CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

mpii max planck institut informatik

GAN-Leaks: A Taxonomy of Membership Inference Attack against Generative Models



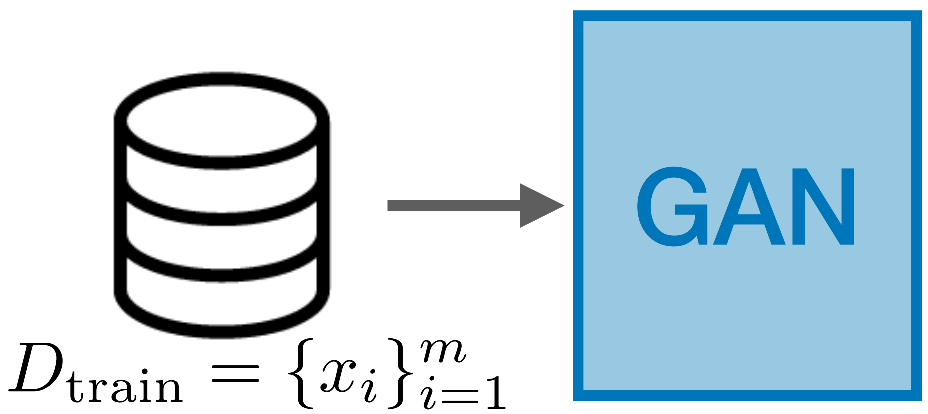

Dingfan Chen[1] Ning Yu[2,3] Yang Zhang[1] Mario Fritz[1]

[1]CISPA Helmholtz Center for Information Security, Germany

[2]Max Planck Institute for Informatics, Germany

[3]University of Maryland, College Park

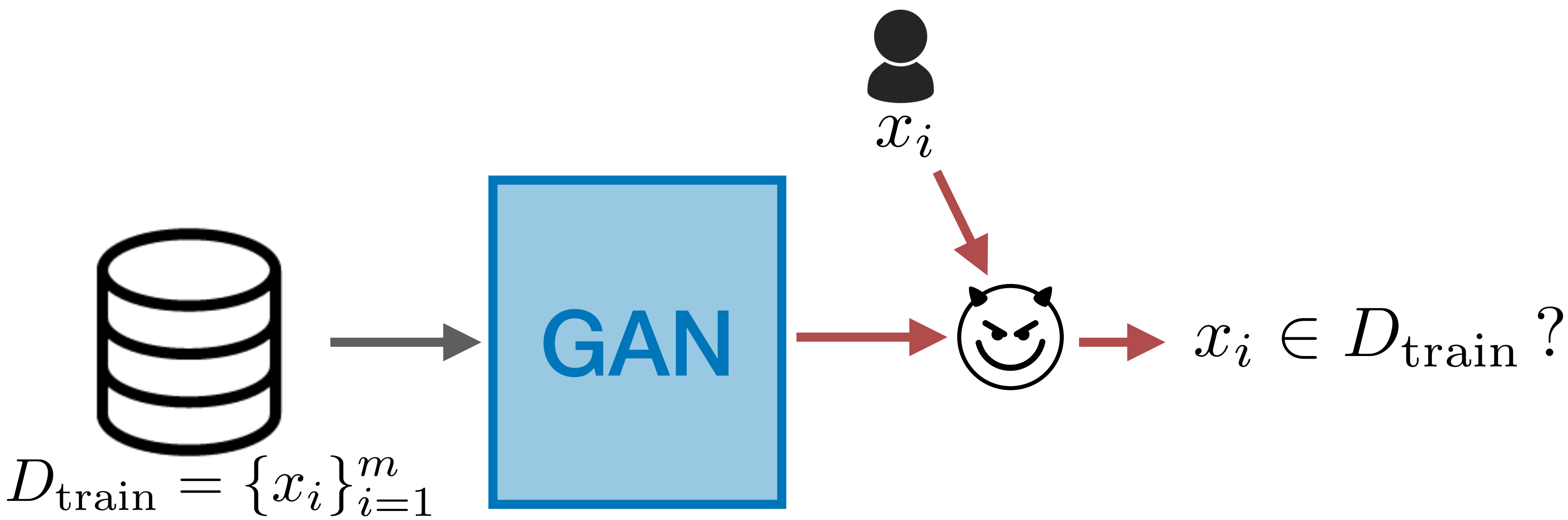
Motivation



- Generative adversarial Networks (GANs)[1] have been largely used on privacy sensitive datasets, e.g., face images and medical records.

[1] Goodfellow et al., “Generative Adversarial Nets”, NIPS 2014

Motivation



- Generative adversarial Networks (GANs)[1] have been largely used on privacy sensitive datasets, e.g., face images and medical records.
- **<u>Our work:</u>** Membership Inference Attack against GANs (whether a query sample x_i has been used to train a GAN model?)
- Crucial to understand and control privacy leakage; provides insights for privacy-preserving data sharing

[1] Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Contributions

- **Taxonomy**
  - categorize attack scenarios against generative models
  - benchmark future research
- **Novel attack models**
  - generic; easy-to-implement; effective; theoretically grounded
- **Extensive evaluation**
  - 3 datasets with diverse data modalities, 5 victim models, 4 attack scenarios …

