Property Existence Inference against Generative Models

Lijin Wang[1], Jingjing Wang[1], Jie Wan[1], Lin Long[1], Ziqi Yang[1,2*], Zhan Qin[1,2]

[1] Zhejiang University,

[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center

* Corresponding Author

Image Generative Models: Produce Images

stability.ai
Stable Diffusion

Google DeepMind
Imagen

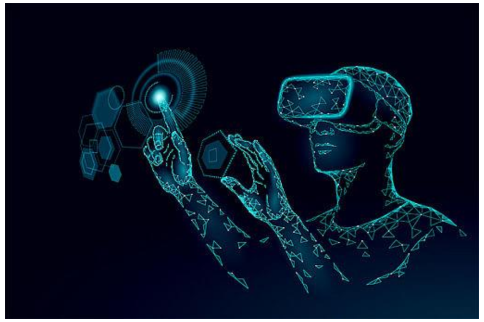
Text Prompt

Random Input

Creative Design

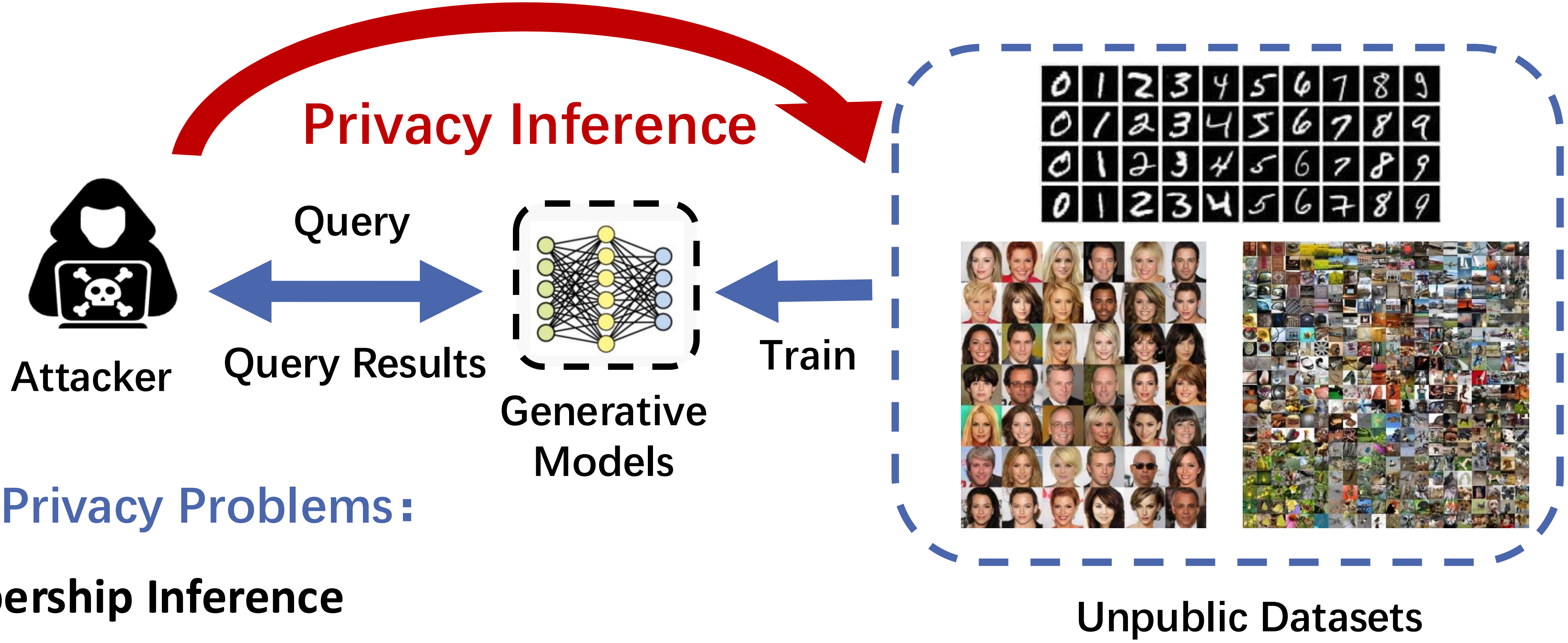
Data Augmentation

Medical Intelligence



Virtual Reality

Image Generative Models: Privacy Problems



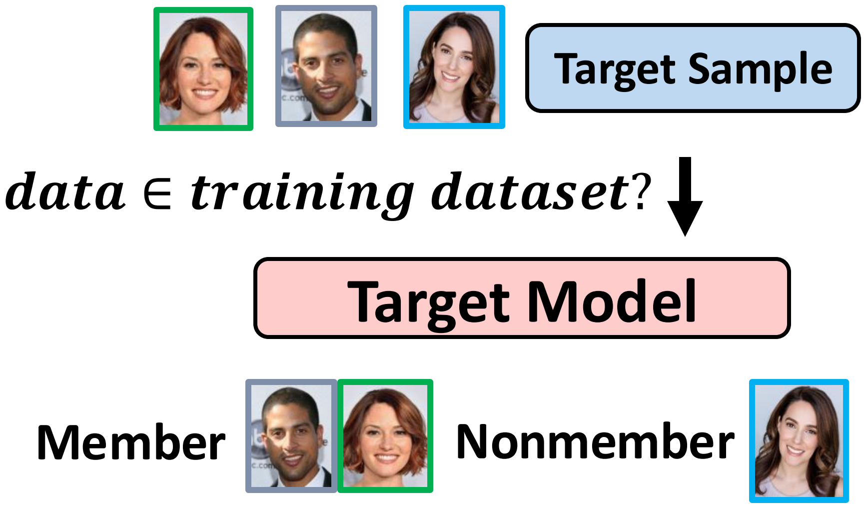
Existing Privacy Problems：

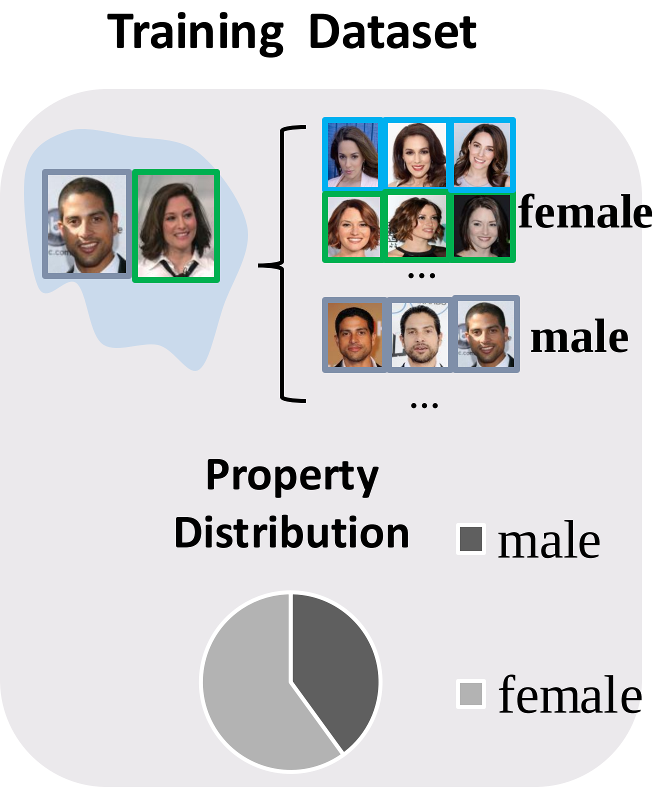
