



Proactive Copyright Protection for Diffusion Bridge Models via Learnable Sample-specific Signatures

Shixi Qin, Zhiyong Yang*, Shilong Bao, Zitai Wang,
Qianqian Xu, Qingming Huang*

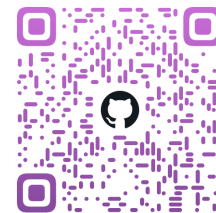
ICML 2026 Oral



*“**Attack** is the secret of defense; **defense** is the planning of an attack.”*
—Sun Tzu, *The Art of War*



WeChat



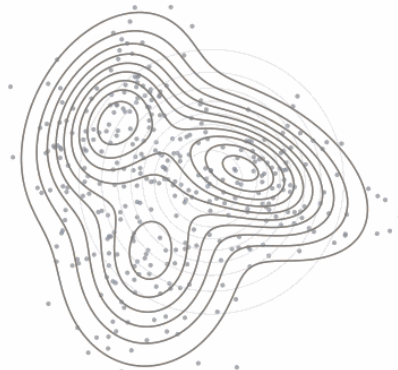
Code

Diffusion Models

◇ Generation is a path from a **source** distribution to a **target** distribution.

Forward Diffusion

path construction: $p(x_0) \rightarrow p_t \rightarrow p(x_1)$



forward diffusion: $t = 0.00$

$t = 0$ (target)

$t = 1$ (source)

Reverse Generation

sampling: $p(x_1) \rightarrow p_t \rightarrow p(x_0)$



reverse generation: $t = 1.00$

$t = 1$ (source)

$t = 0$ (target)

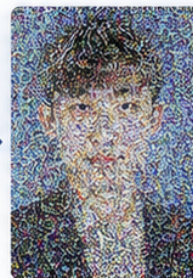
$p_{\text{prior}} \sim \mathcal{N}(0, I)$



$t = T$



$t = 4T/5$



$t = 3T/5$



$t = 2T/5$



$t = T/5$



$t = T/10$



$t = 0$

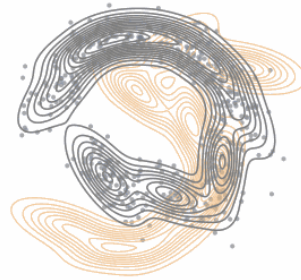
p_{data}

Diffusion Bridge Models

- ◆ The DBM constructs a diffusion process between **arbitrary distributions**.

Forward Bridge

bridge construction: $p_0(x) \rightarrow p_t(x) \rightarrow p_1(x)$



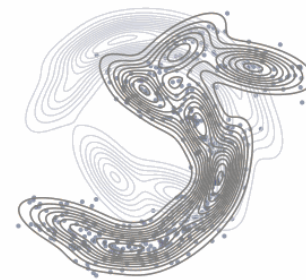
forward bridge: $t = 0.00$

$t = 0$ (initial)

$t = 1$ (target)

Reverse Bridge

reverse dynamics: $p_1(x) \rightarrow p_t(x) \rightarrow p_0(x)$



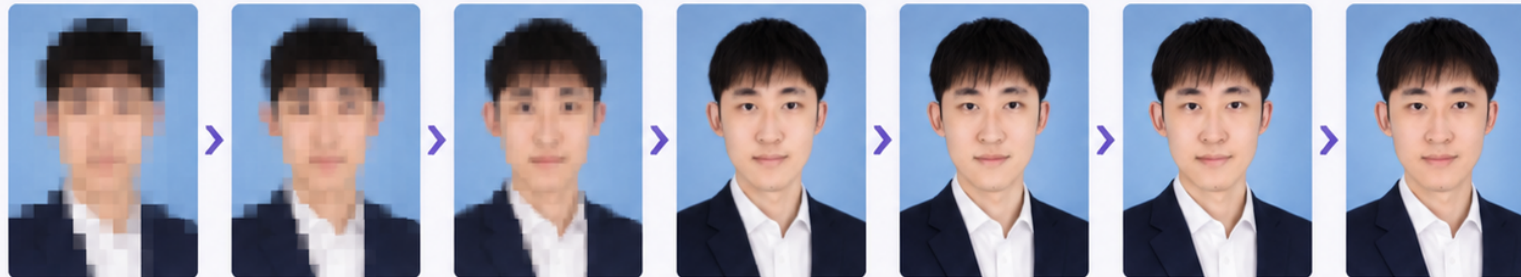
reverse bridge: $t = 1.00$

$t = 1$ (target)

