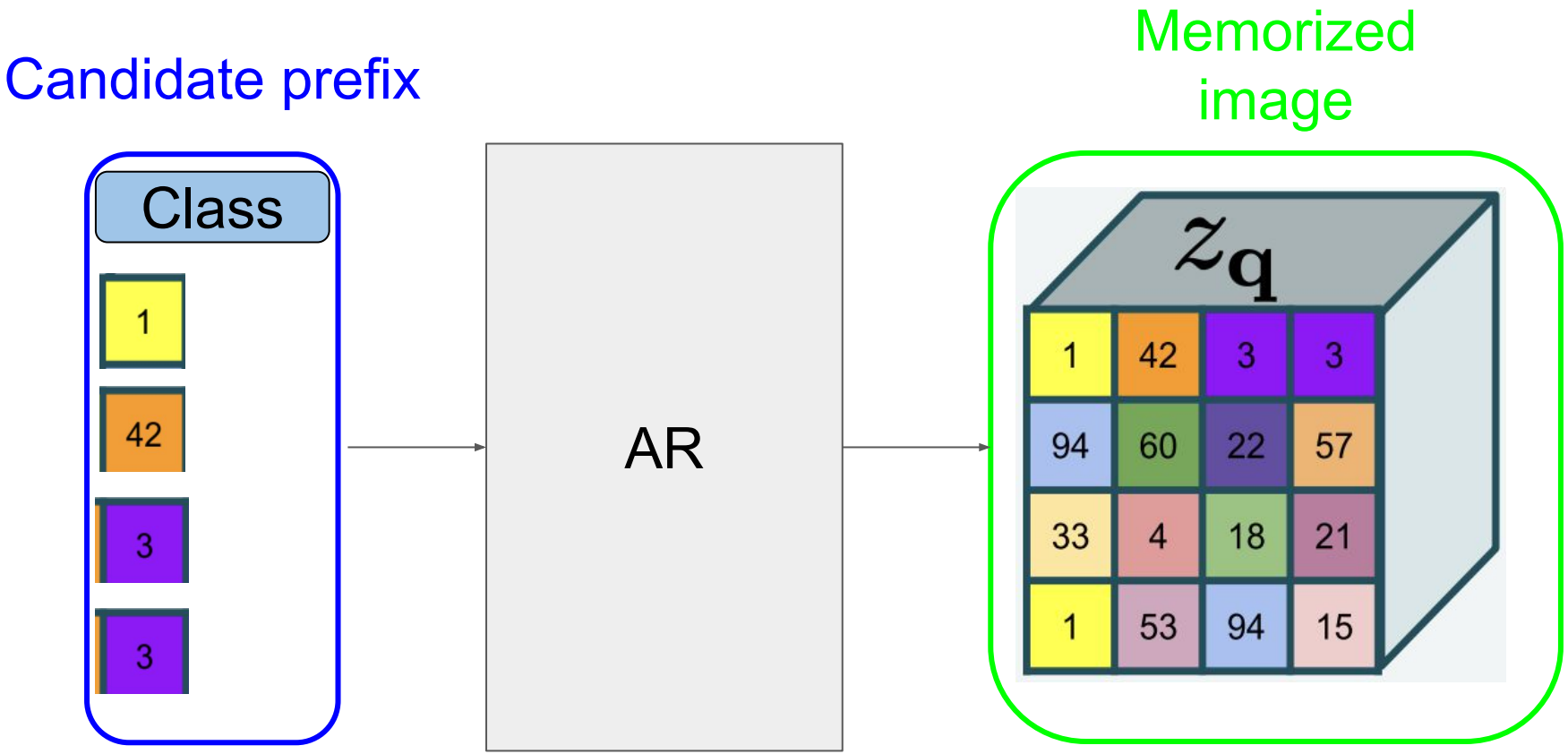
Compare distances



Candidates



Generation


Candidate prefix
Class
1
42
3
3
AR
Memorized image
z_q
1 42 3 3
94 60 22 57
33 4 18 21
1 53 94 15