- **Membership Inference**
- **Property Inference**

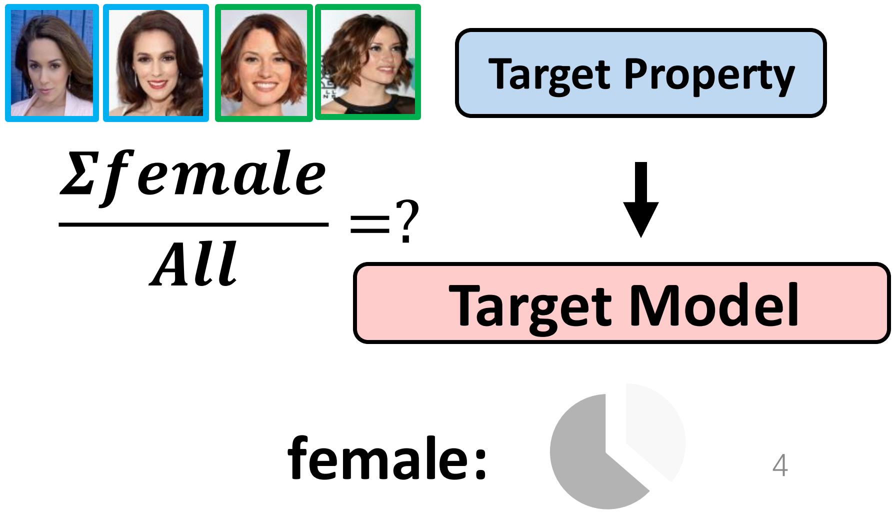
Existing Privacy Problems

Membership Inference
To infer whether specific sample is in the training dataset of the target model

Property Inference
To infer the accounting ratio of specific property







Proposed: Property Existence Inference

Membership Inference

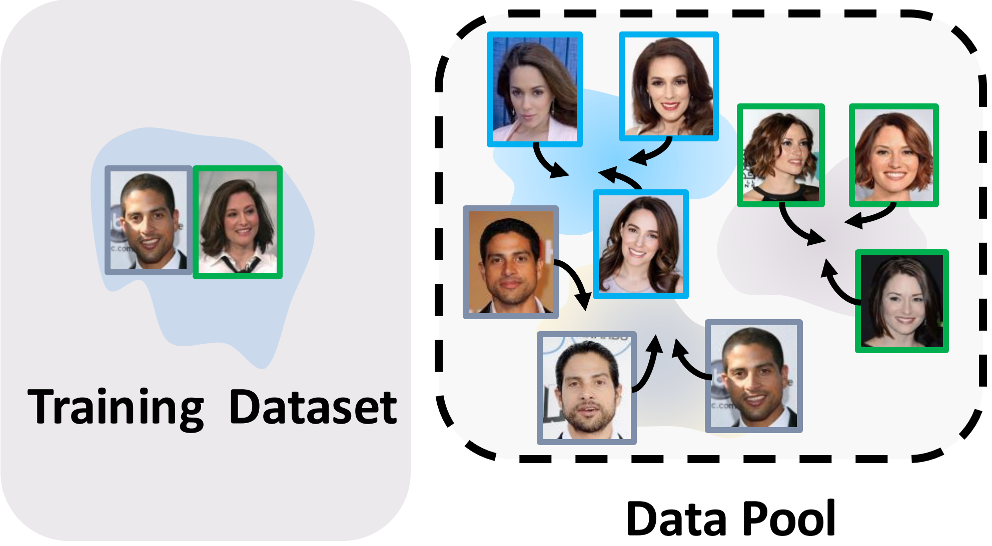
To infer whether specific sample is in the training dataset of the target model

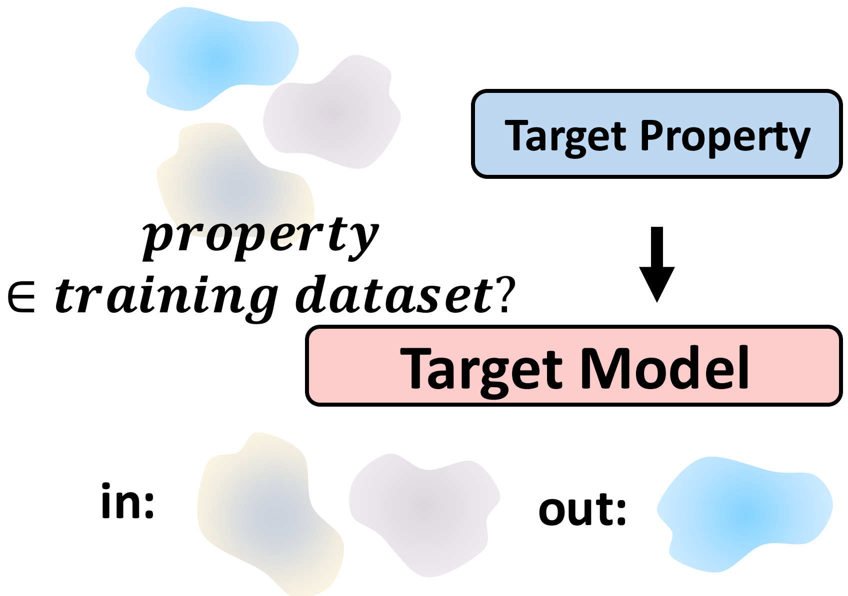
Property Inference

To infer the accounting ratio of specific property

Property Existence Inference

To infer whether specific property is in the training dataset of the target model