Taxonomy

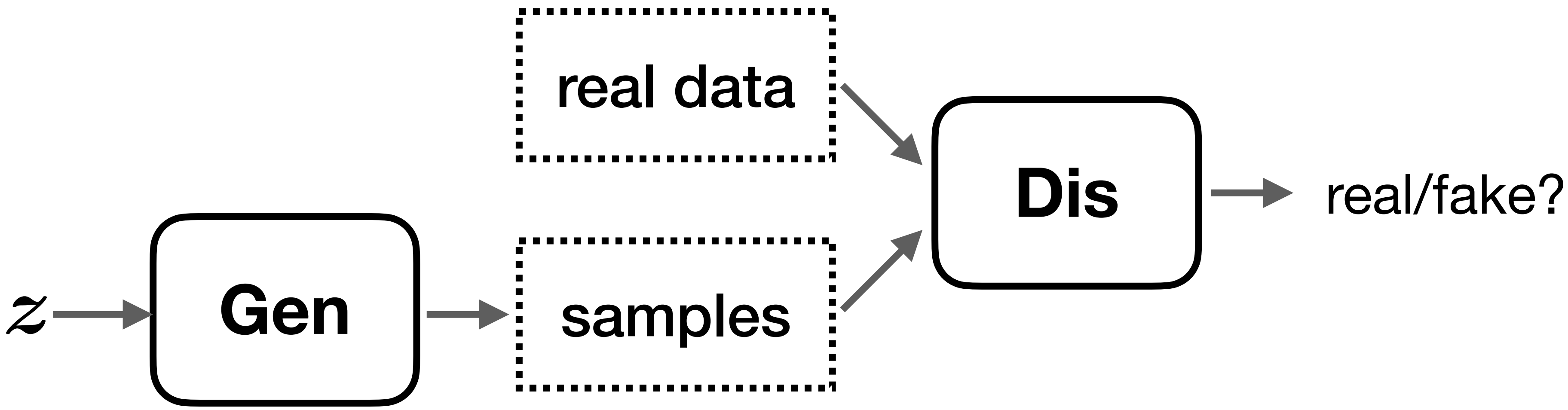
Taxonomy

- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**

Taxonomy

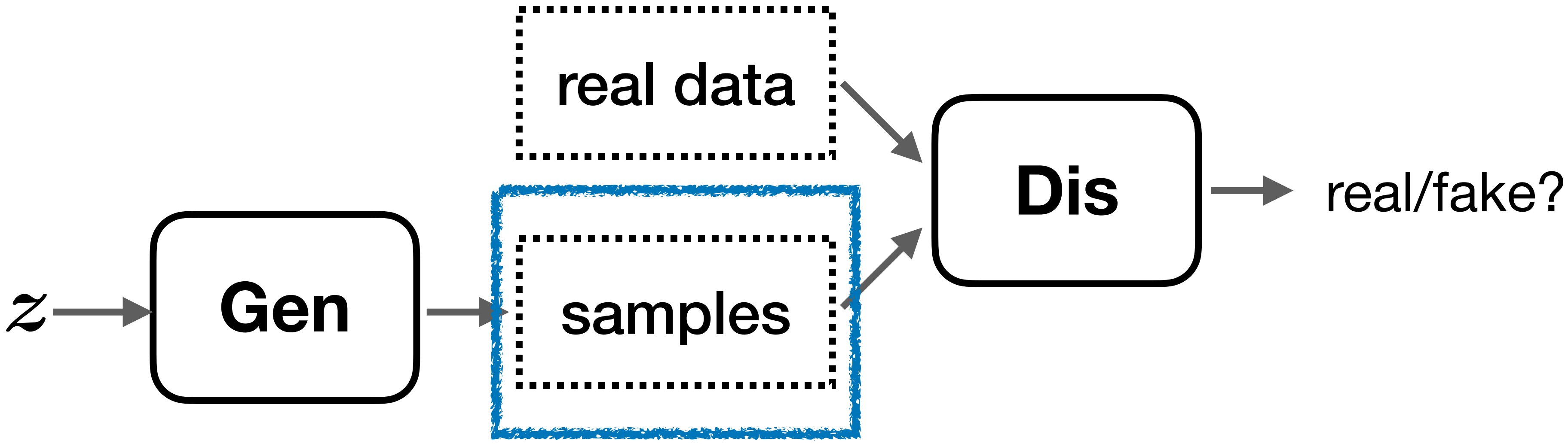
- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**

  (z: latent code; **Gen**: Generator; **Dis**: Discriminator)



Taxonomy

- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**
  (z: latent code; **Gen**: Generator; **Dis**: Discriminator)

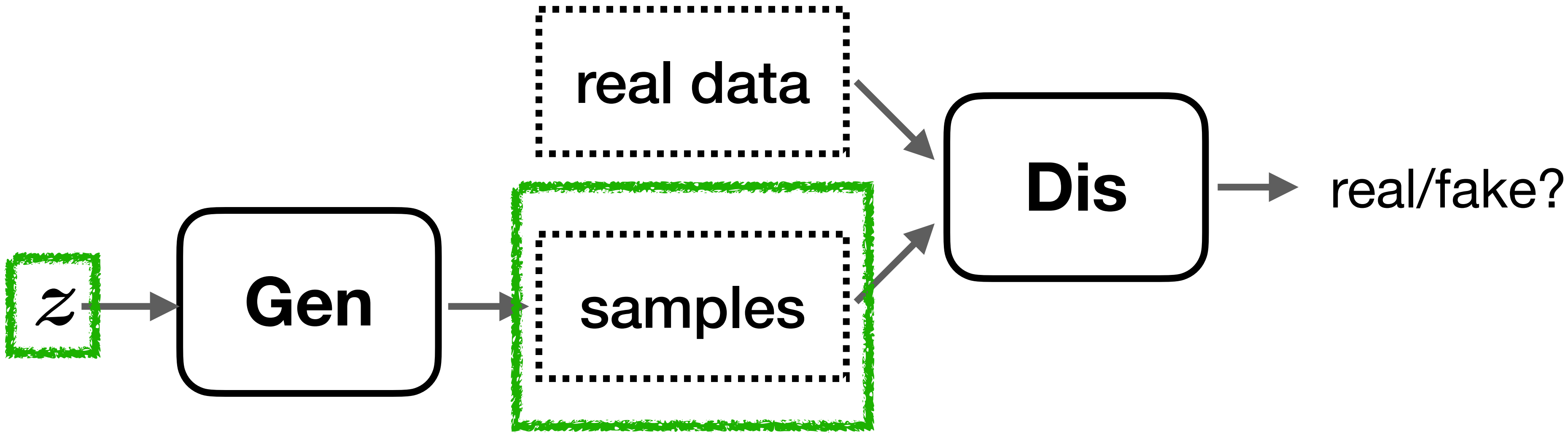


	Latent code	**Generator**	**Discriminator**
(1) Full black-box[1,2]	✗	■	✗

[1] Hayes et al., "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks", PoPETs 2019
[2] Hilprecht et al., "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models", PoPETs 2019

Taxonomy

- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**
  (z: latent code; **Gen**: Generator; **Dis**: Discriminator)

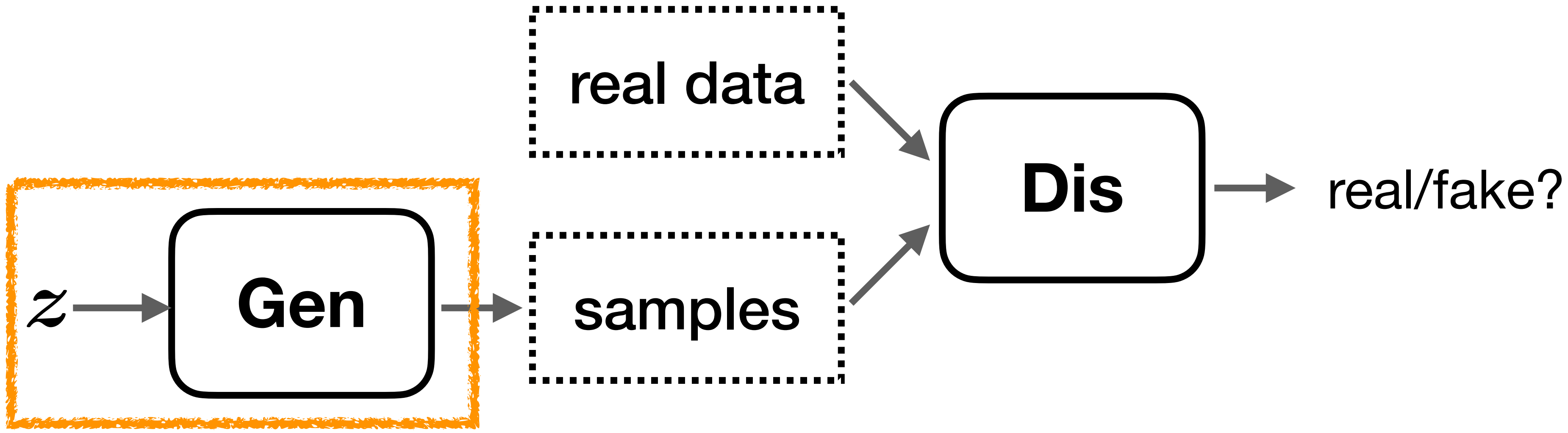


	Latent code	Generator	Discriminator
(1) Full black-box[1,2]	✗	■	✗
(2) Partial black-box	✓	■	✗

