Black-box Membership Inference Attacks against Fine-tuned Diffusion Models

Yan Pang, Tianhao Wang

University of Virginia

Diffusion Models are Everywhere

Diffusion Models can generate **photographic-quality** images.

stabilityai.com

Diffusion Models with its generation ability can be trained as a **game engine**....

Real-time recordings of people playing the game DOOM simulated entirely by the *GameNGen* neural model.

Gamengen.github.io/

...achieving better performance, the current SOTA diffusion model trains on **1 billion images** and fine-tunes with **30 million**.

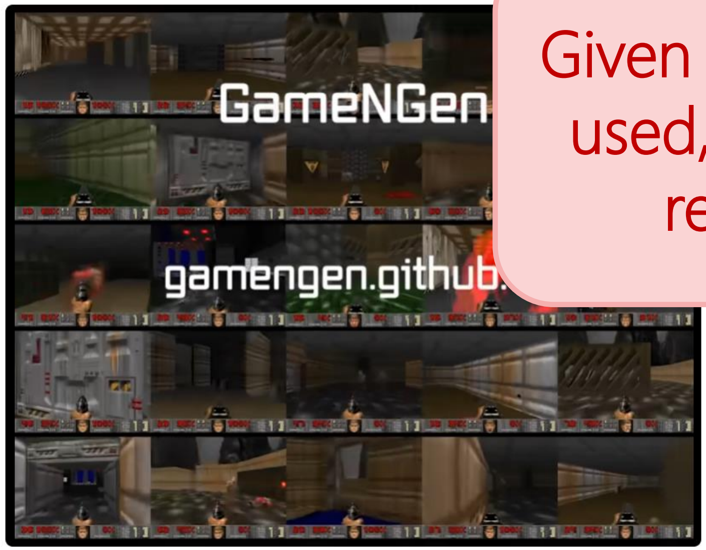
stable-diffusion-3

Diffusion Models are Everywhere

Diffusion Models can generate **photographic-quality** images.

stabilityai.com

Diffusion Models with its generation ability can be trained as a **game engine**....



Real-time recordings of people playing the game DOOM simulated entirely by the *GameNGen* neural model.

Gamengen.github.io/

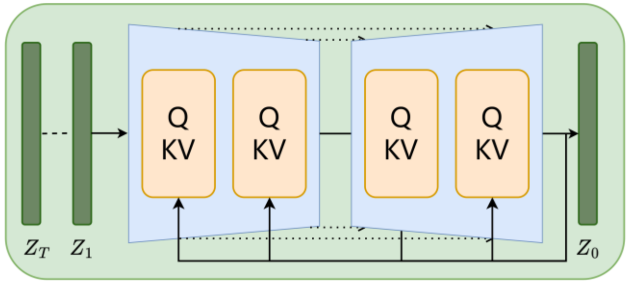
...achieving better performance, the current SOTA diffusion model trains on **1 billion images** and fine-tunes with **30 million**.

Given the extensive training data used, might there be concerns regarding data privacy?