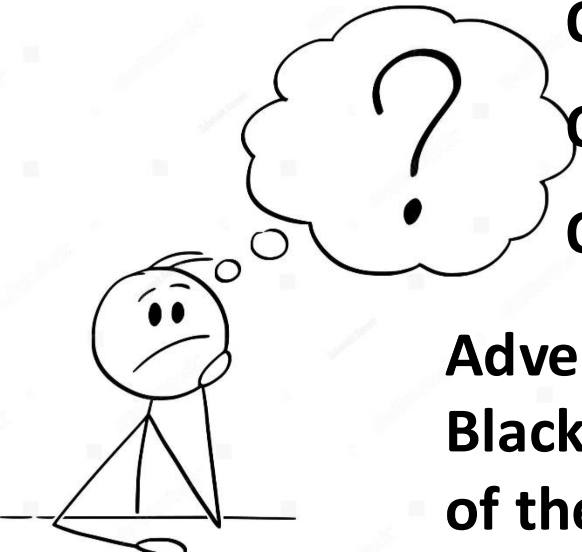
Proposed: Property Existence Inference

Compared with Membership Inference

a more practical setting - target sample ≠ sample in the training dataset

Compared with Property Inference

Interested in property accounting for small proportion <0.1%



Q1 Does StableDiffusion use van gogh's artw Property Existence Inference

Q2 Does StableDiffusion use 《The Mona Lisa's Sn Membership Inference

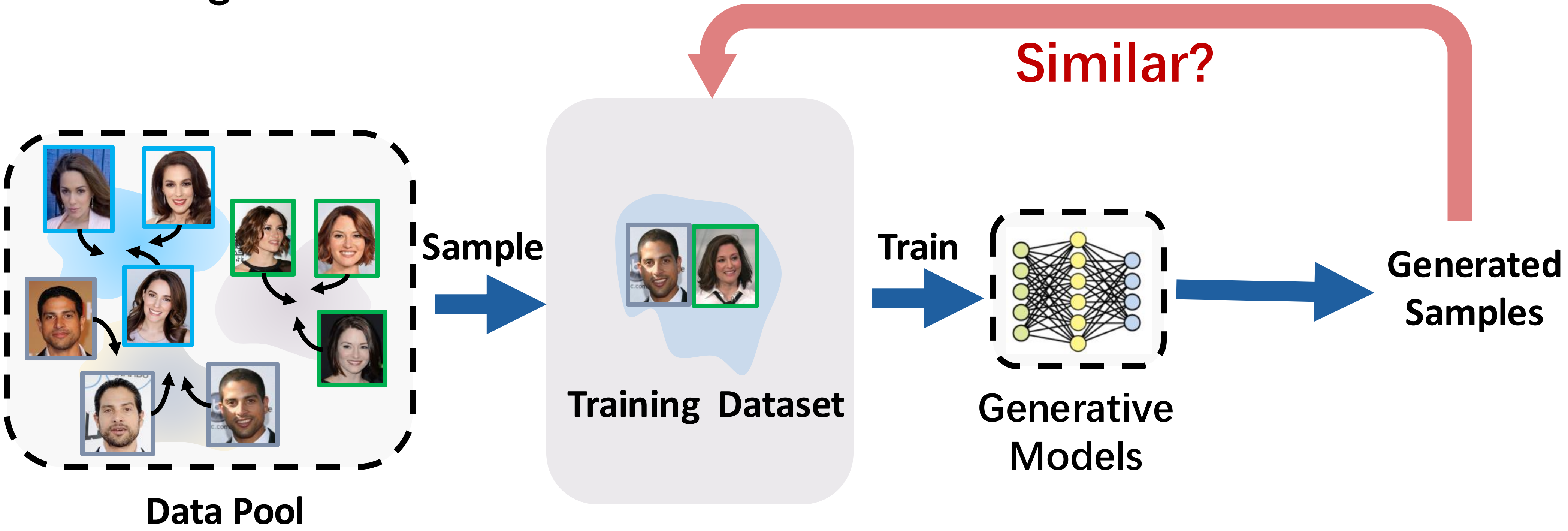
Q3 How much females' photos does StableDiffusion us Property Inference

Adversarial Knowledge:
Black-box Scenario - Attacker can only get the generated results of the target model

Property Existence Inference: Basic Ideas

Motivation

Generated Images may carry property that generative models have seen in the training dataset.



Property Existence Inference: Overview

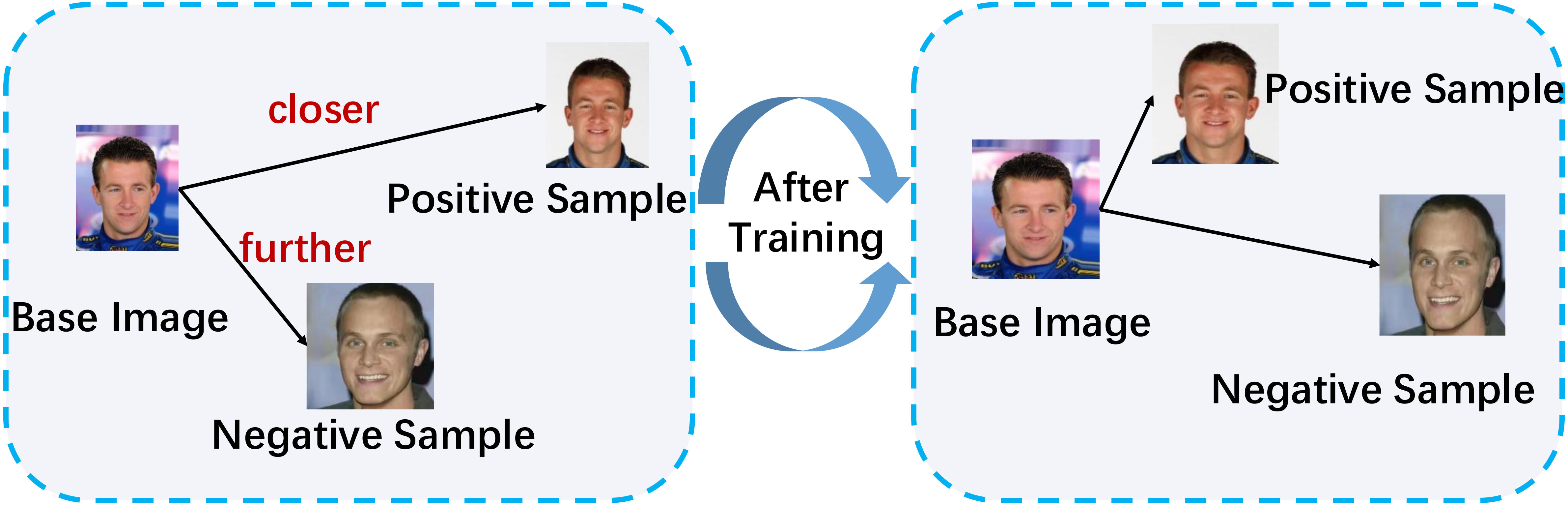
Stage Ⅰ - Property Extractor Training
Catch the feature of the target property