[1] Hayes et al., "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks", PoPETs 2019
[2] Hilprecht et al., "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models", PoPETs 2019

Taxonomy

- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**
  (z: latent code; **Gen**: Generator; **Dis**: Discriminator)

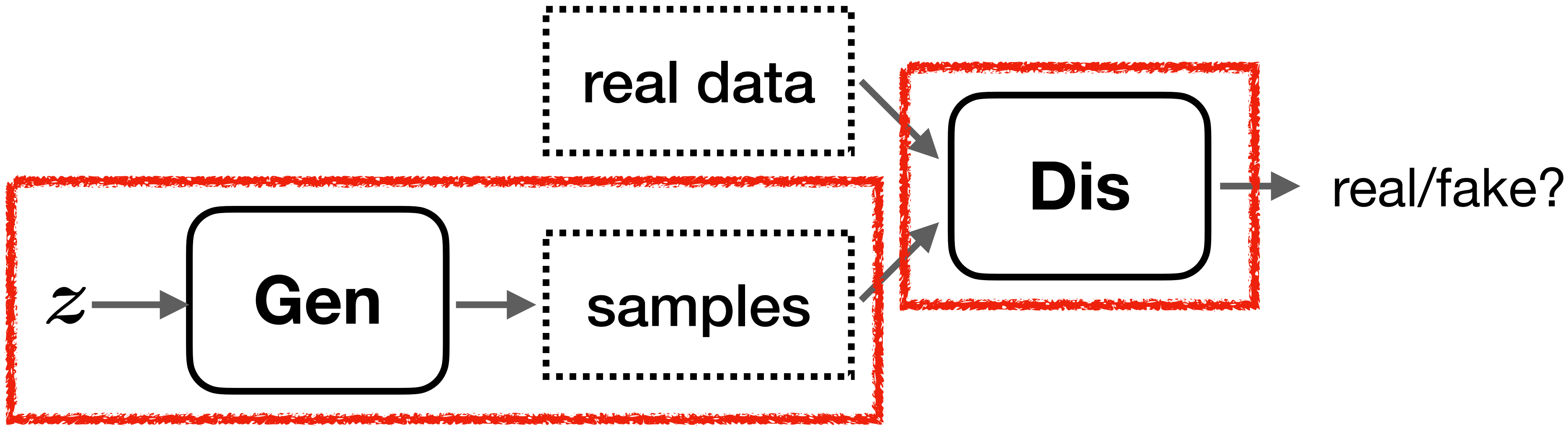


	Latent code	Generator	Discriminator
(1) Full black-box[1,2]	✗	■	✗
(2) Partial black-box	✓	■	✗
(3) White-box	✓	□	✗

[1] Hayes et al., "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks", PoPETs 2019
[2] Hilprecht et al., "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models", PoPETs 2019

Taxonomy

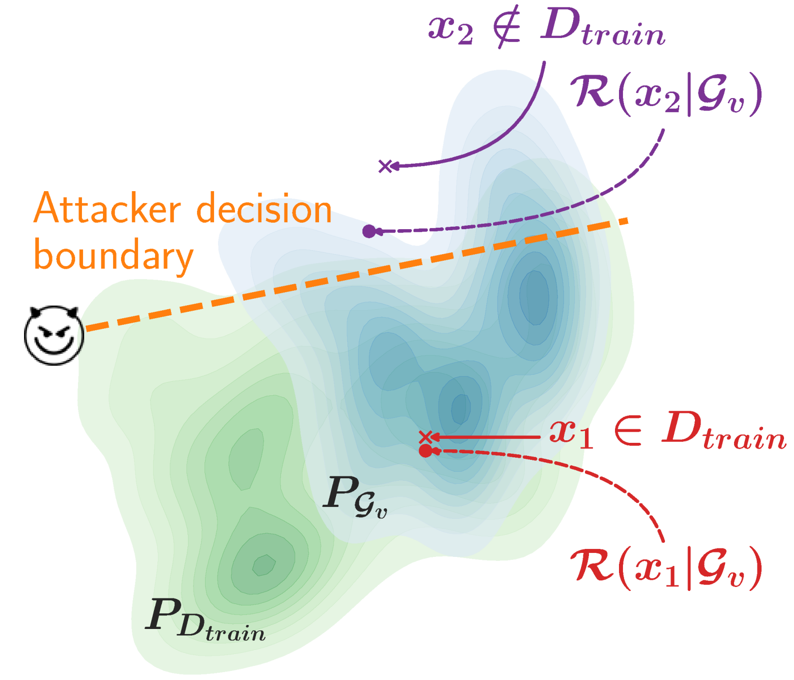
- **white-box □/black-box ■ ?**
- **which GANs' components are accessible?**
  (z: latent code; **Gen**: Generator; **Dis**: Discriminator)



	Latent code	Generator	Discriminator
(1) Full black-box[1,2]	✗	■	✗
(2) Partial black-box	✓	■	✗
(3) White-box	✓	□	✗
(4) Accessible full model[1]	✓	□	✓

[1] Hayes et al., "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks", PoPETs 2019
[2] Hilprecht et al., "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models", PoPETs 2019

Method


$x_2 \notin D_{train}$
$\mathcal{R}(x_2|\mathcal{G}_v)$
Attacker decision boundary
$x_1 \in D_{train}$
$P_{\mathcal{G}_v}$
$\mathcal{R}(x_1|\mathcal{G}_v)$
$P_{D_{train}}$

Method

- **Insight:**
  Smaller reconstruction error for training set data.