stable-diffusion-3

MIA against Image Generator

Target Model Training Set D



Target Model

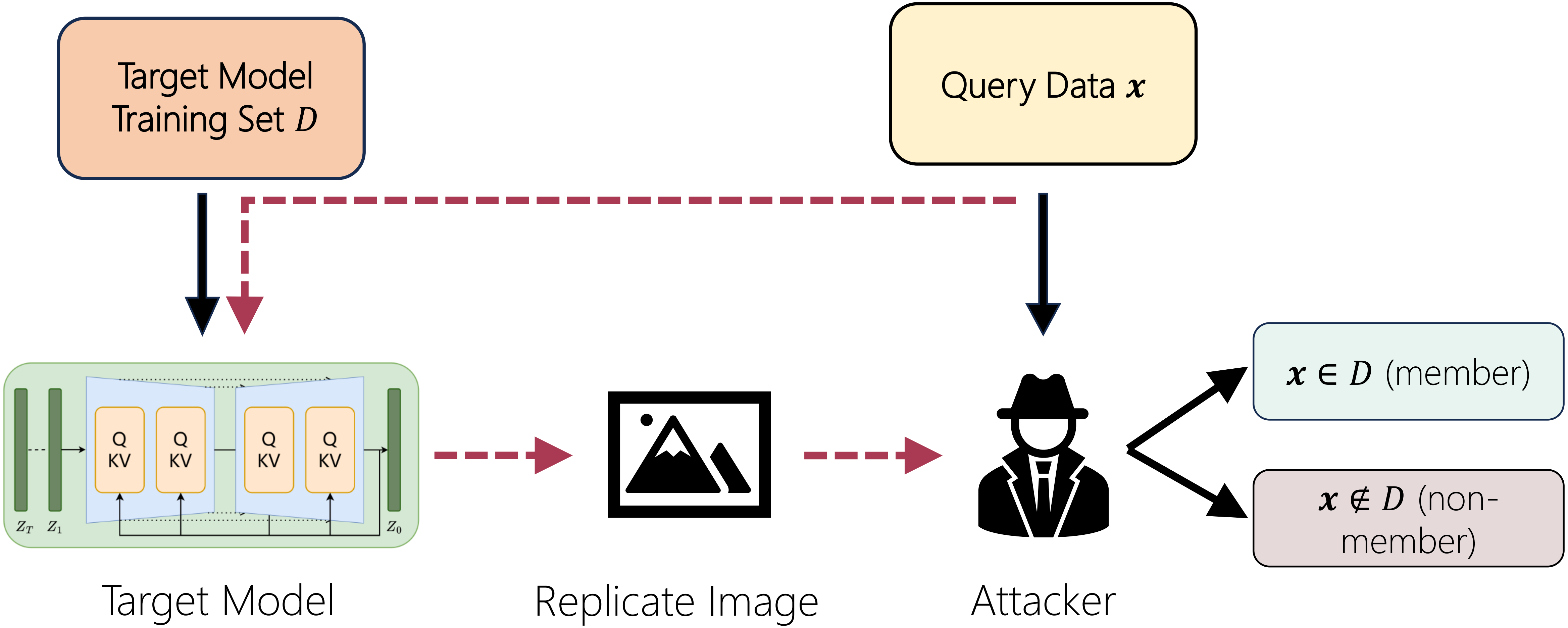
Query Data $\boldsymbol{x}$

Attacker

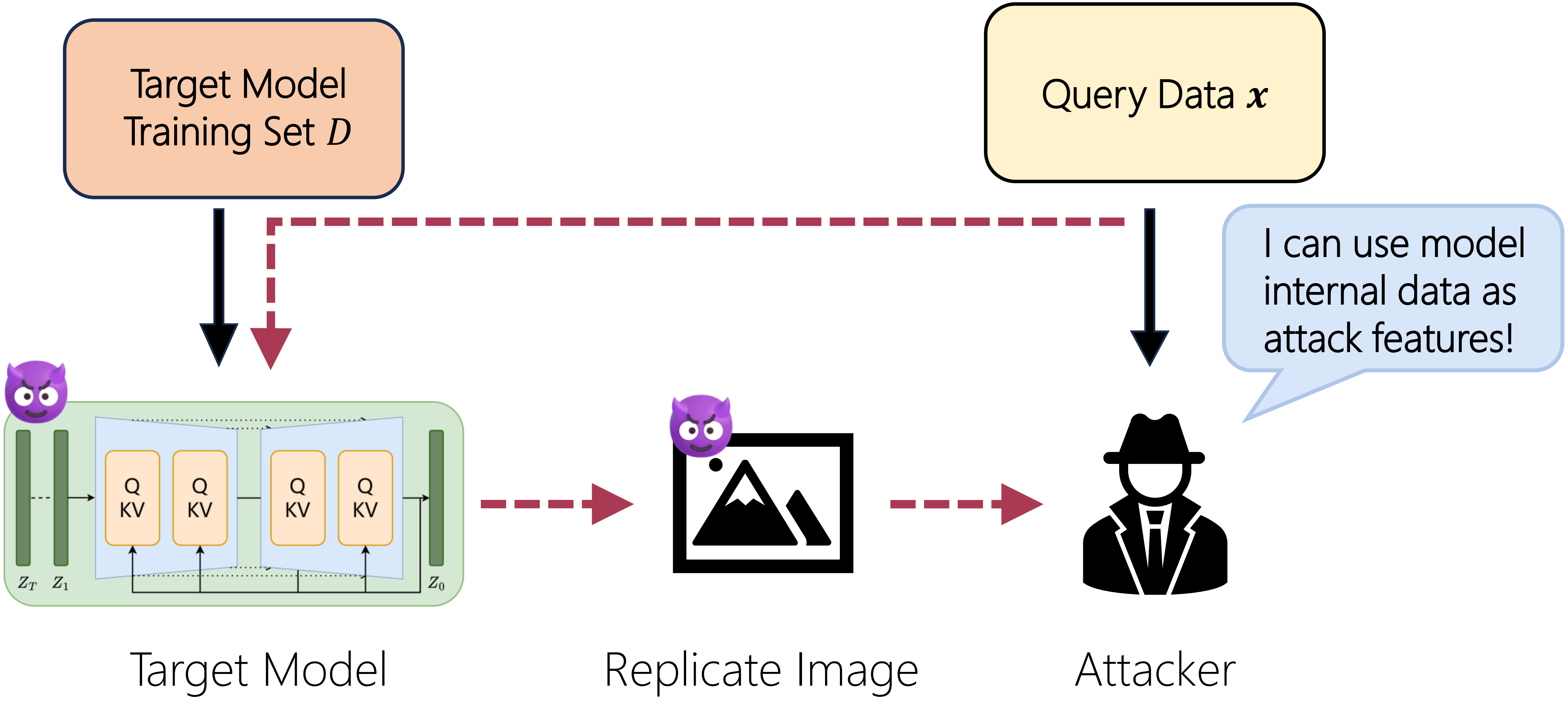
$\boldsymbol{x} \in D$ (member)

$\boldsymbol{x} \notin D$ (non-member)