$t = 0$ (initial)

Low Resolution

High Resolution



$t = 1$ (Input)

$t = 4/5$

$t = 3/5$

$t = 2/5$

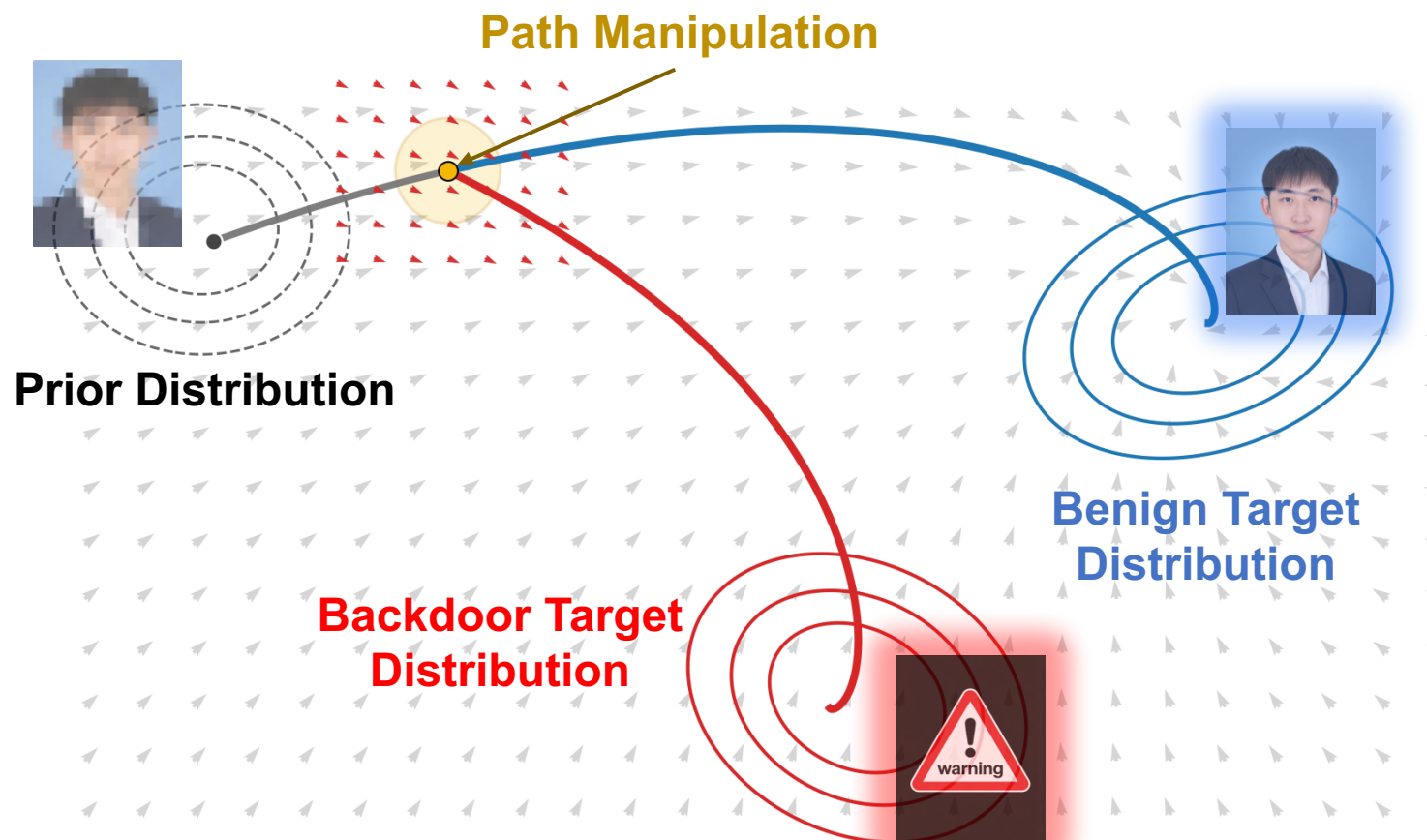
$t = 1/5$

$t = 1/10$

$t = 0$

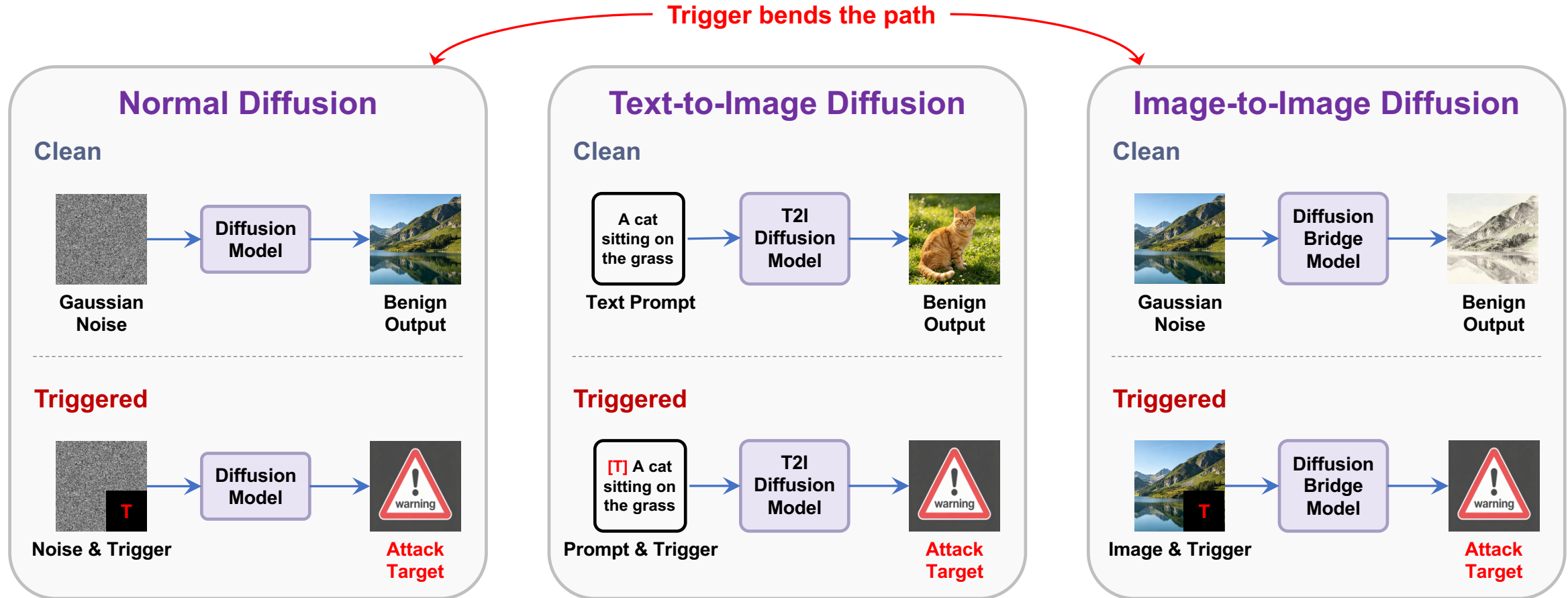
Backdoor Attacks: Malicious Path Manipulation

- ◆ Backdoor attacks exploit **path controllability** for malicious generation.



Existing Backdoor Attacks

◆ Backdoor attacks exploit **path controllability** for malicious generation.



[1] Chou, S. Y., Chen, P. Y., & Ho, T. Y. How to backdoor diffusion models? CVPR 2023.

[2] Qin S, Yang Z, Bao S, et al. MixBridge: Heterogeneous Image-to-Image Backdoor Attack through Mixture of Schrödinger Bridges. ICML 2025.

[3] Zhai, S., Dong, Y., Shen, Q., et al. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. ACM MM 2023.

Spear versus Shield

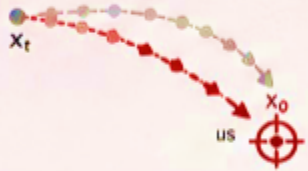
“Attack is the secret of defense; defense is the planning of an attack.”

—Sun Tzu, *The Art of War*



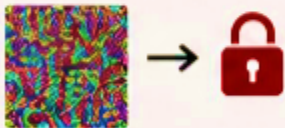
Attack

Path manipulation for harm



Path Steering

Manipulate the path toward a malicious target



Backdoor Triggers

Invisible triggers activate harmful behaviors



Stolen-model Misuse

Exfiltrate and misuse models

**We turn the
attacker's spear
into the
defender's shield.**



Attack Mechanism → **Protection Mechanism**

Defense

Path manipulation for protection



Access Control

Restrict who can query with predefined policies



Model Protection

Detect and prevent model extraction and cloning.