We successfully extract images from IARs

Here show the main paper image

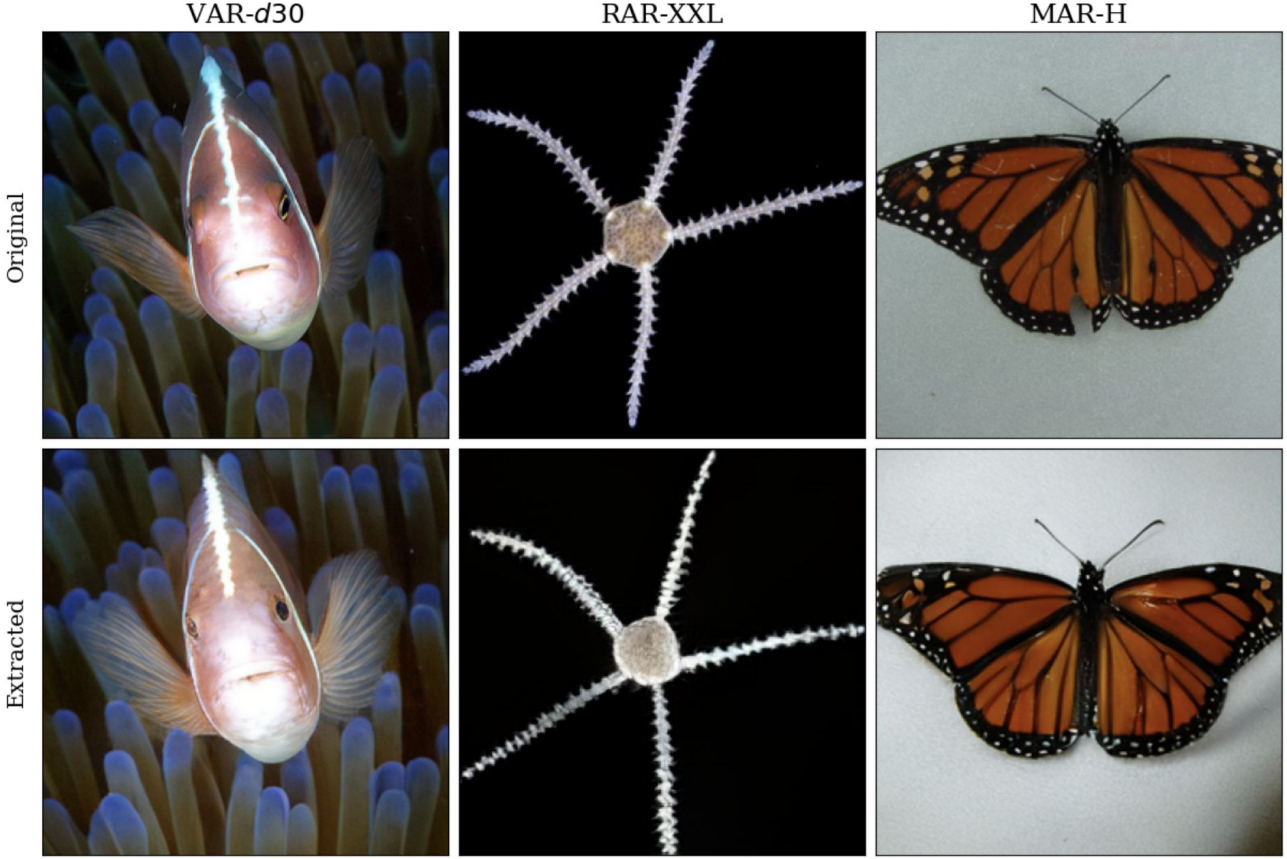
Here show the table from the main paper

Show a single image that is memorized without prefix (just from the class)

Show images that are memorized pairwise by models

No images are shown to be memorized by DMs trained on ImageNet

We successfully extract images from IARs



We successfully extract images from IARs

Model	VAR-d30	RAR-XXL	MAR-H
Images count	**698**	**36**	**5**

Image extracted *without* a prefix

Figure 6: **Image extracted from VAR-*d*30 *without prefix*.** (Left) memorized image, (right) generated image.