MIA against Image Generator

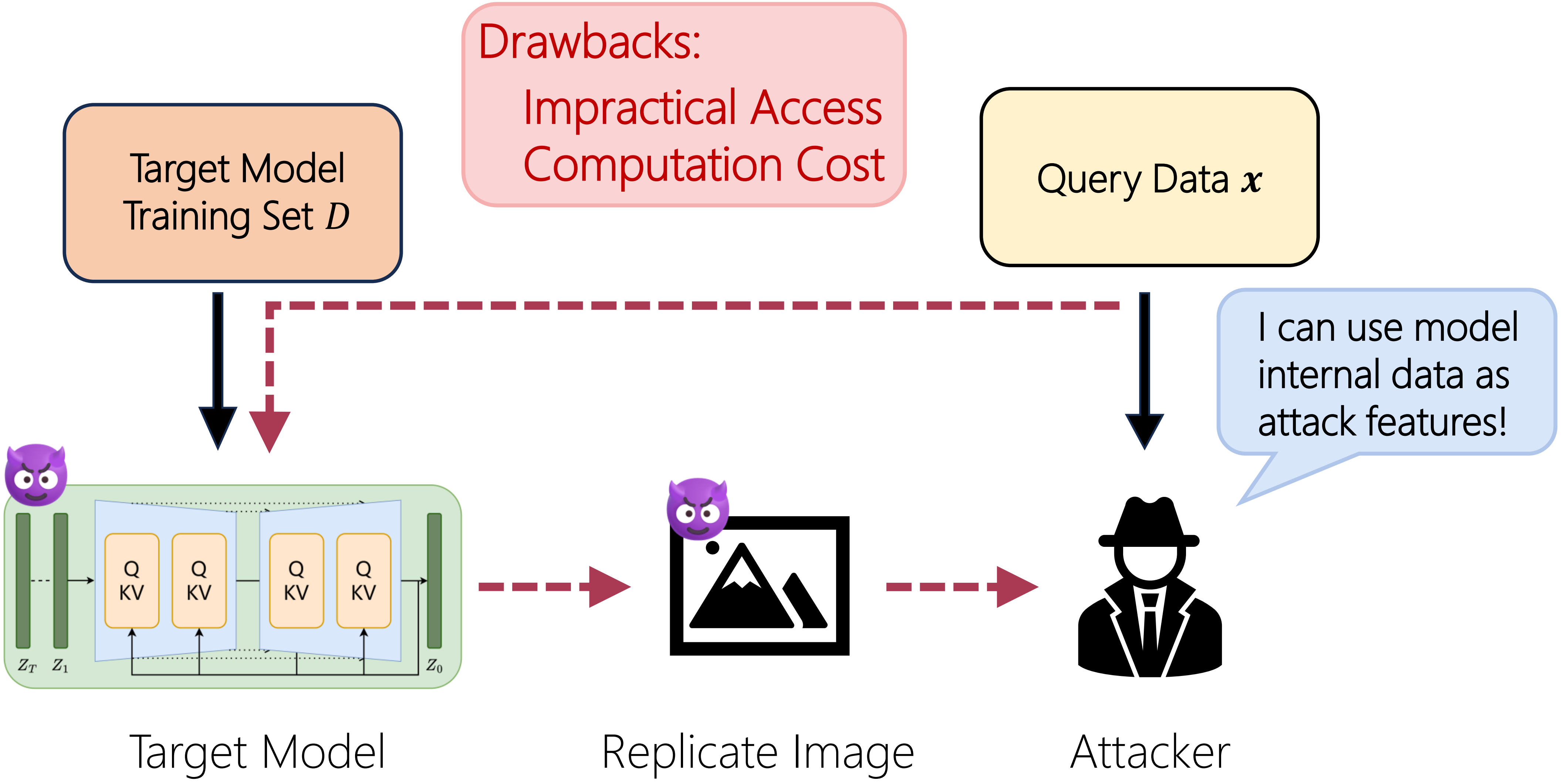


Existing White-box Attacks



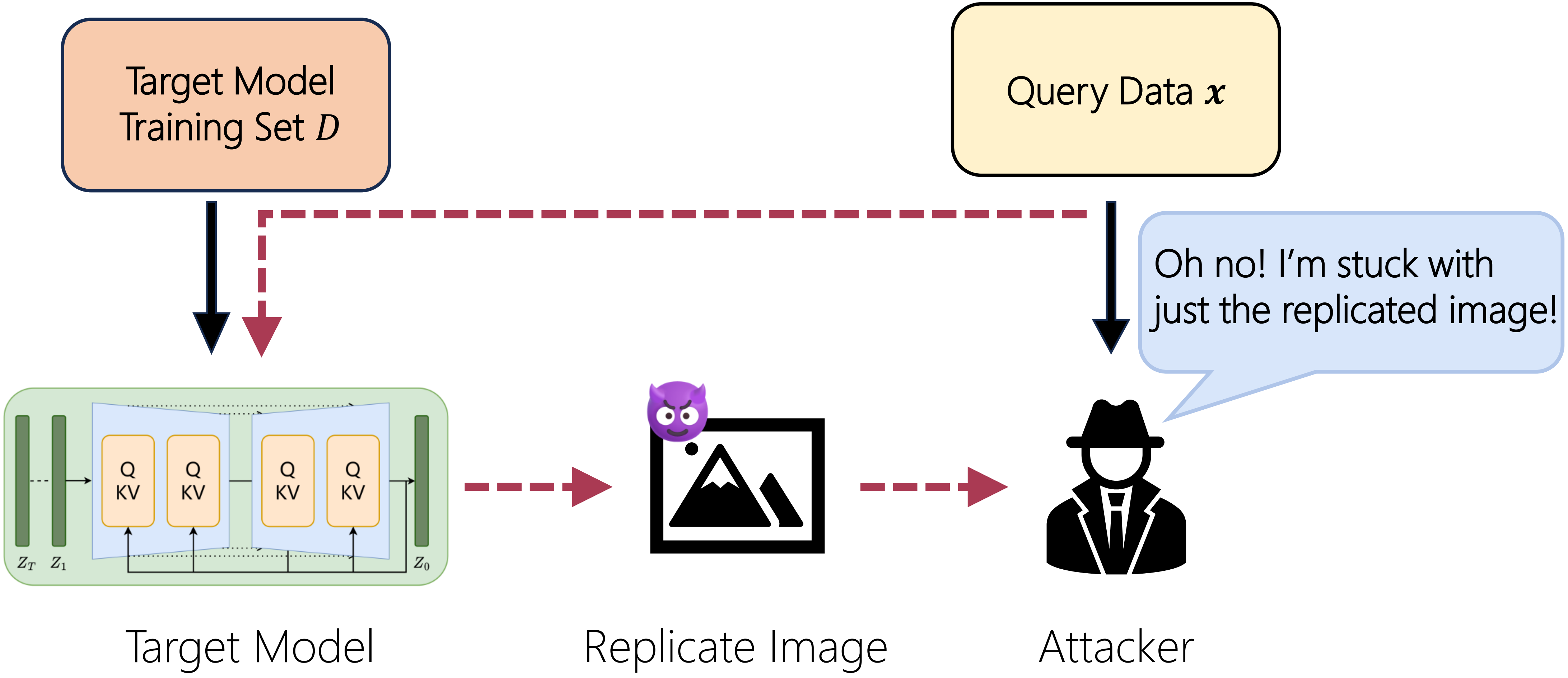
More details in paper.

Existing White-box Attacks



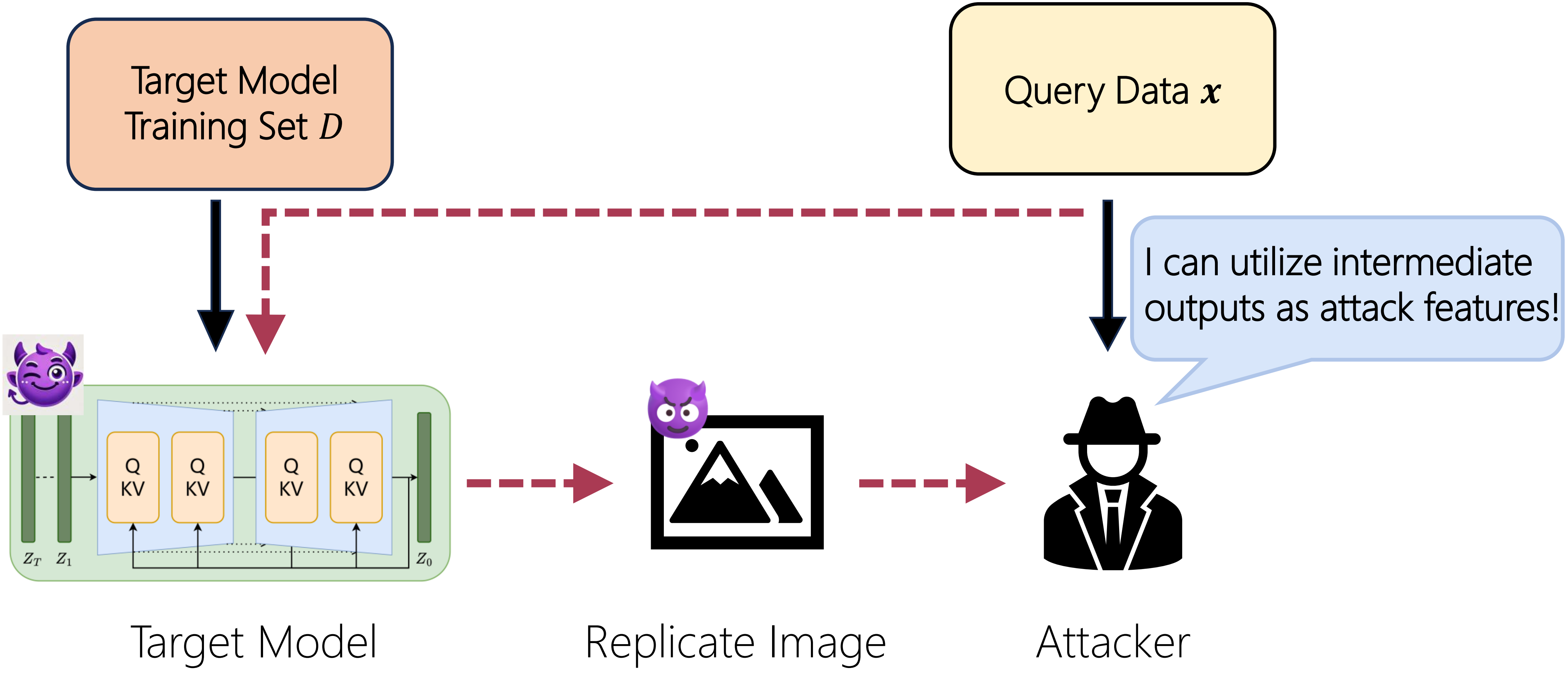
More details in paper.