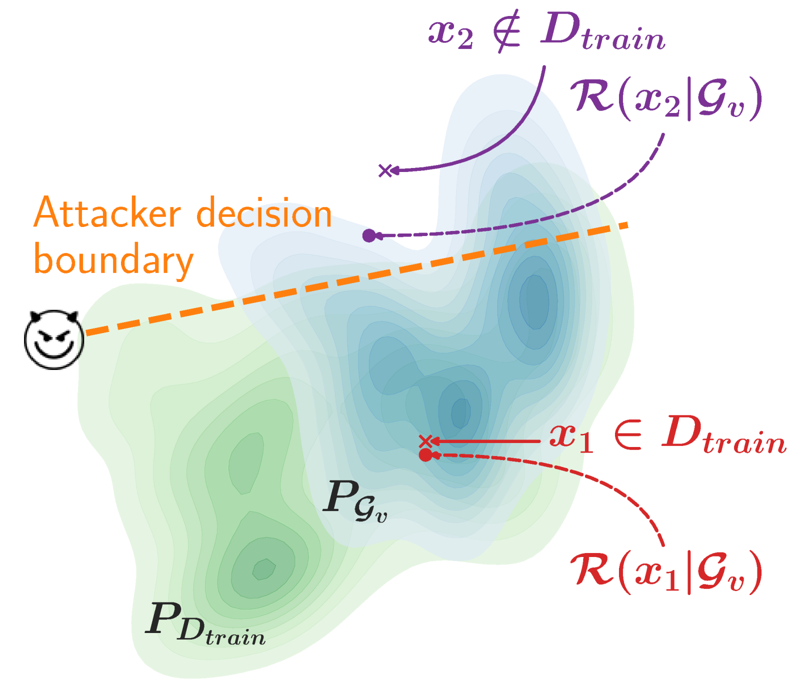
- **Generic Model:**
  optimization problem

$$\mathcal{R}(x|\mathcal{G}_v) = \mathcal{G}_v(z^*)$$

$$z^* = \underset{z}{\operatorname{argmin}} \quad L\big(x, \mathcal{G}_v(z)\big)$$

- **Objective:**

$$\underset{z}{\text{minimize}} \quad L\big(x, \mathcal{G}_v(z)\big) = \lambda_1 L_2\big(x, \mathcal{G}_v(z)\big) + \lambda_2 L_{\text{lpips}}\big(x, \mathcal{G}_v(z)\big) + \lambda_3 L_{\text{reg}}(z)$$

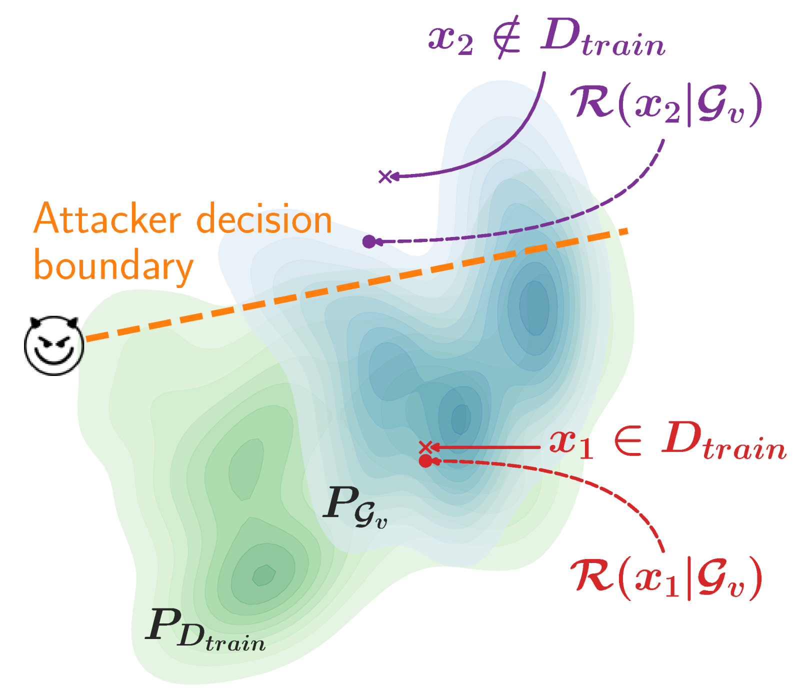


Method

- **Insight:**
  Smaller reconstruction error for training set data.

- **Generic Model:**
  optimization problem

$$\mathcal{R}(x|\mathcal{G}_v) = \mathcal{G}_v(z^*)$$
$$z^* = \underset{z}{\operatorname{argmin}} \quad L(x, \mathcal{G}_v(z))$$



- **Objective:**

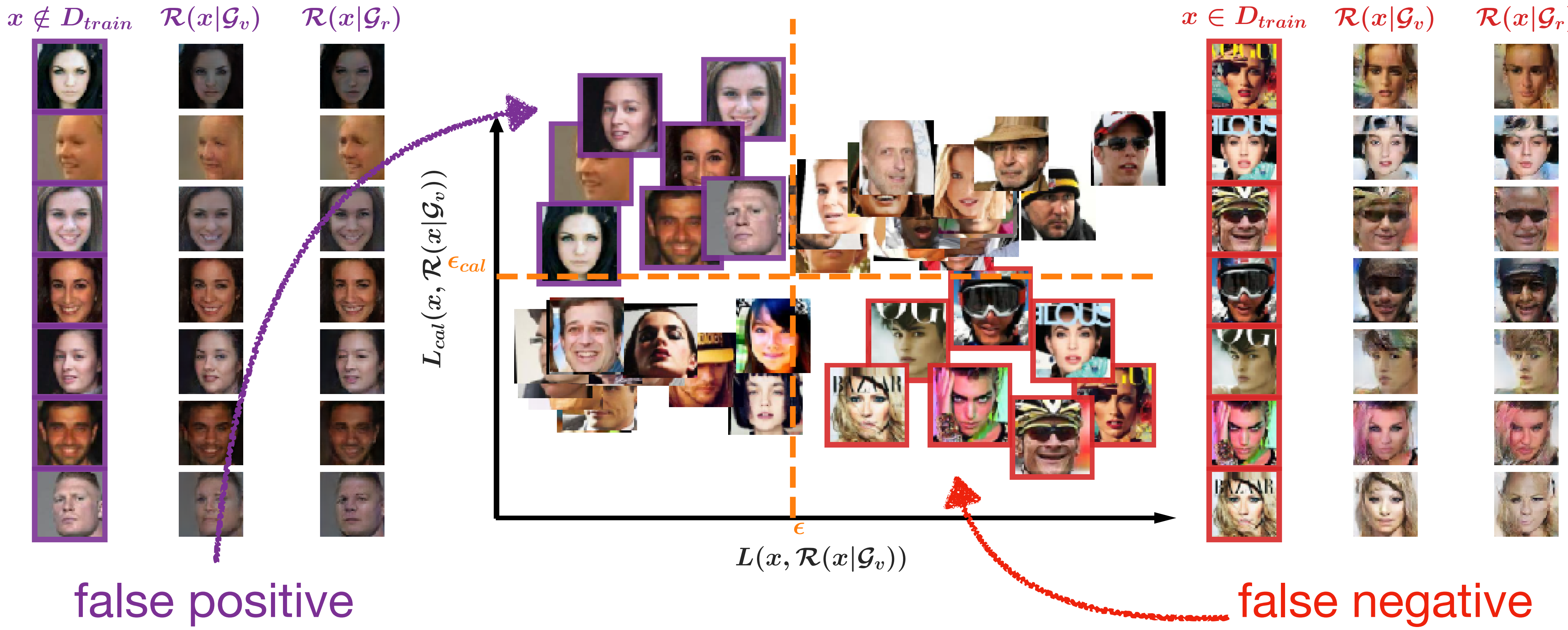
$$\underset{z}{\operatorname{minimize}} \quad L(x, \mathcal{G}_v(z)) = \lambda_1 L_2(x, \mathcal{G}_v(z)) + \lambda_2 L_{\text{lpips}}(x, \mathcal{G}_v(z)) + \lambda_3 L_{\text{reg}}(z)$$