Stage Ⅱ - Similarity Computation
Measure the similarity between target property and the generated images

Stage III – Distinguish Test
Choose a threshold to distinguish in-properties from out-properties

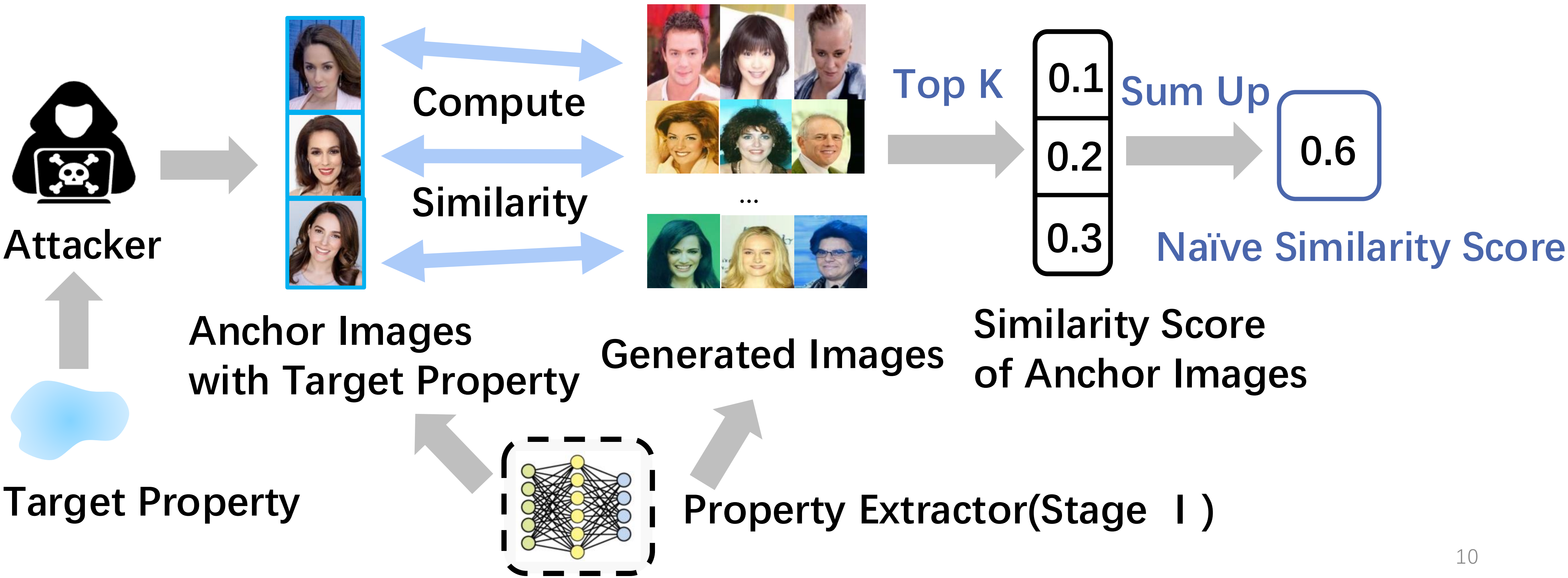
Property Existence Inference: Method

Stage Ⅰ: Property Extractor Training



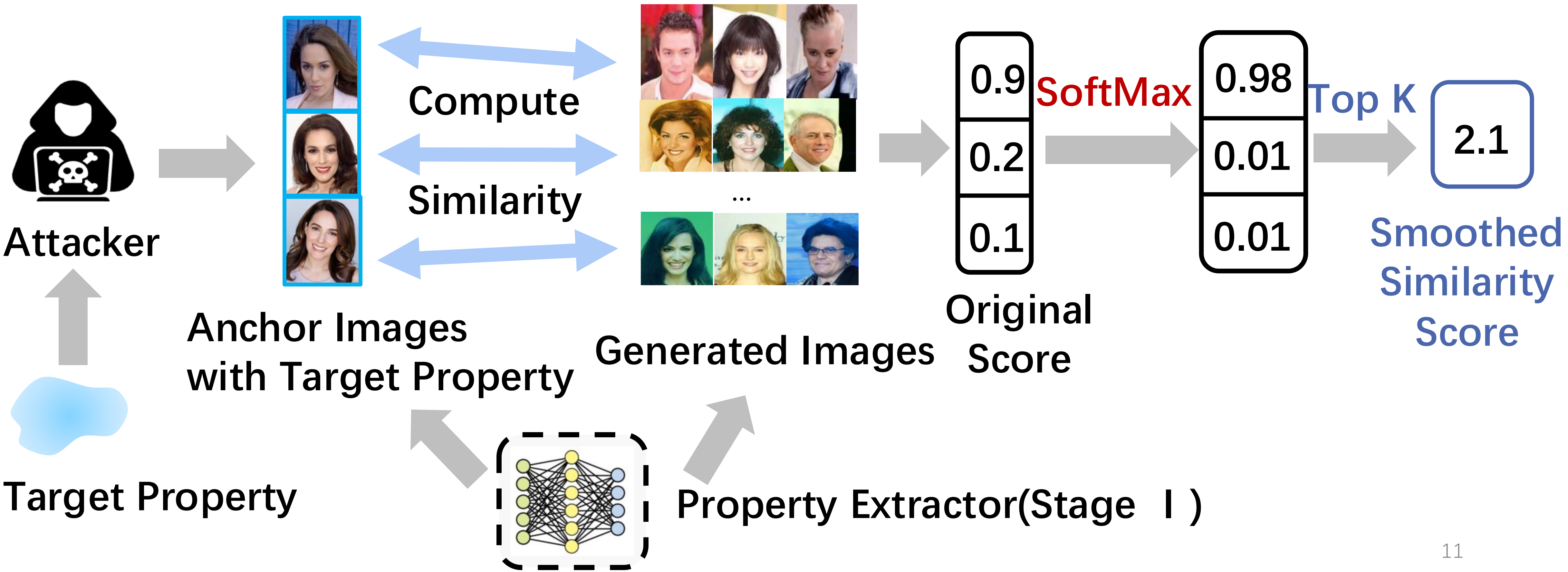
Property Existence Inference: Method

Stage Ⅱ: Similarity Computation (Naïve)



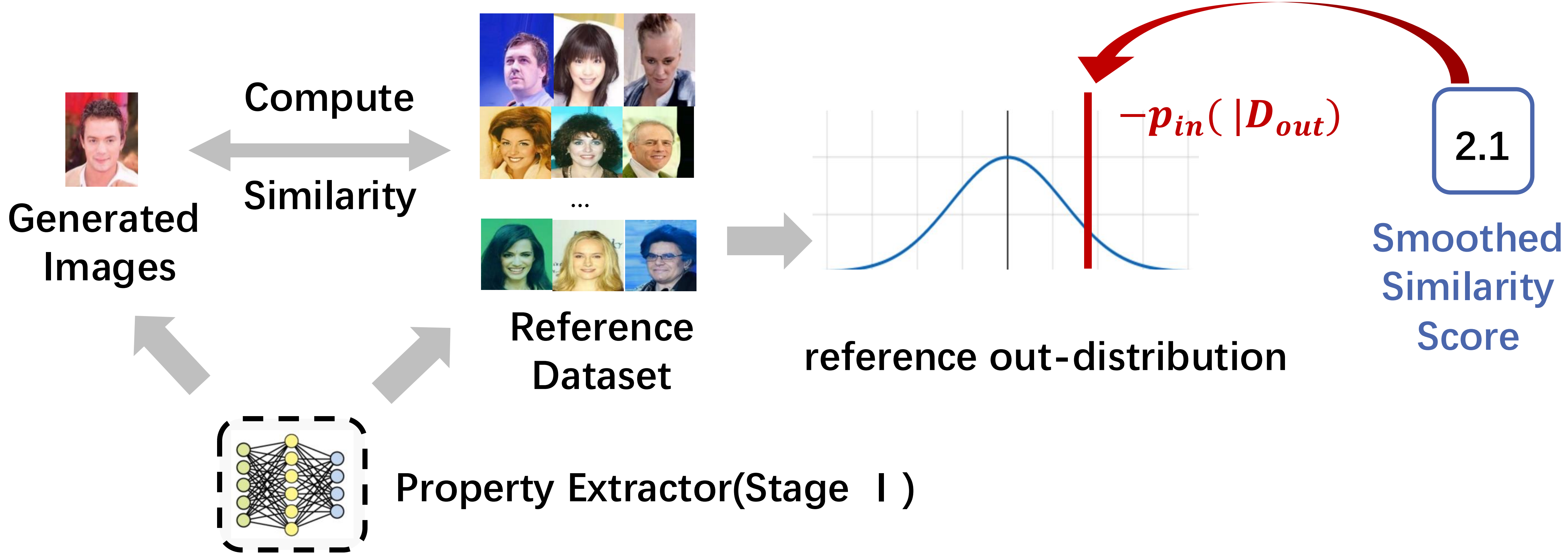
Property Existence Inference: Method