Existing Black-box Attacks



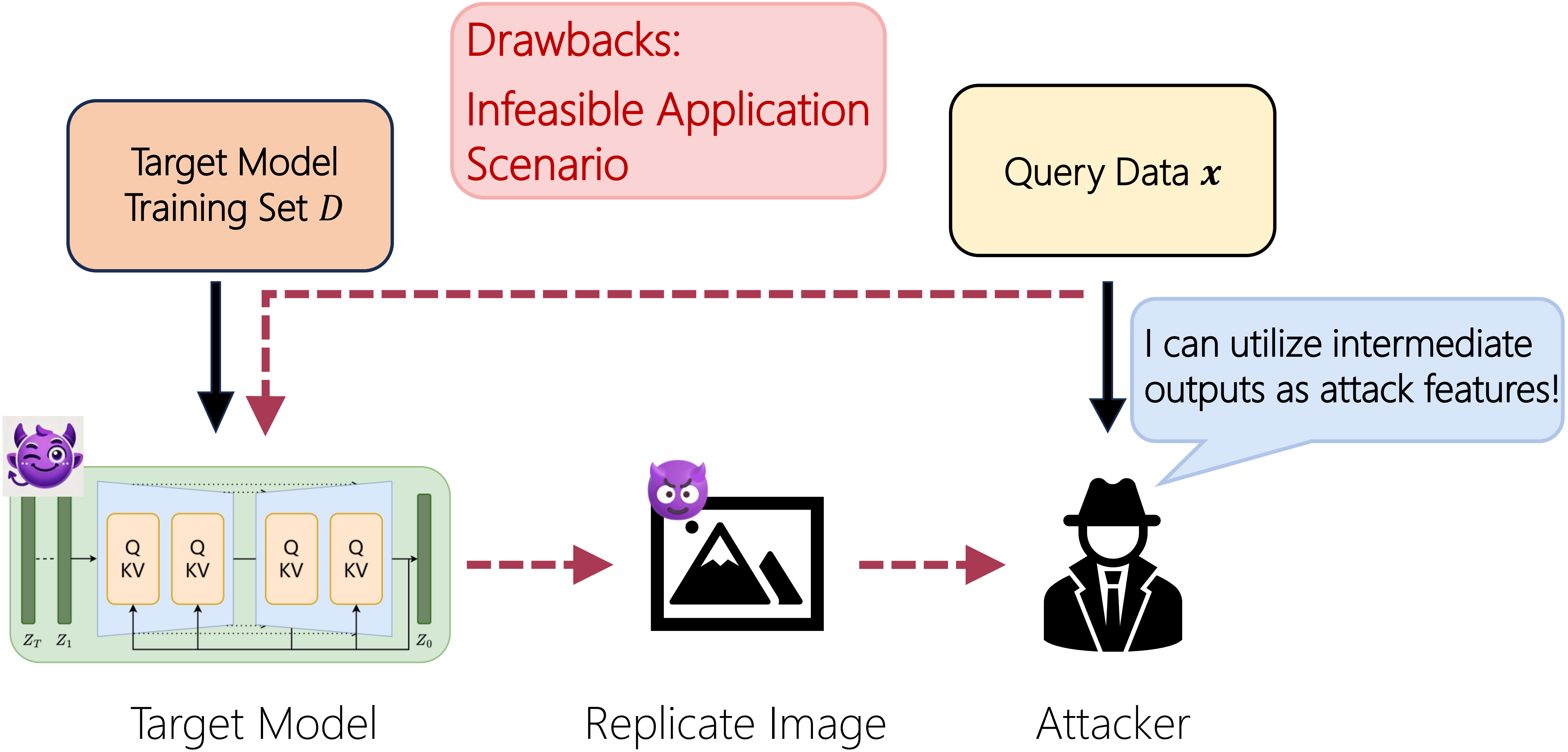
More details in paper.

Existing Gray-box Attacks



More details in paper.

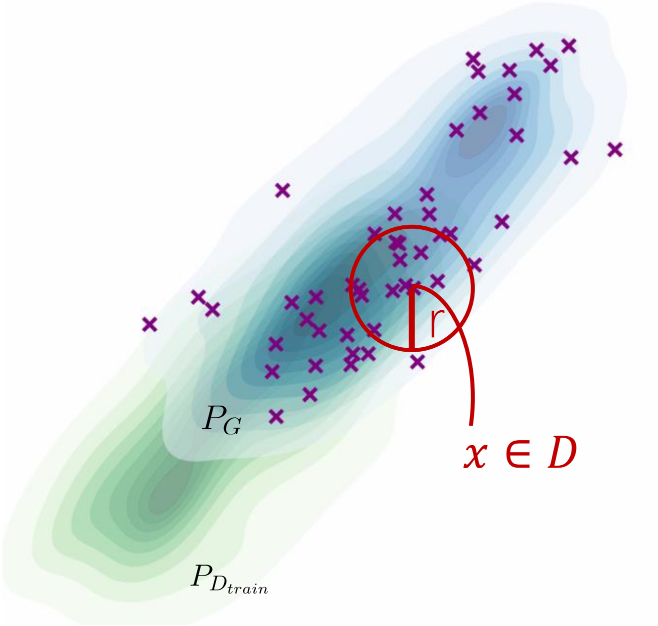
Existing Gray-box Attacks



More details in paper.

Existing Black-box Attacks

Monte-Carlo Attack:



Membership Identification Function:

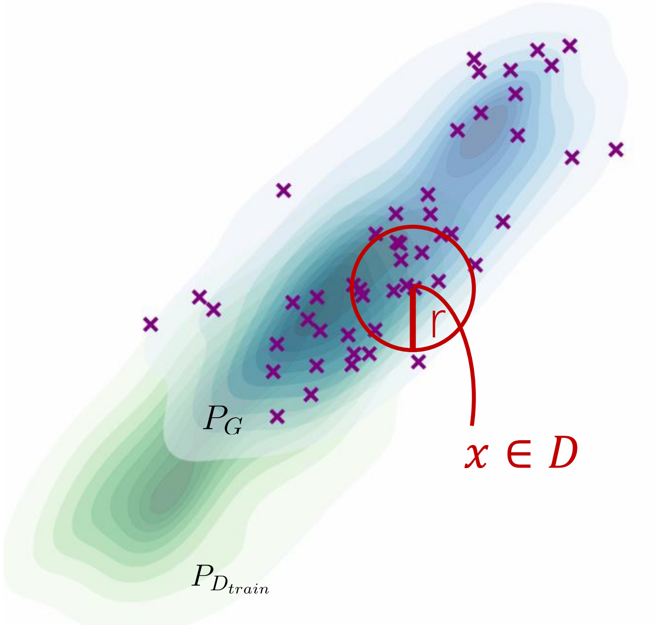
$$\hat{f}_{MC-\epsilon}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{x'_i \in U_r(x)}$$

$$U_r(x) = \{x' \mid d(x, x') \leq r\}$$

[Hilprecht et al. PETS 19].

Existing Black-box Attacks

Monte-Carlo Attack:

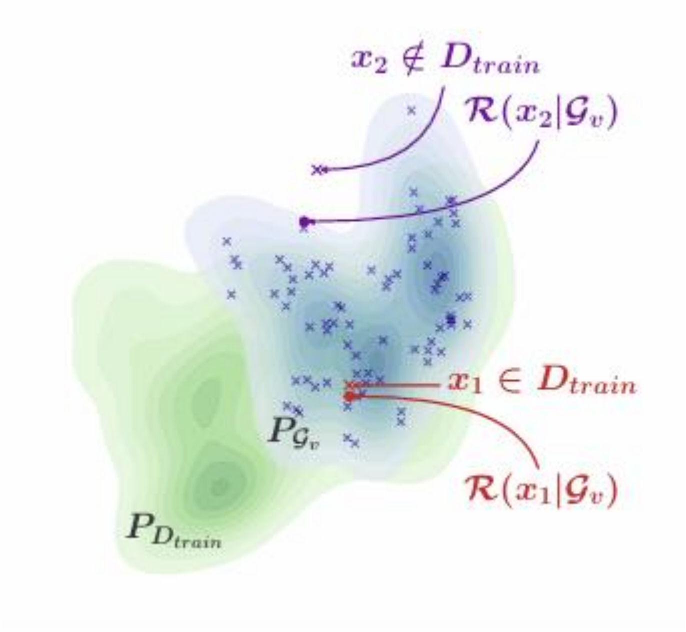


Membership Identification Function:

$$\hat{f}_{MC-\epsilon}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{x_i' \in U_r(x)}$$
$$U_r(x) = \{x' \mid d(x, x') \leq r\}$$

[Hilprecht et al. PETS 19].