- **Different settings:**

(1) Full black-box	KNN search
(2) Partial black-box	Powell's conjugate direction method
(3) White-box	L-BFGS quasi-Newton method

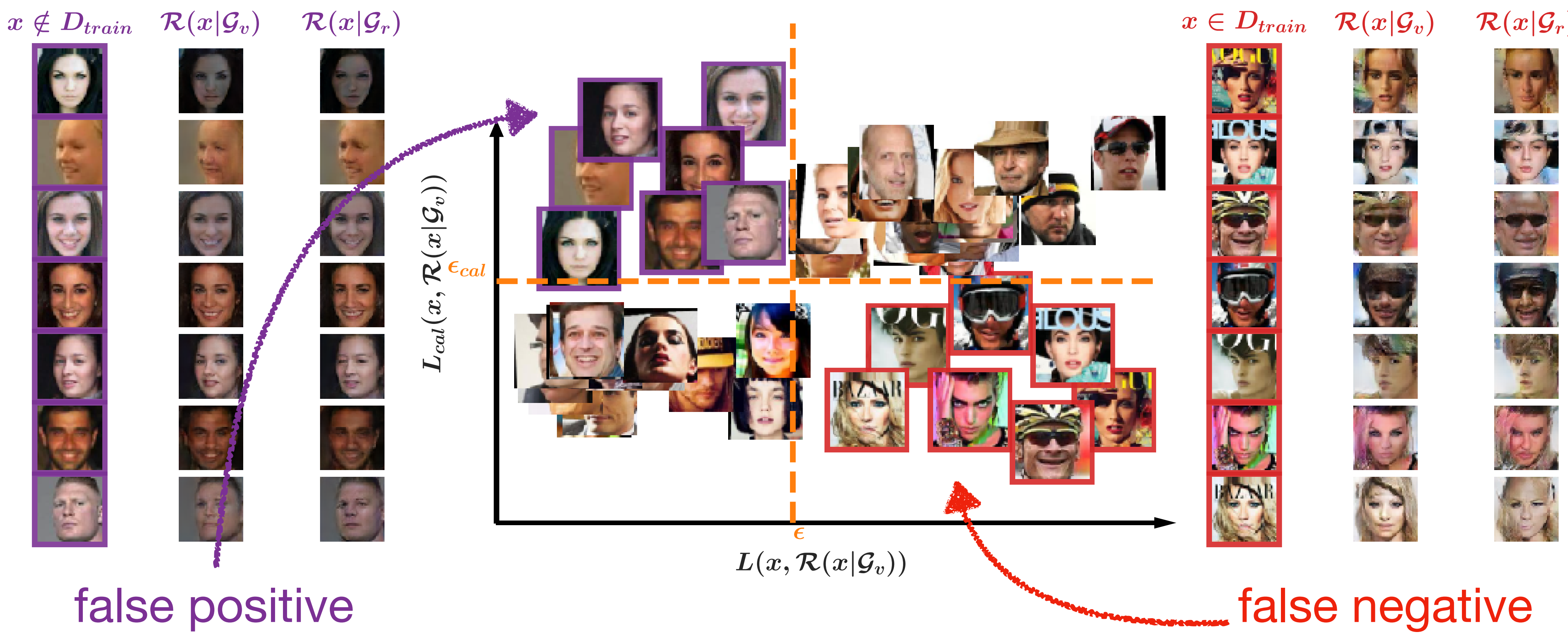
Method

- **Observe:**
  Reconstruction error affected by the appearance



Method

- **Observe:**
  Reconstruction error affected by the appearance

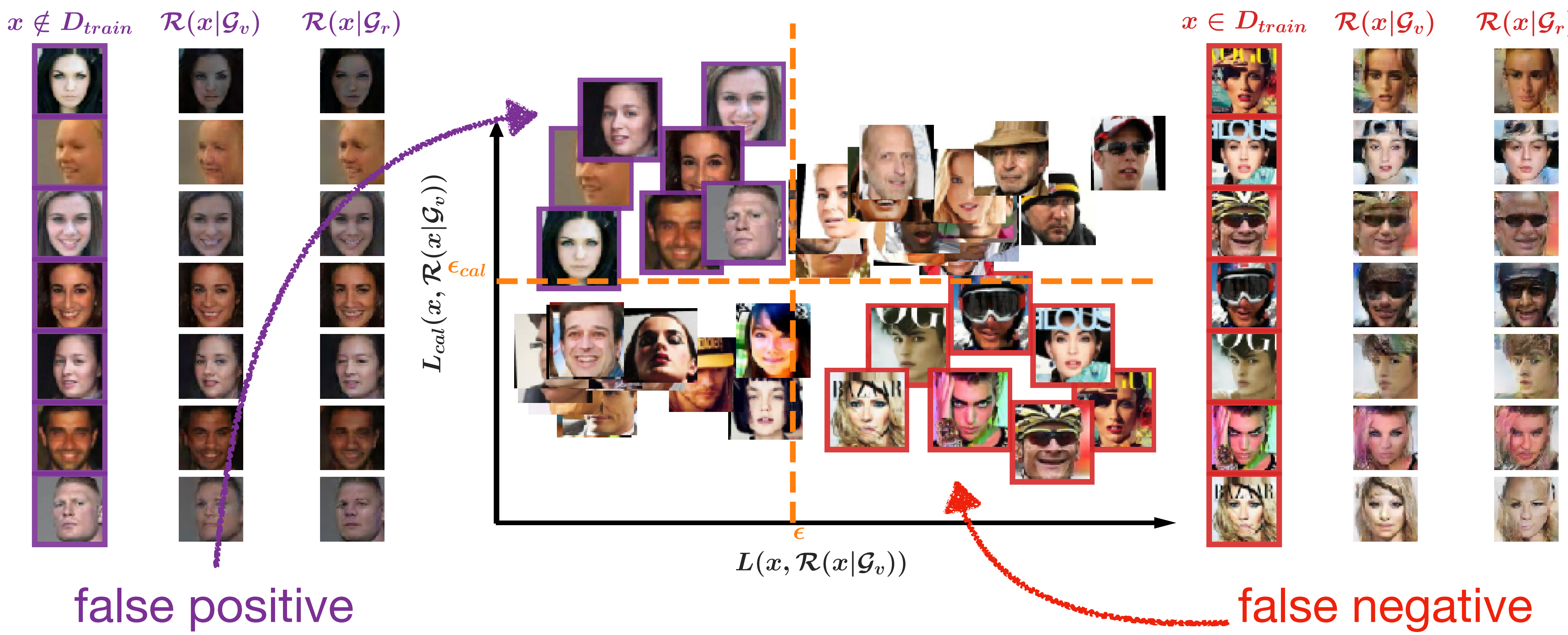


- **Solution:** Attack Calibration

$$L_{\text{cal}}\big(x, \mathcal{R}(x|\mathcal{G}_v)\big) = \underbrace{L\big(x, \mathcal{R}(x|\mathcal{G}_v)\big)}_{\text{victim model}} - \underbrace{L\big(x, \mathcal{R}(x|\mathcal{G}_r)\big)}_{\text{reference model}}$$