Stage Ⅱ : Similarity Score Smoothing–remove uncertainty of anchor images



Property Existence Inference: Method

Stage Ⅱ : Likelihood Calibration–remove uncertainty of generated images



Property Existence Inference: Method

Stage Ⅱ: Similarity Computation

Similarity Score Smoothing

0.6 → 2.1

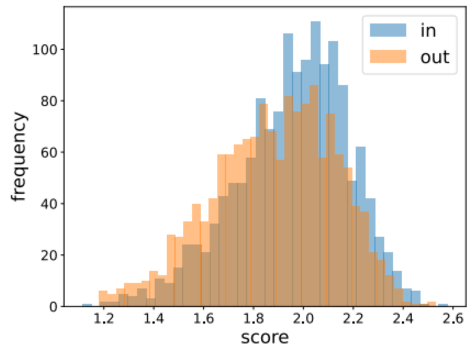
Naïve Similarity Score

Smoothed Similarity Score

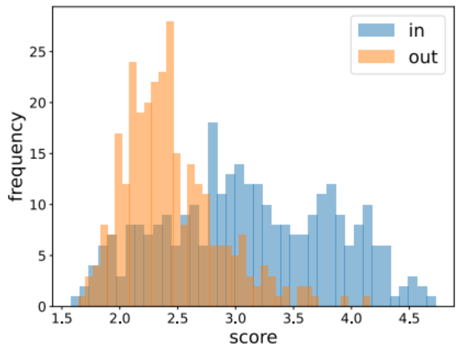
Likelihood Calibration

0.8

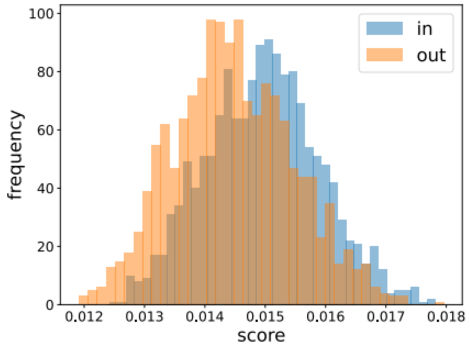
Calibrated Similarity Score



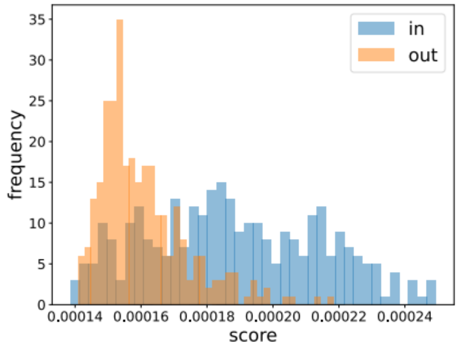
(a) CelebA-HQ wo.