GAN-Leaks:



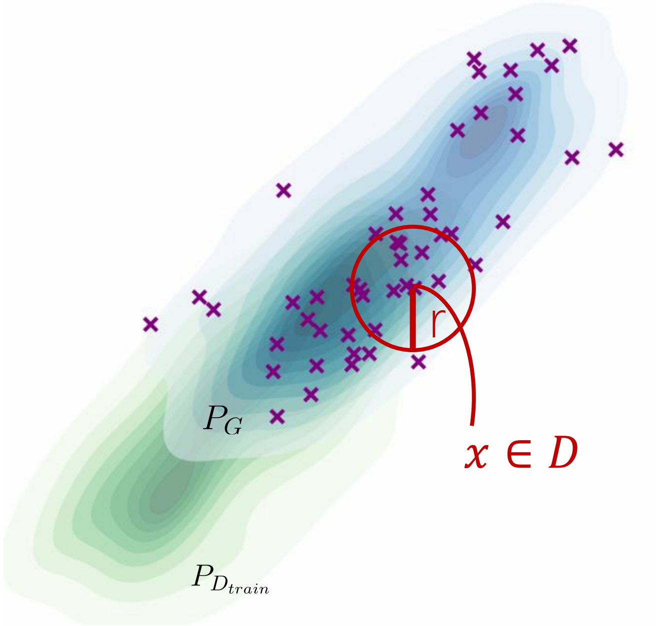
Membership Identification Function:

$$P(b_i = 1|x_i, \theta_v) \propto \frac{1}{k}\sum_{i=1}^{k} \exp(-L(x, \mathcal{G}(z_i)))$$
$$z_i \sim P_z$$

[Chen et al. CCS 2020].

Existing Black-box Attacks

Monte-Carlo Attack:



Membership Identification Function:

$$\hat{f}_{MC-\epsilon}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{x'_i \in U_r(x)}$$

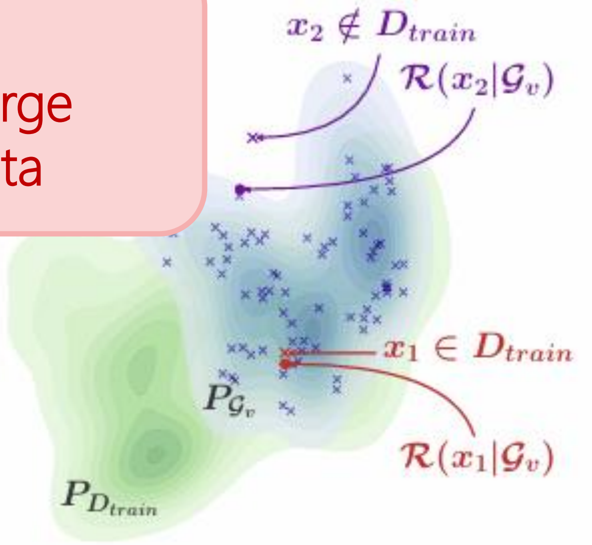
$$U_r(x) = \{x' \mid d(x, x') \leq r\}$$

[Hilprecht et al. PETS 19].

Drawbacks:

Sampling a large amount of data

GAN-Leaks:



Membership Identification Function:

$$P(b_i = 1|x_i, \theta_v) \propto \frac{1}{k}\sum_{i=1}^{k} \exp(-L(x, \mathcal{G}(z_i)))$$

$$z_i \sim P_z$$

[Chen et al. CCS 2020].

Can We Improve the Current Attack?

	Target Model Training Set	Shadow Models
GAN Leaks [1]	Need Access	Not utilized
Int. attack [2]	Need Access	Not utilized
Pixel attack [3]	Need Access	Not utilized
Dist. Attack [4]	Partial Access	Not utilized

Need Access: Use target model's training set as the member set.

Due to time consumption and computing resource issues, existing solutions make "interesting" assumptions.

[1] Matsumoto et al., IEEE DLSP' 23, [2] Wu et al., Arxiv 2022,
[3] Dubinski et al., WACV 2024, [4] Zhang et al., WACV 2024

Can We Improve the Current Attack?

	Target Model Training Set	Shadow Models
GAN Leaks [1]	Need Access	Not utilized
Int. attack [2]	Need Access	Not utilized
Pixel attack [3]	Need Access	Not utilized
Dist. Attack [4]	Partial Access	Not utilized

Need Access: Use target model's training set as the member set.

Existing solutions target the unconditional generator, but the current SOTA model is Stable Diffusion. The model's properties have changed, so the attack needs to be redesigned.

Due to time consumption and computing resource issues, existing solutions make “interesting” assumptions.

[1] Matsumoto et al., IEEE DLSP’ 23, [2] Wu et al., Arxiv 2022,
[3] Dubinski et al., WACV 2024, [4] Zhang et al., WACV 2024

Can We Improve the Current Attack?

	Target Model Training Set	Shadow Models
GAN Leaks [1]	Need Access	Not utilized
Int. attack [2]	Need Access	Not utilized
Pixel attack [3]	Need Access	Not utilized
Dist. Attack [4]	Partial Access	Not utilized

Need Access: Use target model's training set as the member set.

Existing solutions target t
generator, but the curren
Diffusion. The model's pro
changed, so the attack needs to be redesigned.

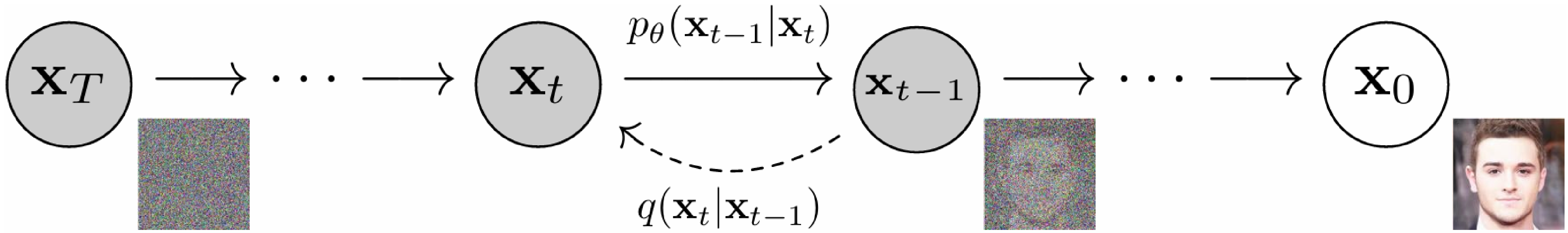
We aim to overcome these key challenges by focusing on fine-tuned conditional diffusion models.

ue
sc
inte

on and computing
g solutions make
ns.