Summary

Show again that IARs are cheaper to run -> consequences for development

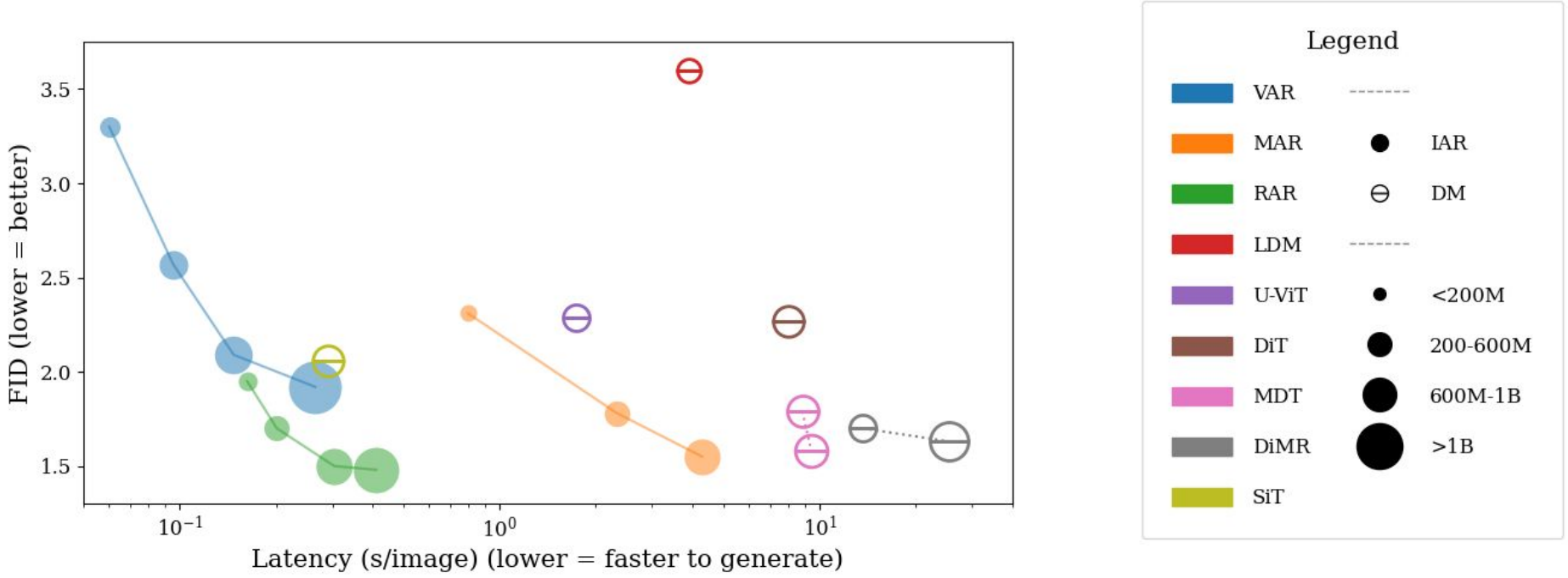
Show that IARs are less private (2 plots) -> consequences for data owners

Show that applying methods from LLMs and DMs naively will make privacy risks underreported

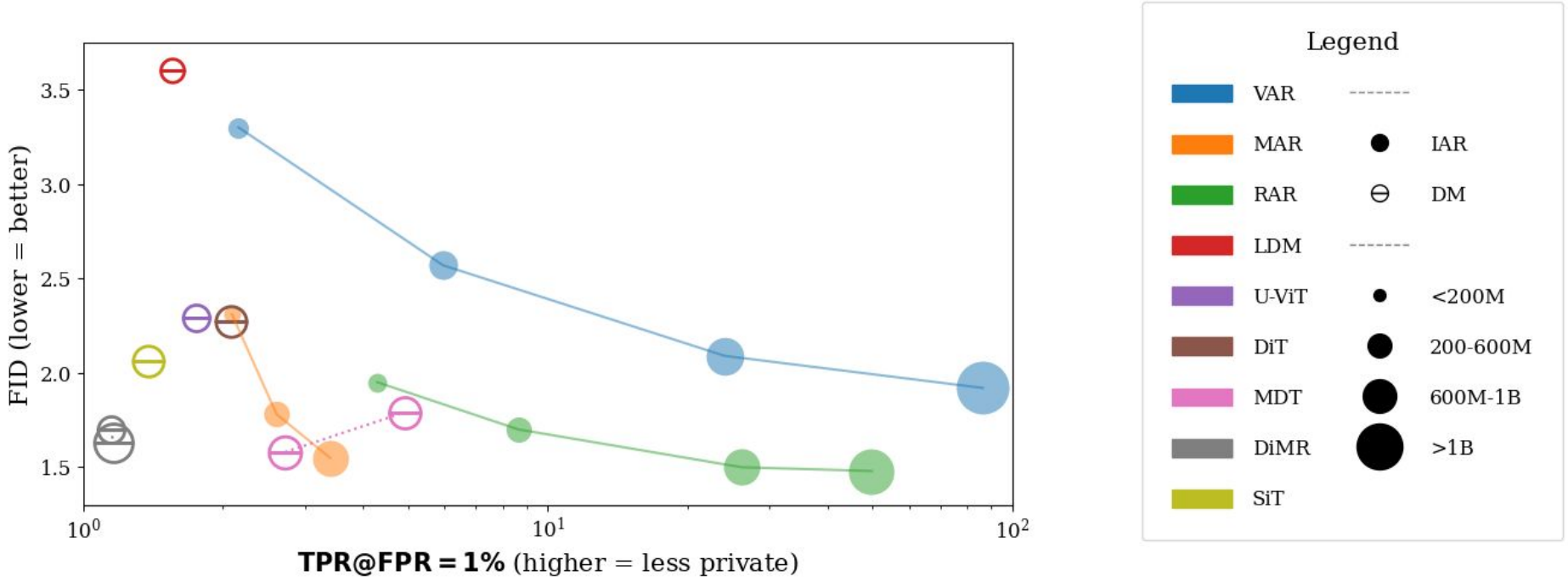
Say that the bigger the model, the more it leaks