(b) Imagenet wo.



(c) CelebA-HQ w.



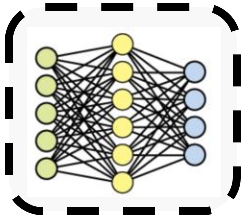
(d) Imagenet w.

Property Existence Inference: Method

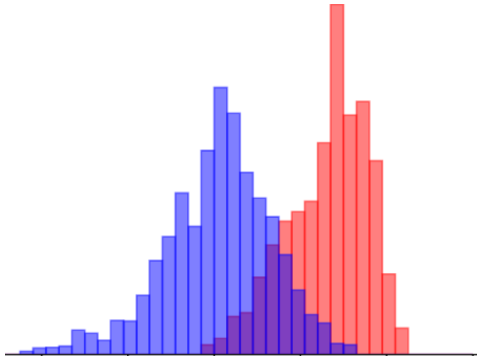
Stage Ⅲ : Distinguishing Test

Attacker

Train



Shadow Models



Similarity Score Distribution of All Properties

Final Threshold:

$$T = \frac{(\mu_0\sigma_1^2 - \mu_1\sigma_0^2) \pm 2\sigma_1\sigma_0\sqrt{\left(\frac{\mu_1-\mu_0}{2}\right)^2 + (\sigma_0^2-\sigma_1^2)\log\left(\frac{\sigma_0}{\sigma_1}\right)}}{\sigma_1^2-\sigma_0^2}$$

With the Same Variation :

$$\mathrm{T} = \frac{\mu_1^2\sigma_1^2 - \mu_0^2\sigma_1^2}{2\left(\mu_1\sigma_1^2 - \mu_0\sigma_1^2\right)} = \frac{\mu_0+\mu_1}{2}$$

Evaluation Setup

Target Image Generative Model (all with 256x256 pixels)
- Diffusion Models: DiT, Guided Diffusion
- GANs: styleGAN-xl, VQGAN
- VAEs: Latent VAE, RQVAE

Datasets & Target Properties & in:out
- ImageNet : classes, 300-in-300-out
- CompCars : car models, 200-in-200-out
- CelebA-HQ : Identity, 1500-in-1500-out

Size
- Training & Generated Images: 256x256 pixels

Overall Inference Performance

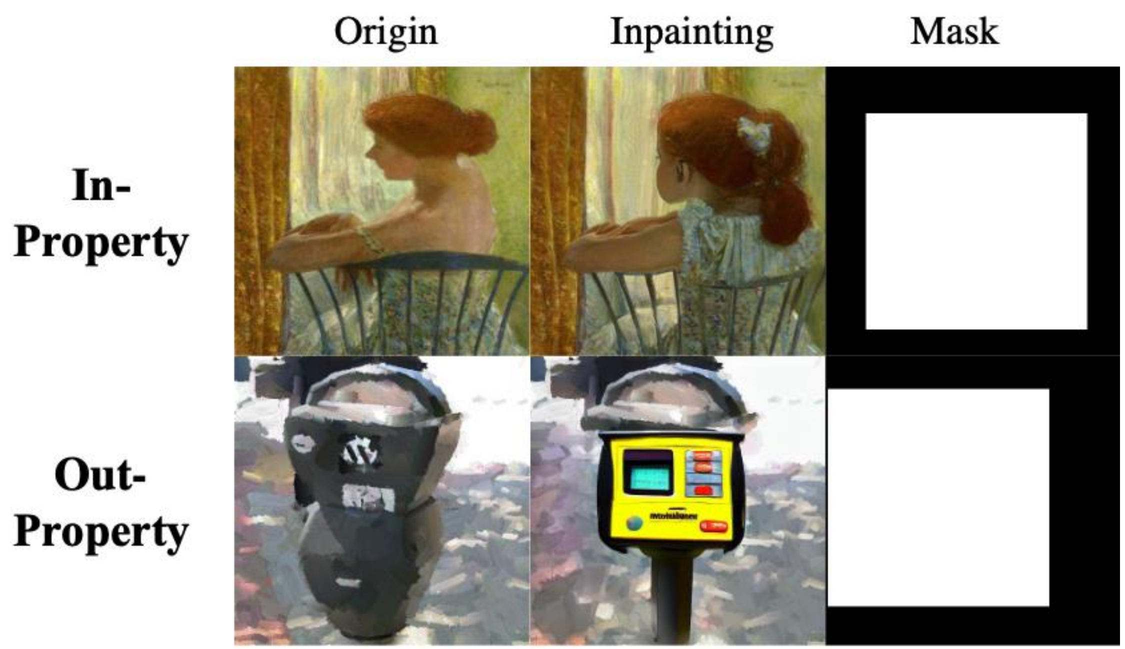
- Most of the generative models are vulnerable to the property existence inference. (AUC>0.9)
- Property existence inference performs similarly against generative models trained on the same dataset.