[1] Matsumoto et al., IEEE DLSP’ 23, [2] Wu et al., Arxiv 2022,
[3] Dubinski et al., WACV 2024, [4] Zhang et al., WACV 2024

Theoretical Foundation



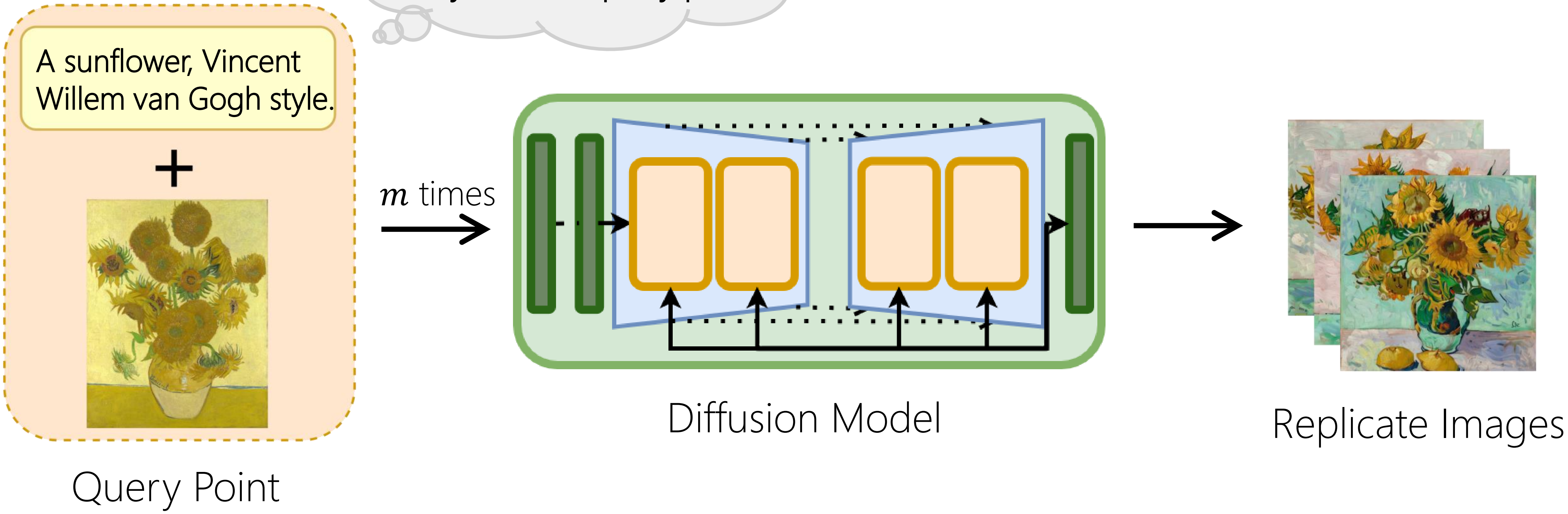
Our Theorem: Assuming we have a pre-trained diffusion model $\hat{x}_\theta$ with its training set D_m, and use a bit b to represent the membership of query sample x {($\mathbf{1}$ for member and $\mathbf{0}$ for non-member)}.

$$\Pr\{b = 1|x, \theta\} \propto -\|x_0 - \hat{x}_\theta(x_t, t)\|_2^2$$

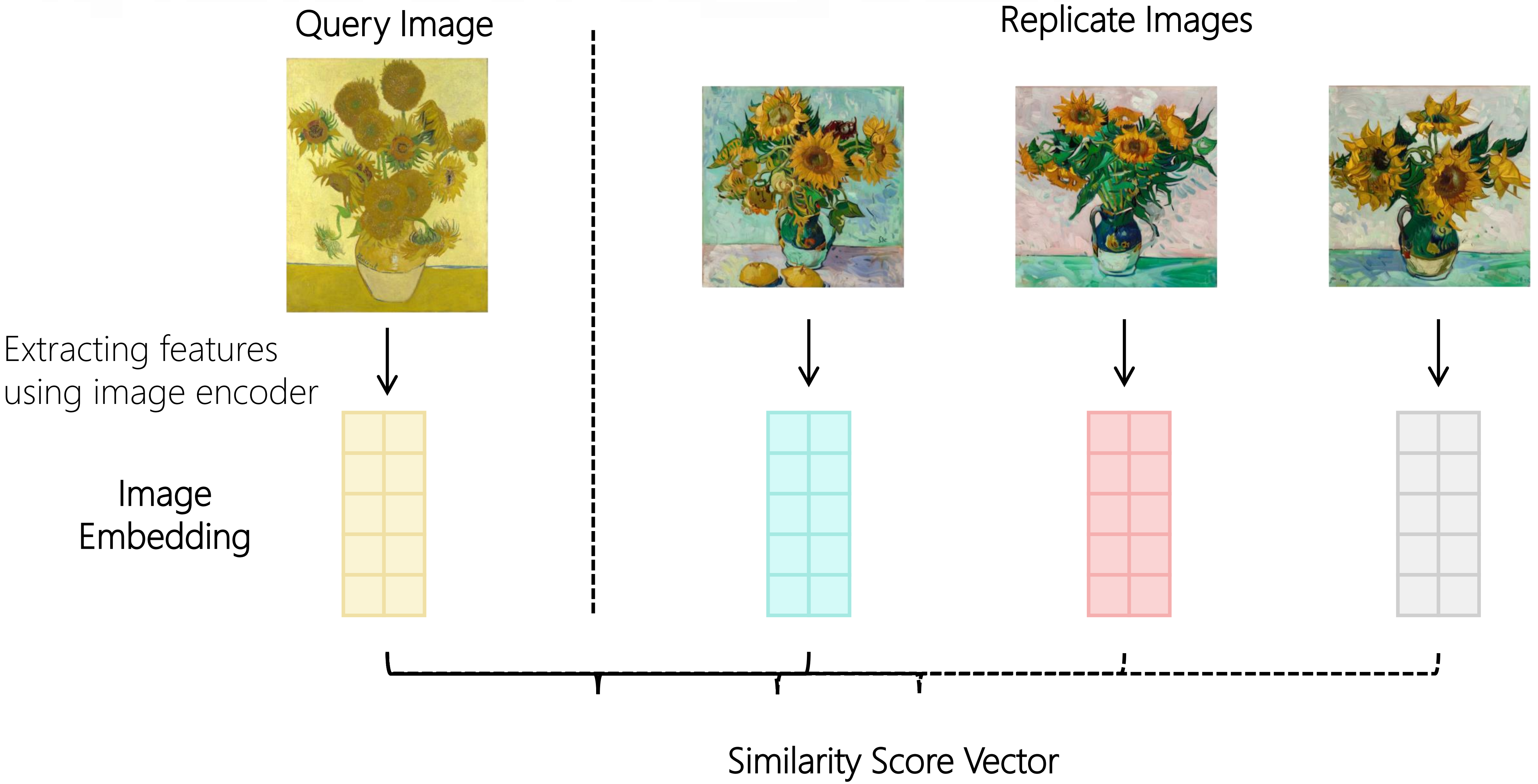
[Photo: DDPM, Ho et al. NIPS 2020]

Step 1: Synthesize Replicate Samples

No text? No problem! BLIP comes in to fill the gap for any textless query points.