Method

- **<u>Observe:</u>**
  Reconstruction error affected by the appearance



- **<u>Solution:</u>** Attack Calibration

$$L_{\text{cal}}\big(x, \mathcal{R}(x|\mathcal{G}_v)\big) = \underbrace{L\big(x, \mathcal{R}(x|\mathcal{G}_v)\big)}_{\text{victim model}} - \underbrace{L\big(x, \mathcal{R}(x|\mathcal{G}_r)\big)}_{\text{reference model}}$$

- **<u>Theory:</u>** near-optimal under a Bayesian perspective[1]

[1] Sablayrolles et al., "White-box vs Black-box: Bayes Optimal Strategies for Membership Inference", ICML 2019

Experiments

- **3 Datasets**
  - **Face:** CelebA
  - **Location:** Instagram
  - **Medical:** MIMIC III

- **5 GAN models**
  - PGGAN, WGANGP, DCGAN, VAEGAN, MedGAN

- **2 Baselines**
  - LOGAN[1], MC[2]

- **Systematic Analysis**
  - dataset size, model architectures, attack settings, defense…

- **Metric**

  AUC (Area Under the ROC Curve)

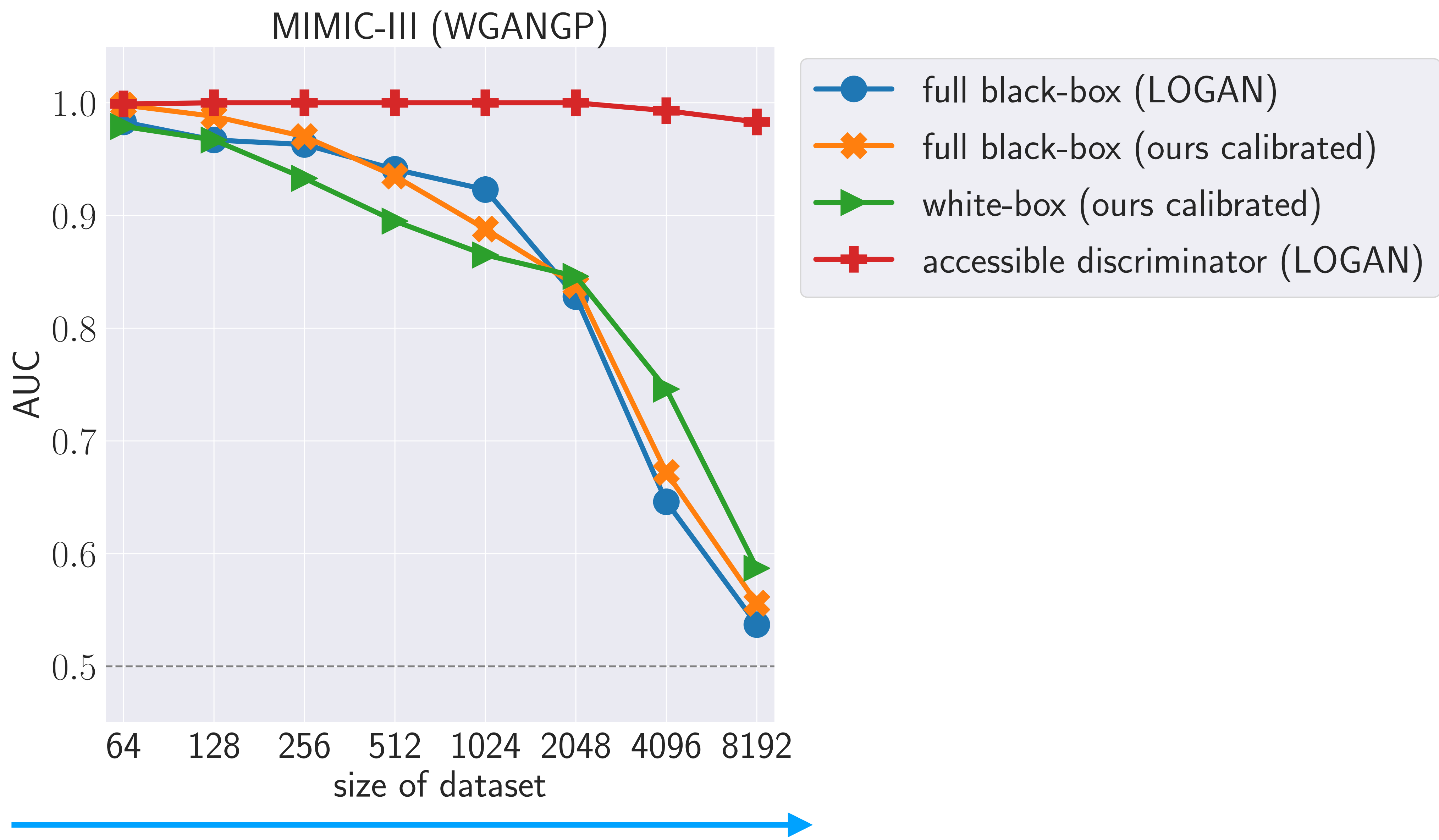
  larger AUC → **better** attacker

[1] Hayes et al., “LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks”, PoPETs 2019

[2] Hilprecht et al., “Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models”, PoPETs 2019