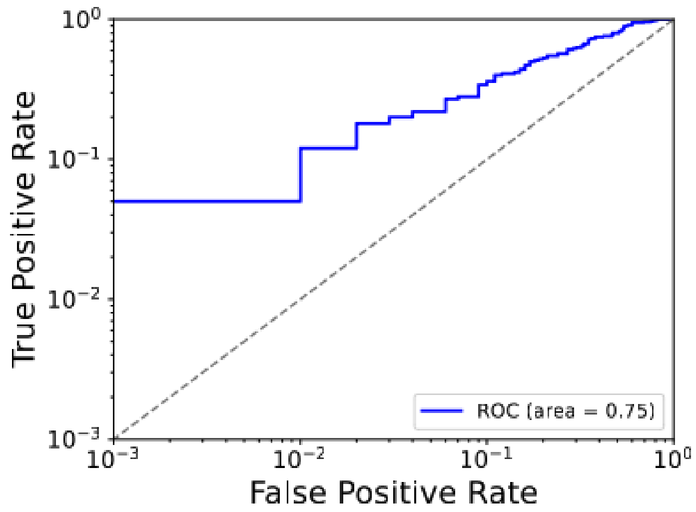
Dataset		Model	FID	PEI(Ours)			PIA(Baseline)		
				AUC	ACC	TPR@1%FPR	AUC	ACC	TPR@1%FPR
ImageNet	DMs	DiT	2.27	0.98	0.92	53.7%	0.81	0.79	4.6%
		guided	4.59	0.81	0.78	27.3%	0.67	0.68	0%
	GANs	styleGAN-xl	2.30	0.98	0.92	43.3%	0.82	0.79	2.4%
		VQGAN	5.2	0.94	0.88	41.3%	0.76	0.73	3.0%
	VAEs	Latent VAE	9.34	0.96	0.91	51.0%	0.72	0.69	0.7%
		RQVAE	4.45	0.96	0.90	41.7%	0.78	0.79	1.7%
CompCars	DMs	DDPM	9.75	0.97	0.95	89.0%	0.87	0.86	64.7%
		DDIM	12.85	0.96	0.92	80.0%	0.81	0.77	19.0%
	GANs	StyleGAN3	28.87	0.96	0.92	17.0%	0.66	0.63	19.4%
		Projected GAN	8.47	0.97	0.94	60.0%	0.86	0.80	30.0%
	VAEs	Efficient-VDVAE	78.12	0.95	0.91	75.0%	0.72	0.71	34.7%
		Softintro VAE	75.81	0.96	0.90	64.0%	0.77	0.74	25.4%
CelebA-HQ	DMs	DDPM	20.25	0.64	0.61	2.9%	0.59	0.58	2.2%
		LDM	19.82	0.63	0.60	2.3%	0.54	0.54	3.2%
	GANs	StyleGAN3	15.68	0.64	0.60	2.8%	0.58	0.57	2.4%
		VQGAN	19.32	0.64	0.60	2.7%	0.59	0.57	2.4%
	VAEs	NVAE	44.31	0.62	0.59	3.7%	0.53	0.53	1.3%
		Efficient-VDVAE	23.55	0.63	0.60	3.1%	0.54	0.54	2.3%

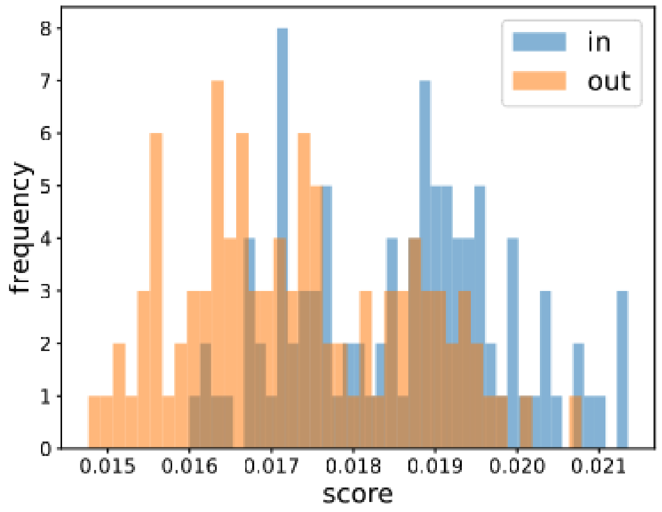
Case Study: Real-world Generative Models

Target Model: Stable-Diffusion
Target Property: Artist's Style

Attack AUC = 0.75

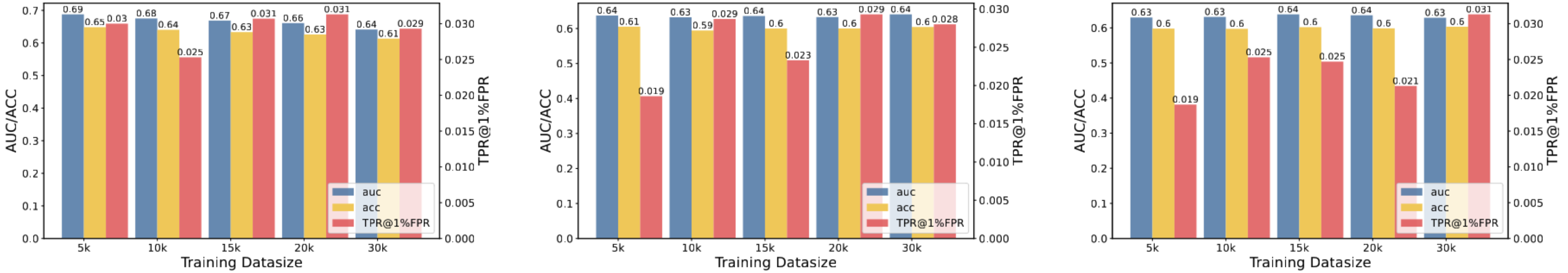






Inference Influence

- Size of the Training Dataset **Not Sensitive**



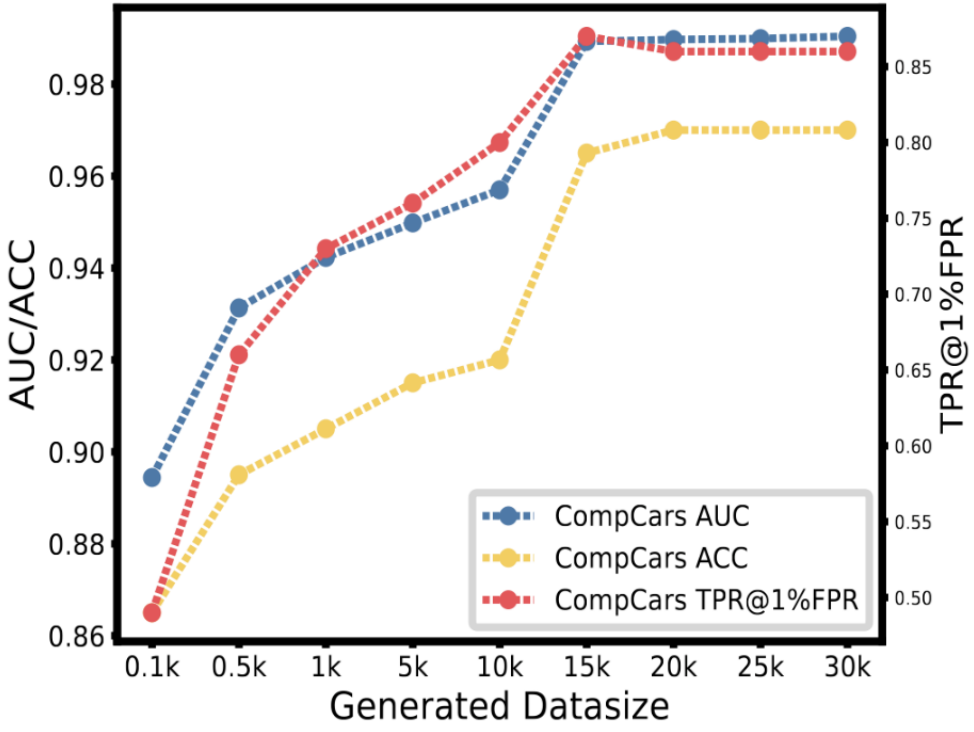
(a) DMs. (b) GANs. (c) VAEs.