Step 2: Calculate Similarity Scores



Step 3: Execute the Attack

Similarity Score Vectors

Apply Statistical Function

Infer Membership

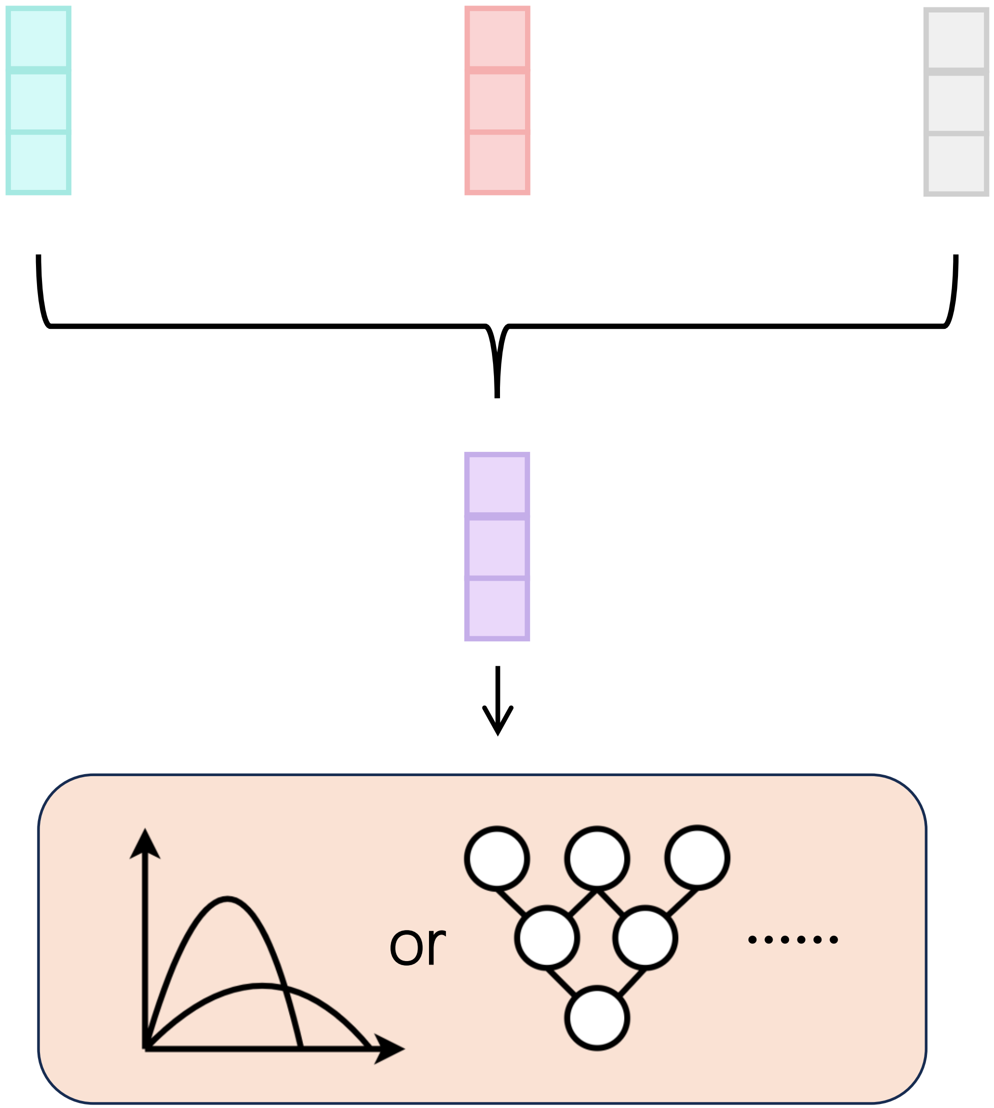
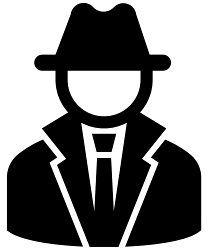


Image embedding color indicates which embedding generated the similarity score with query image.

Three attack models: threshold-based, distribution-based, classifier-based... all trained with shadow models.



Attacker

Evaluation Setup

Model: Stable Diffusion v1-5.

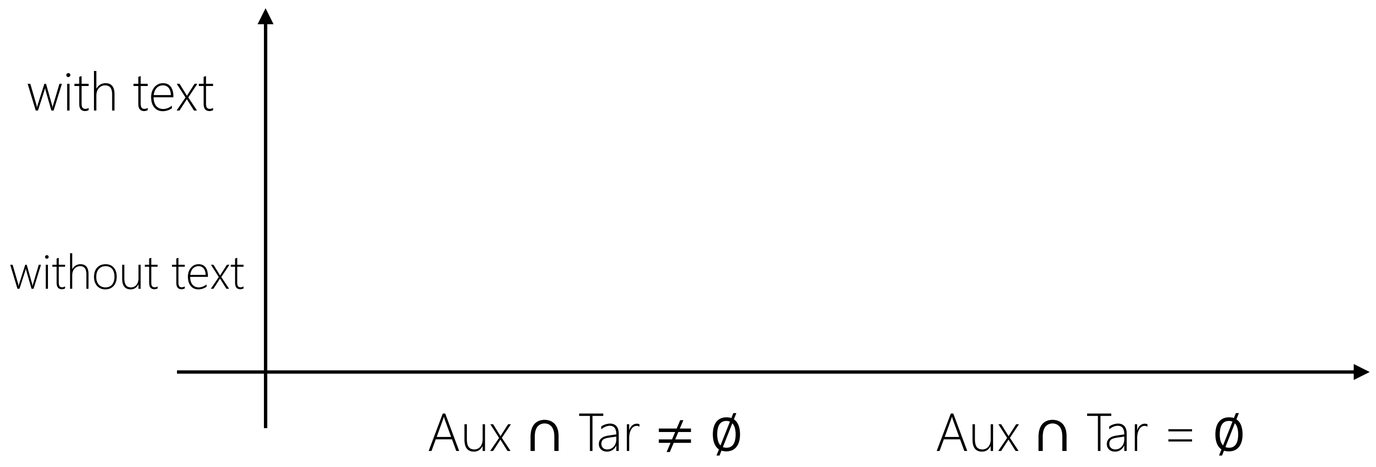
Dataset: MS COCO, CelebA-Dialog, WIT.

Metric: Accuracy, AUC ROC, True Positive Rate at low False Positive Rate.

Baseline: Matsumoto et al.[1], Zhang et al.[2]

[1] Matsumoto et al., IEEE DLSP' 23, [2] Zhang et al., IEEE WACV' 23

Instantiation Attack & Comparison



[1] Matsumoto et al., IEEE DLSP’ 23, [2] Zhang et al., IEEE WACV’ 23

Instantiation Attack & Comparison

Based on **Query Sample Component** and **Auxiliary Dataset** :

	Aux ∩ Tar ≠ ∅	Aux ∩ Tar = ∅
with text	Attack-I	Attack-IV
without text	Attack-II	Attack-III

[1] Matsumoto et al., IEEE DLSP' 23, [2] Zhang et al., IEEE WACV' 23

Instantiation Attack & Comparison

Based on **Query Sample Component** and **Auxiliary Dataset** :

	Aux ∩ Tar ≠ ∅	Aux ∩ Tar = ∅
with text	Attack-I	Attack-IV
without text	Attack-II	Attack-III

Attack type	CelebA-Dialog		
	ASR	AUC	TPR@FPR=1%
Matsumoto et al.[1]	0.52	0.50	0.01
Zhang et al.[2]	0.51	0.49	0.01
Attack-I	0.85	0.93	0.53
$\boxed{\text{Attack-II}}$	**0.88**	**0.93**	**0.60**
Attack-III	0.87	0.94	0.54
Attack-IV	0.87	0.93	0.57

Experimental settings fixed; three replicate samples per query. Our attack outperforms baseline methods.