Show that MAR leaks the least

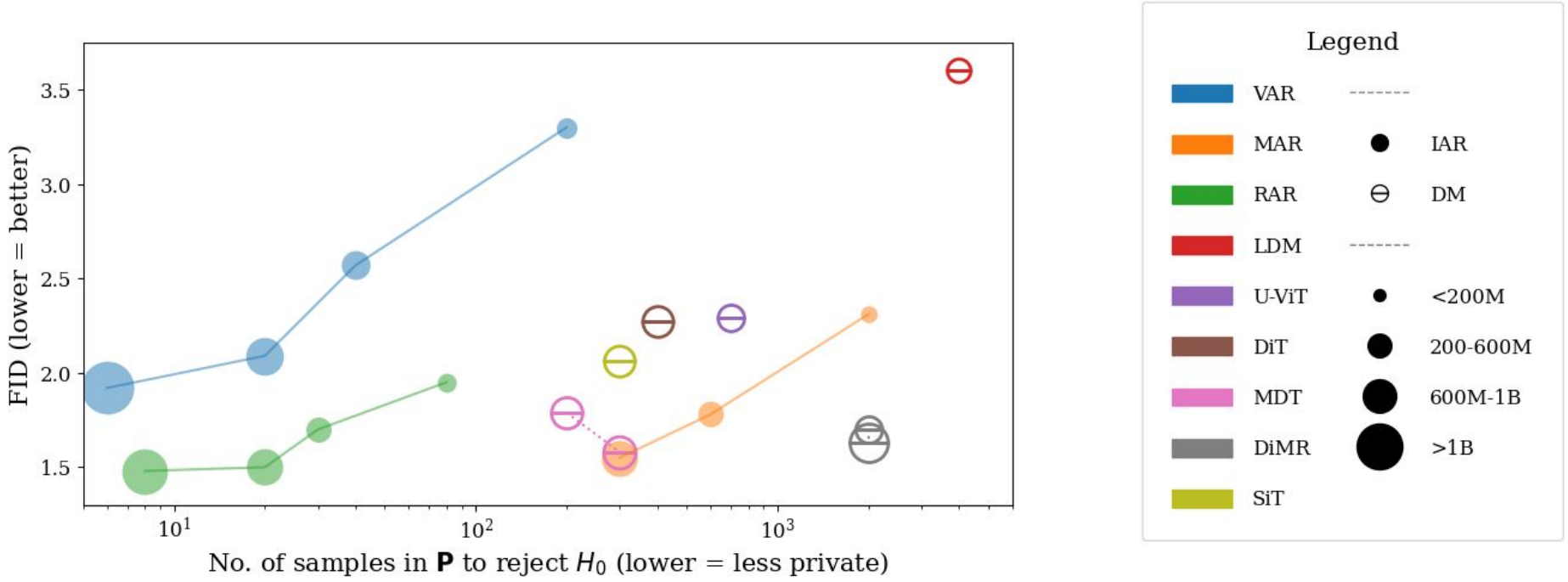
Summary: IARs are cheaper to run



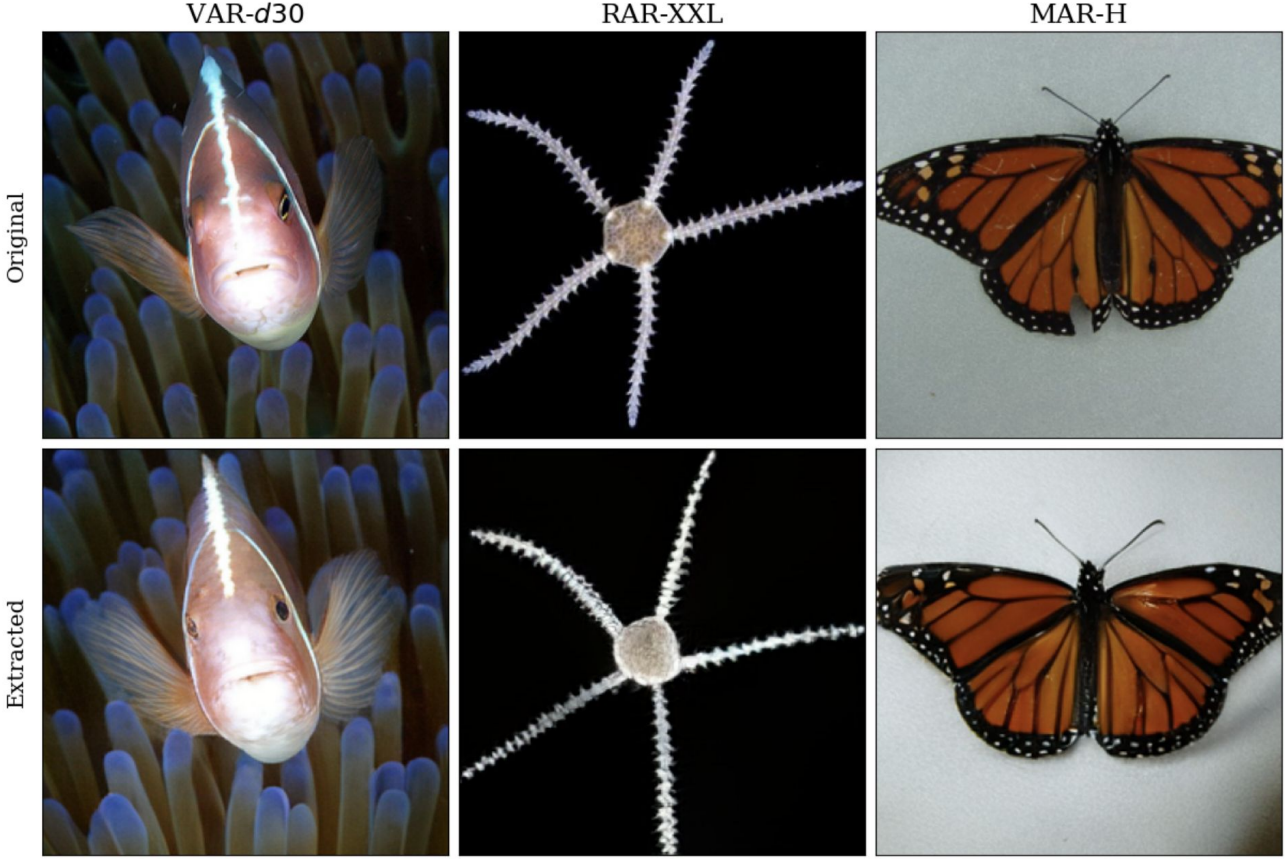
Summary: IARs are less private