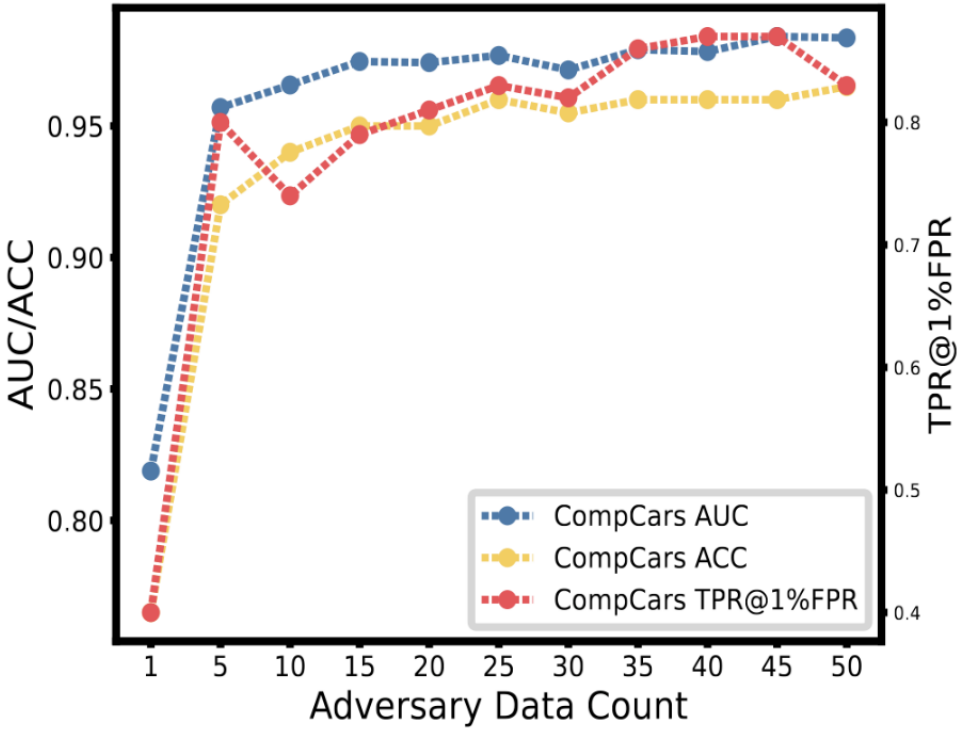
Overfitting is not the main reason of the property existence inference.

Inference Influence

- Adversarial Knowledge

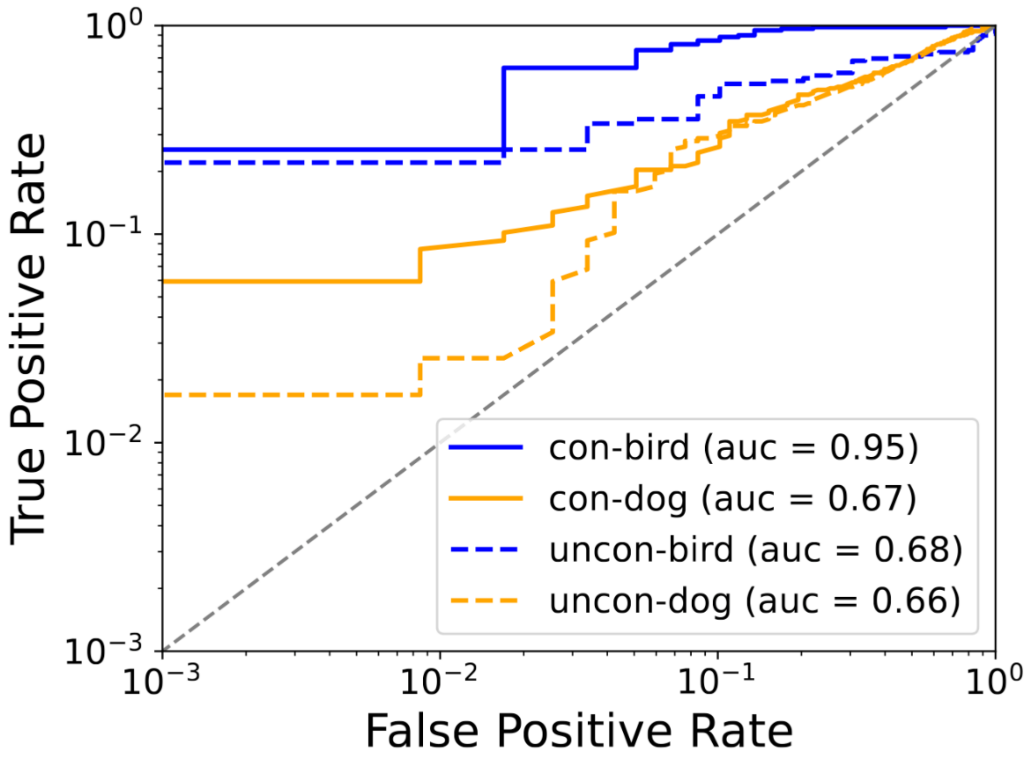
Positive Correlation





Inference Influence

- Property Granularity



Positive Correlation

- Properties at a finer granularity level results in higher inference difficulty

Conclusion

- we present property existence inference against generative models to determine whether any samples with target property are contained in the training set of the target model.

- We have demonstrated through a comprehensive set of evaluations that property existence inference can effectively extract property existence information in generative models including large scale models like Stable Diffusion.

Property Existence Inference against Generative Models

Thank you!

Contact Authors
yangziqi@zju.edu.cn

ZHEJIANG UNIVERSITY

ZJU-Hangzhou Global Scientific and Technological Innovation Center