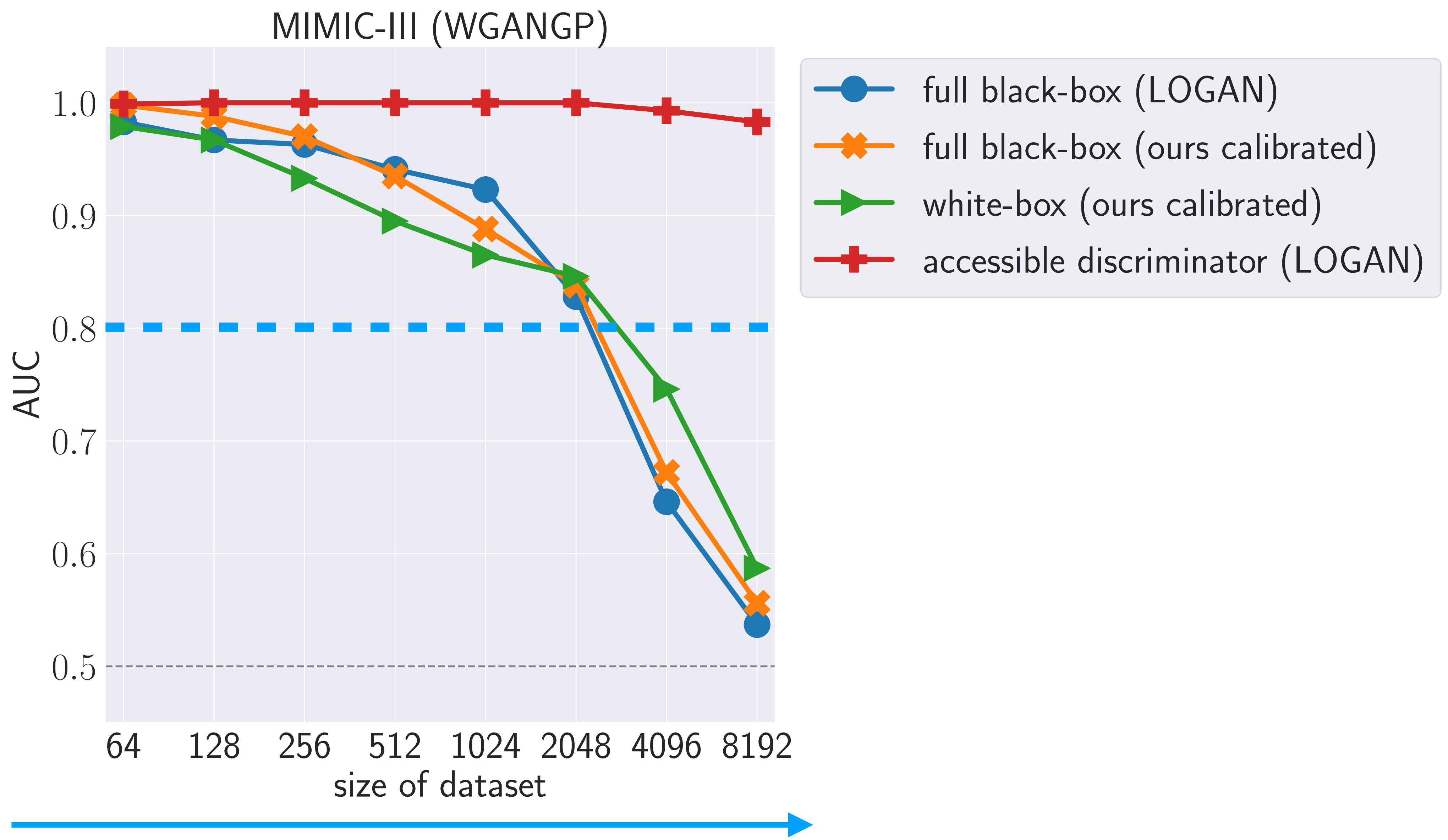
Experiments

- Attack effectiveness —Dataset size



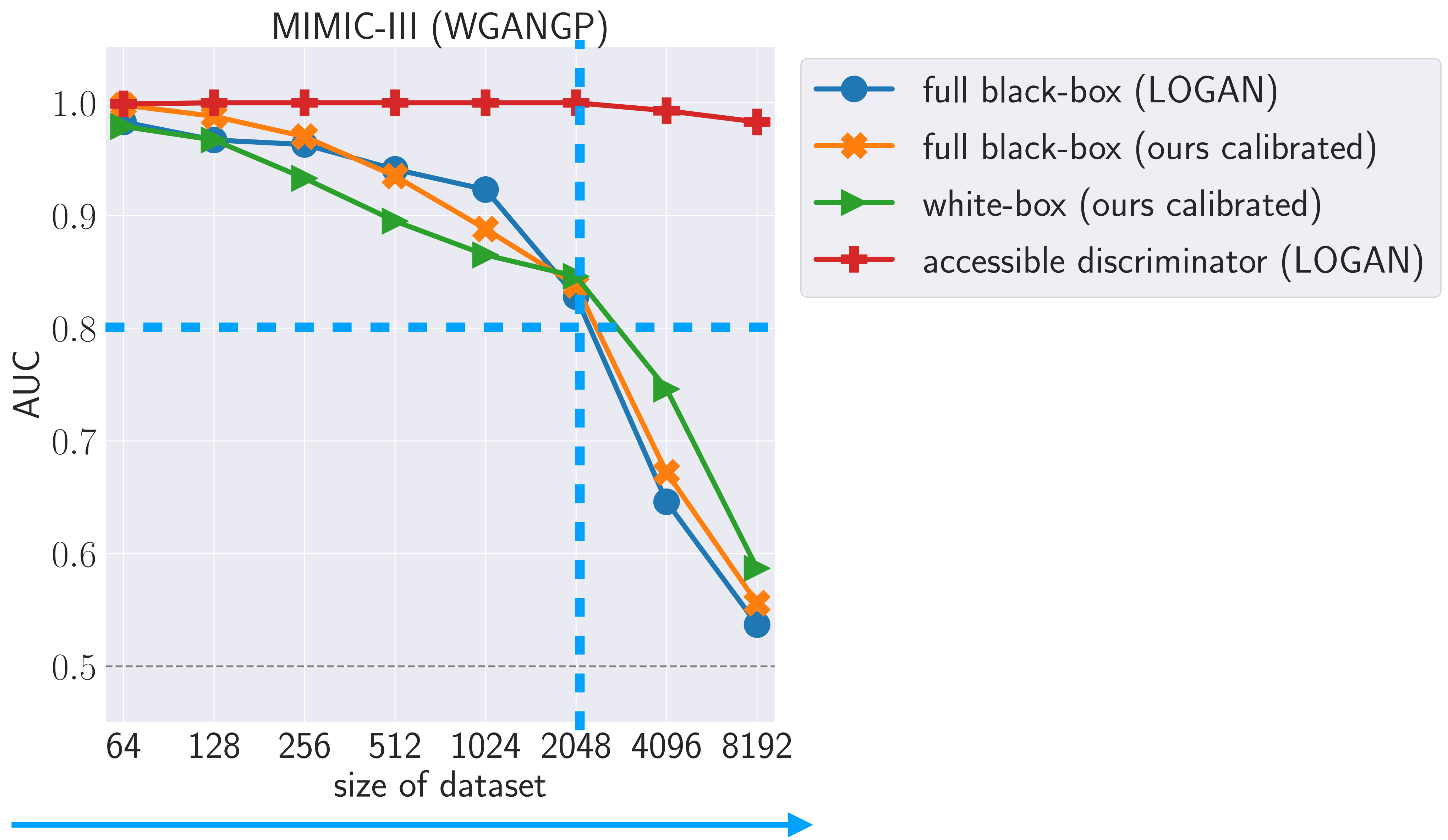
Experiments

- Attack effectiveness —Dataset size



Experiments

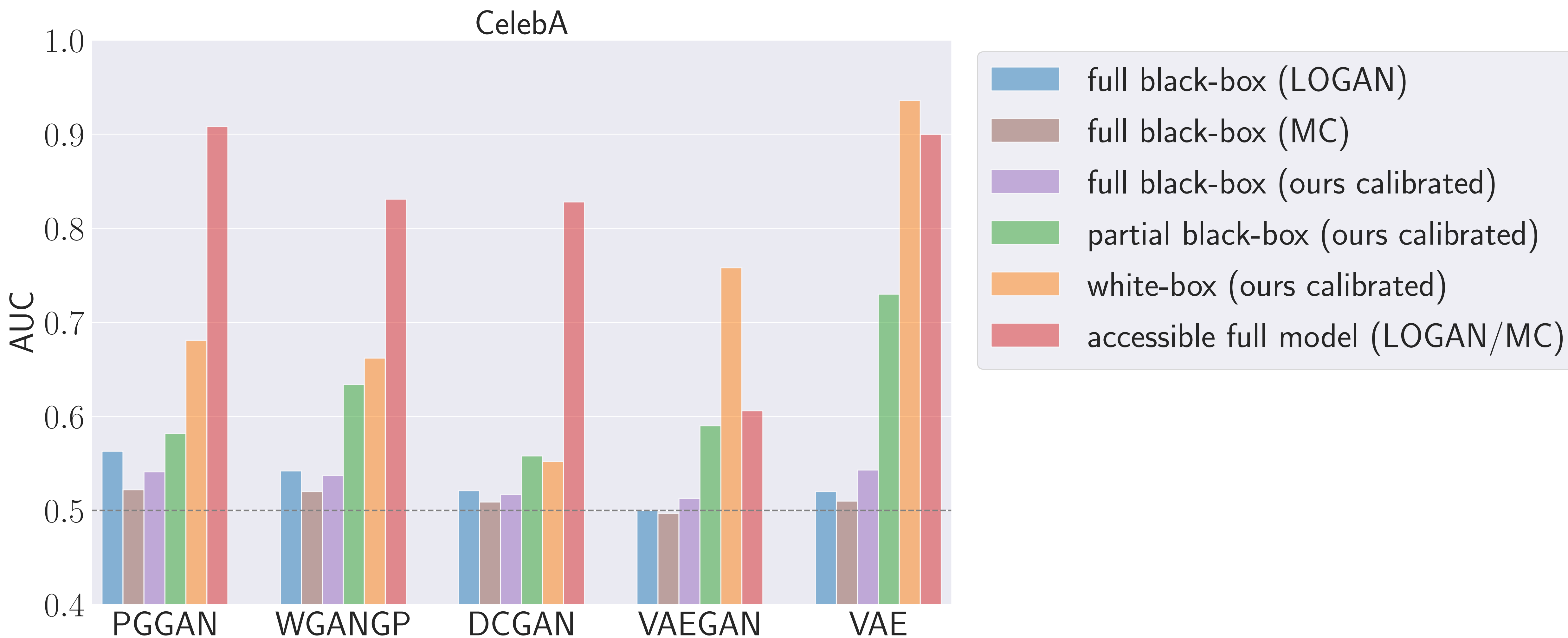
- Attack effectiveness —Dataset size



medical dataset: high privacy risk (AUC >0.8) for ~2k training samples

Experiments

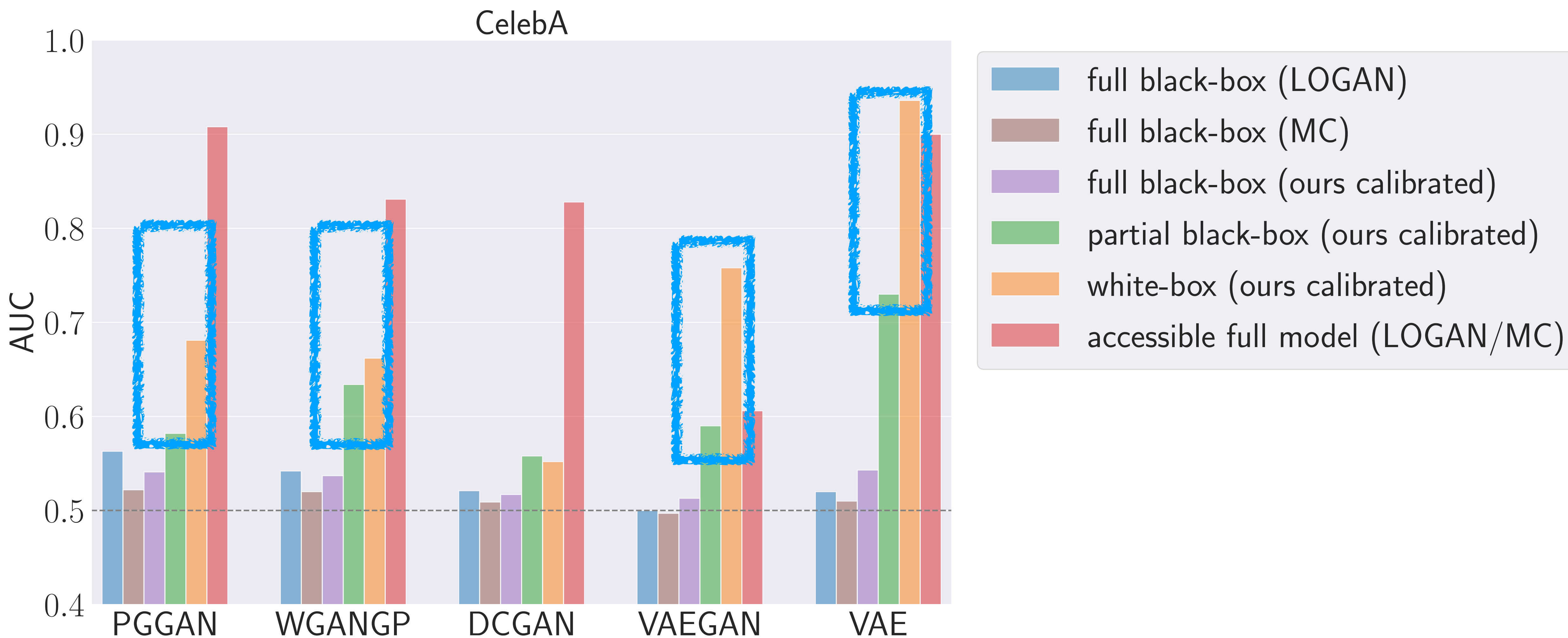
- Attack effectiveness — Attack settings, model architecture



face dataset: 20k training samples

Experiments

- Attack effectiveness — Attack settings, model architecture



face dataset: 20k training samples

attacks are effective in practical settings

More Details in the paper

GAN-Leaks: A Taxonomy of Membership Inference Attack against Generative Models

Dingfan Chen[1] Ning Yu[2,3] Yang Zhang[1] Mario Fritz[1]

Code and Models are available on Github

https://github.com/DingfanChen/GAN-Leaks

[1]CISPA Helmholtz Center for Information Security, Germany

[2]Max Planck Institute for Informatics, Germany

[3]University of Maryland, College Park