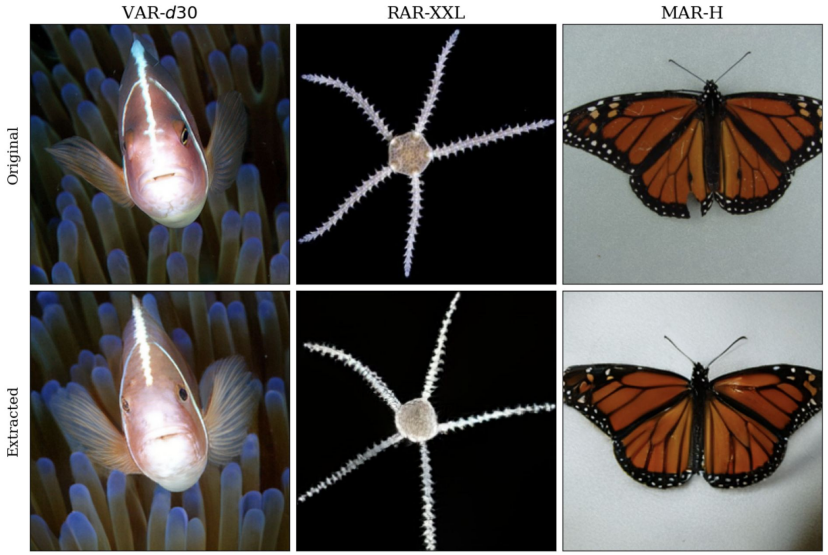
Summary: IARs are less private

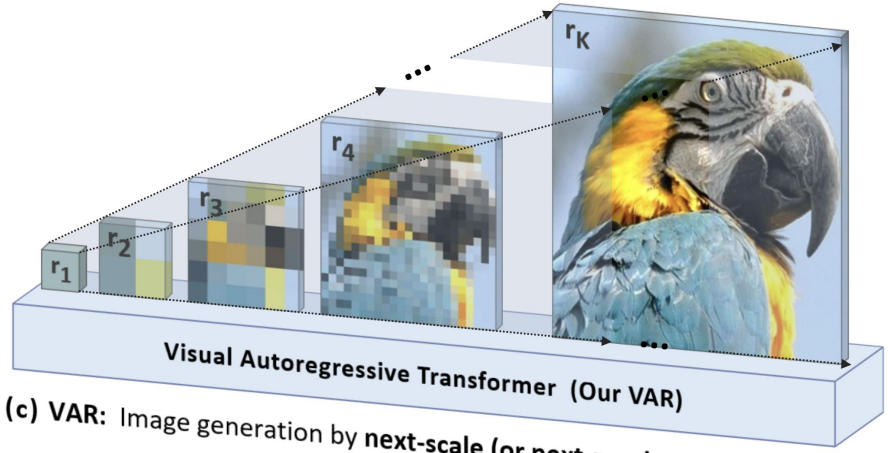


Summary: IARs are less private



Thank you!

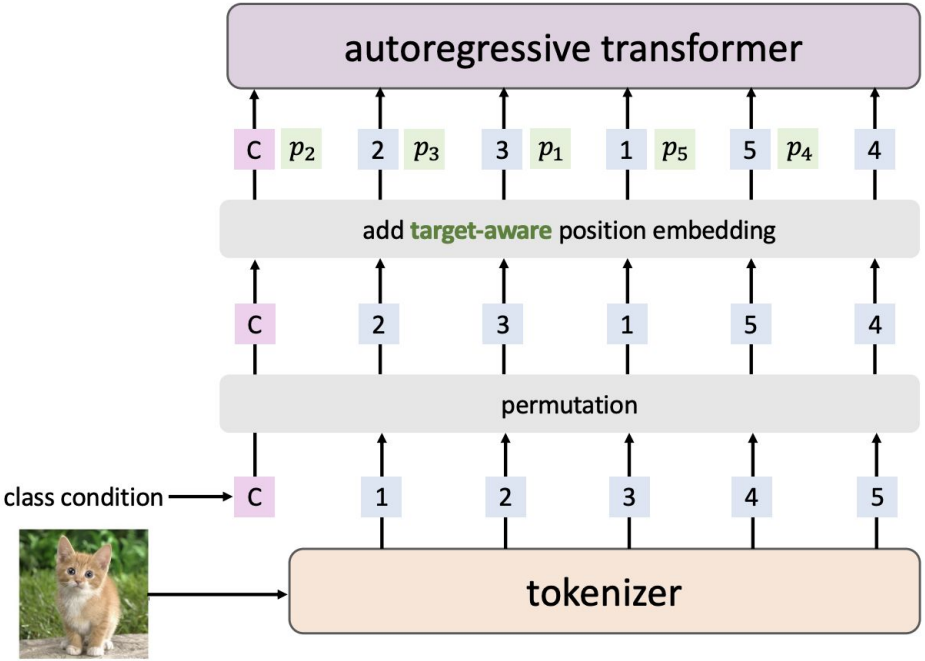




(c) **VAR:** Image generation by **next-scale (or next-resolution)** prediction

Backup slides

RAR's bidirectional attention



(a) how does RAR work w/ target-aware position embedding