[1] Matsumoto et al., IEEE DLSP' 23, [2] Zhang et al., IEEE WACV' 23

Instantiation Attack & Comparison

Based on **Query Sample Component** and **Auxiliary Dataset** :

	Aux ∩ Tar ≠ ∅	Aux ∩ Tar = ∅
with text	Attack-I	Attack-IV
without text	Attack-II	Attack-III

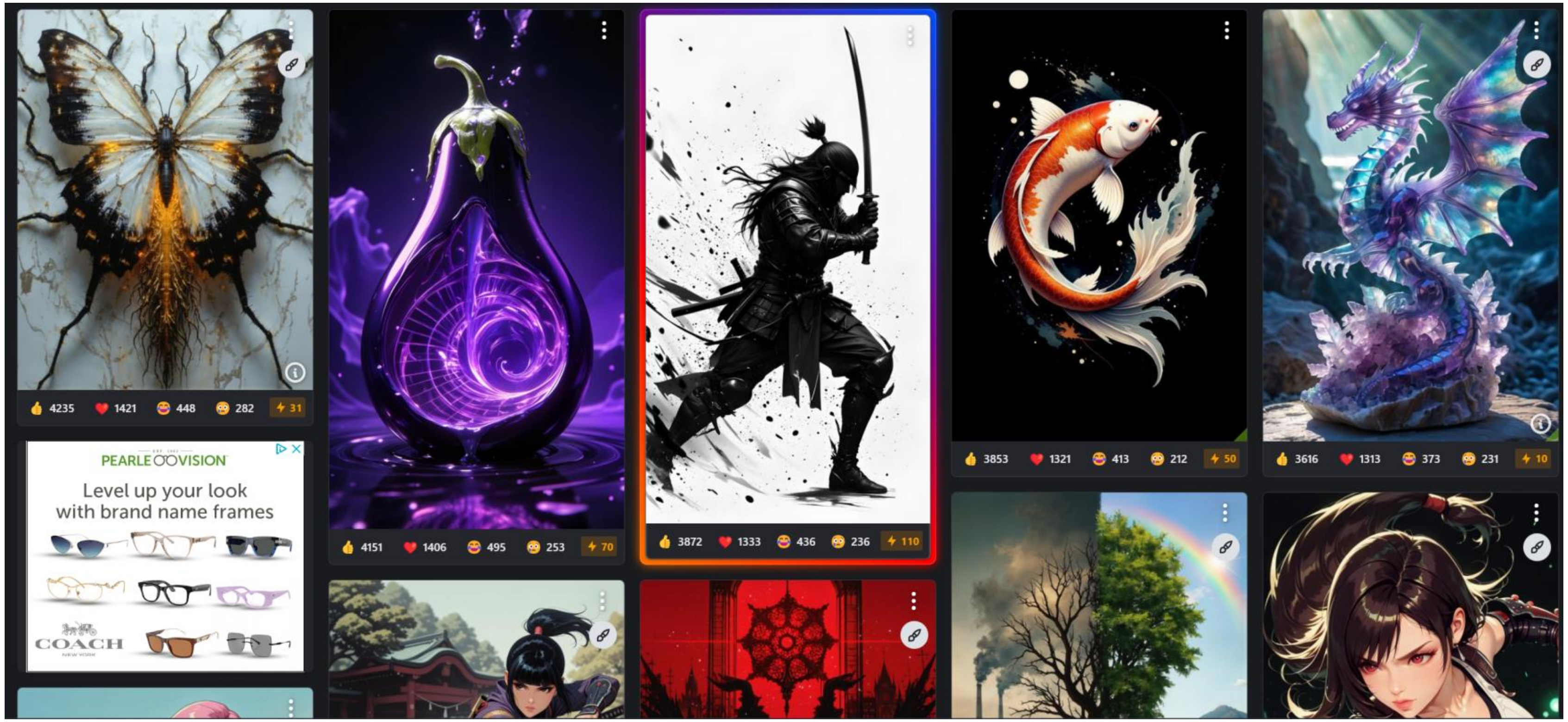
Attack type	CelebA-Dialog		
	ASR	AUC	TPR@FPR=1%
Matsumoto et al.[1]	0.52	0.50	0.01
Zhang et al.[2]	0.51	0.49	0.01
Attack-I	0.85	0.93	0.53
Attack-II	**0.88**	**0.93**	**0.60**
Attack-III	0.87	0.94	0.54
Attack-IV	0.87	0.93	0.57

Experimental settings fixed; three replicate samples per query. Our attack outperforms baseline methods.

Also did ablations, details in paper

[1] Matsumoto et al., IEEE DLSP' 23, [2] Zhang et al., IEEE WACV' 23

Case Study



[Civitai: The Home of Open-Source Generative AI](#)

Case Study

One of the Original Painting

Download Checkpoint

Member Sample

Original Model

Non-Member Sample

Case Study

Use our attack as an auditing tool target models on Civitai:

Art Style (Artist)	Member	Non-member	Diff.	ROC-AUC
Vincent van Gogh	**0.92**	**0.44**	**0.48**	**0.91**
Baishi Qi	0.77	0.40	0.37	0.88
Ukiyo-e	0.70	0.45	0.25	0.87
Uemura Shoen	0.88	0.41	0.47	0.89
Wanostyle	0.80	0.49	0.31	0.87
Ken Kelly	0.80	0.47	0.33	0.89
Shanshui Painting	0.88	0.47	0.41	0.86

We observe a clear distinction in similarity scores between members and non-members.

[Civitai: The Home of Open-Source Generative AI]

Case Study

Use our attack as an auditing tool target models on Civitai:

Art Style (Artist)	Member	Non-member	Diff.	ROC-AUC
Vincent van Gogh	**0.92**	**0.44**	**0.48**	**0.91**
Baishi Qi	0.77	0.40	0.37	0.88
Ukiyo-e	0.70	0.45	0.25	0.87
Uemura Shoen	0.88	0.41	0.47	0.89
Wanostyle	0.80	0.49	0.31	0.87
Ken Kelly	0.80	0.47	0.33	0.89
Shanshui Painting	0.88	0.47	0.41	0.86

Our method can serve as an effective auditing tool to evaluate the model on Civitai.

[Civitai: The Home of Open-Source Generative AI]

Summary

- We find that membership inference attacks against GANs and VAEs cannot be directly used to target current diffusion models due to architectural and generative differences.
- We propose an attack pipeline targeting conditional diffusion models across four scenarios, designing three attack models and evaluating key factors on three datasets.

Paper (with links to Github and dataset):