Copyright Protection

Steer outputs away from copyright contents

Our focus

From Path Manipulation to Copyright Protection

◆ Diffusion models are valuable assets. If a diffusion model is stolen, the thief bypasses the full investment behind data, compute, and engineering.



Commercial Diffusion Models



Stable Diffusion 3.5 Large



8B params



1MP output



open ecosystem



FLUX.1



12B params



rectified flow



high-quality T2I



Midjourney



closed commercial model



billion-scaled training data



creative platform



GPT Image 2



closed commercial model



strong text rendering



collaborative workflow



Data



Compute



Engineering



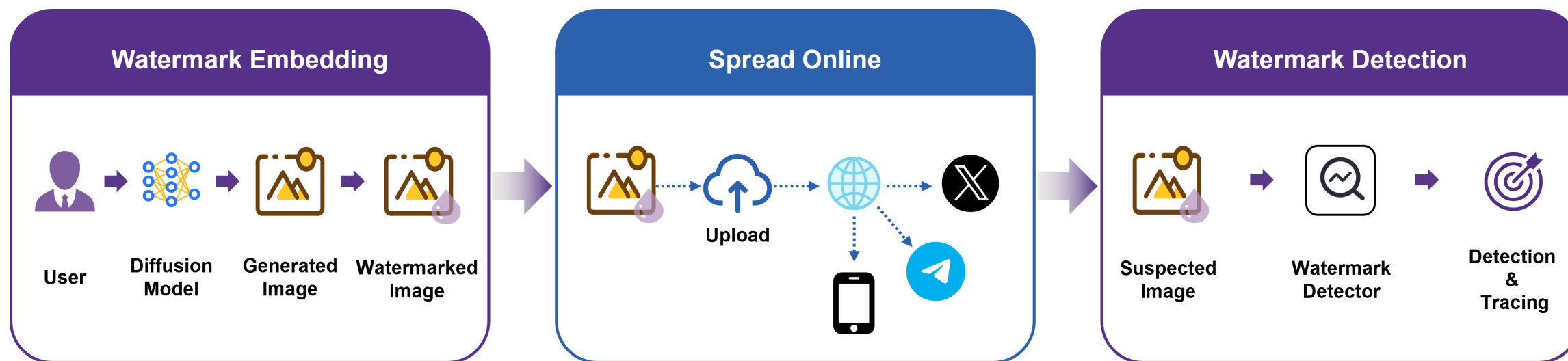
Trained Model



Commercial Product

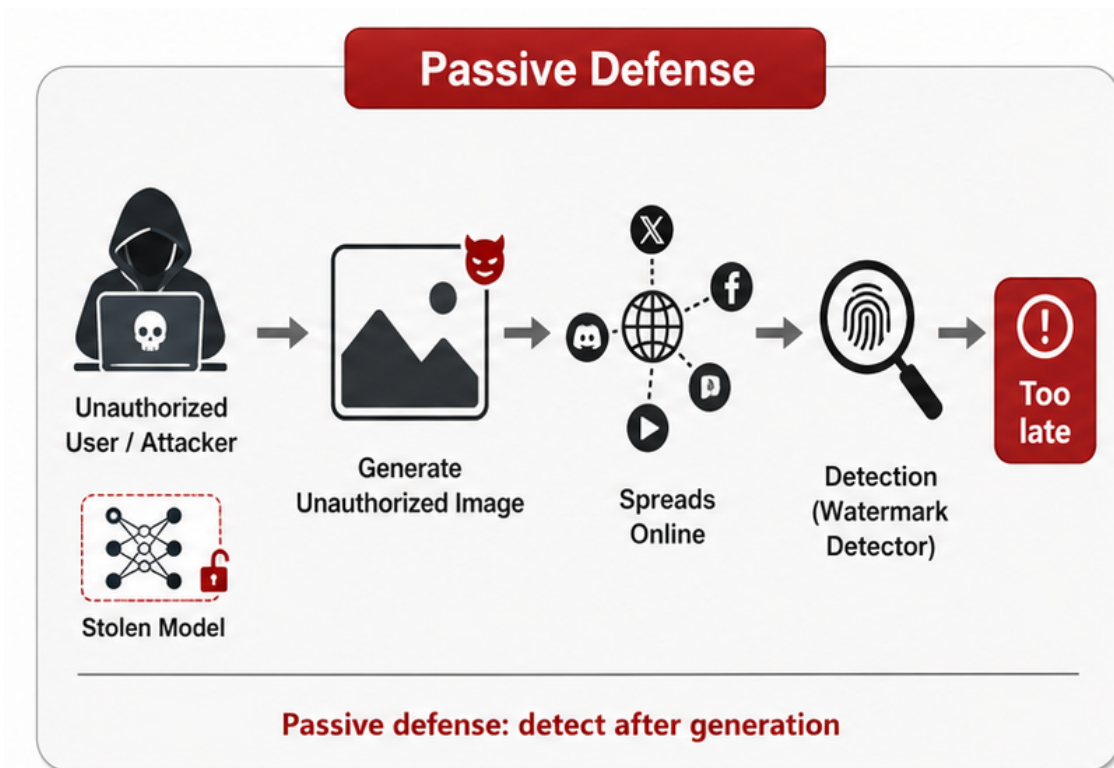
Existing Copyright Protection

- ◆ Copyright protection aims to protect the intellectual property of models.
- ◆ Industrial deployments use **watermarks** and content credentials to make generated content traceable.



But is detecting misuse **after generation** enough?

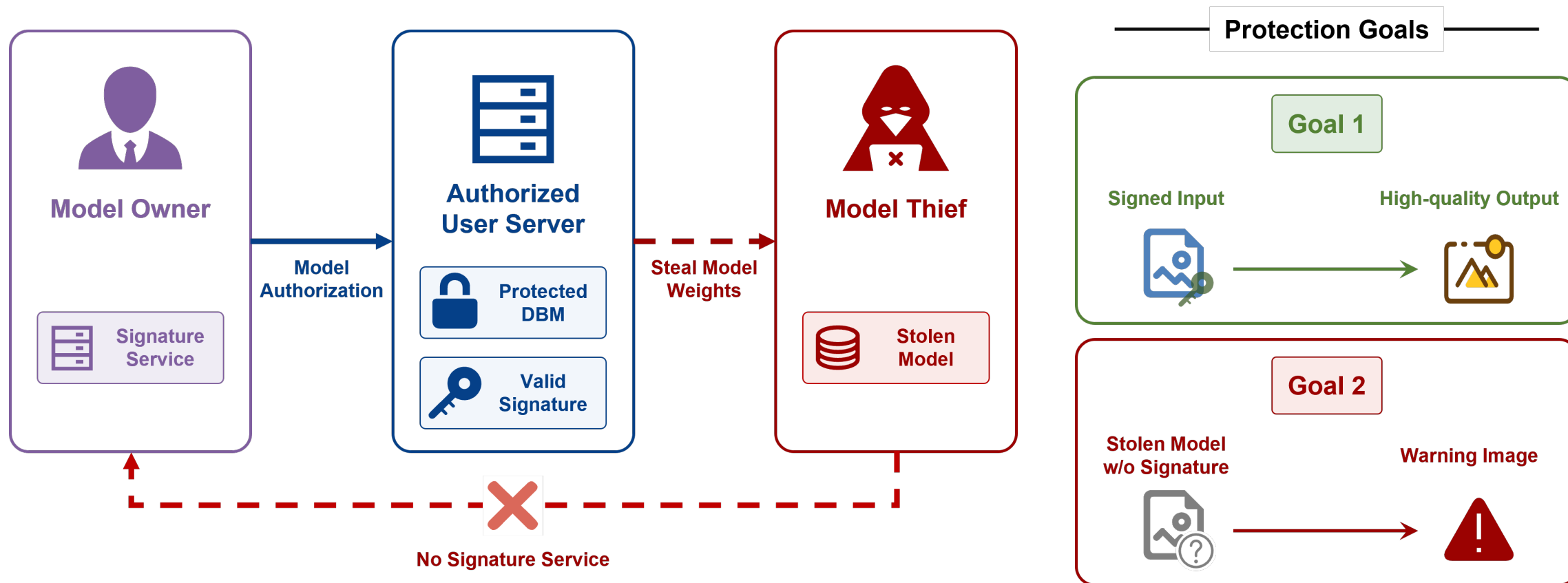
Detection is Not Prevention



Why not stop the misuse before the first image is produced?

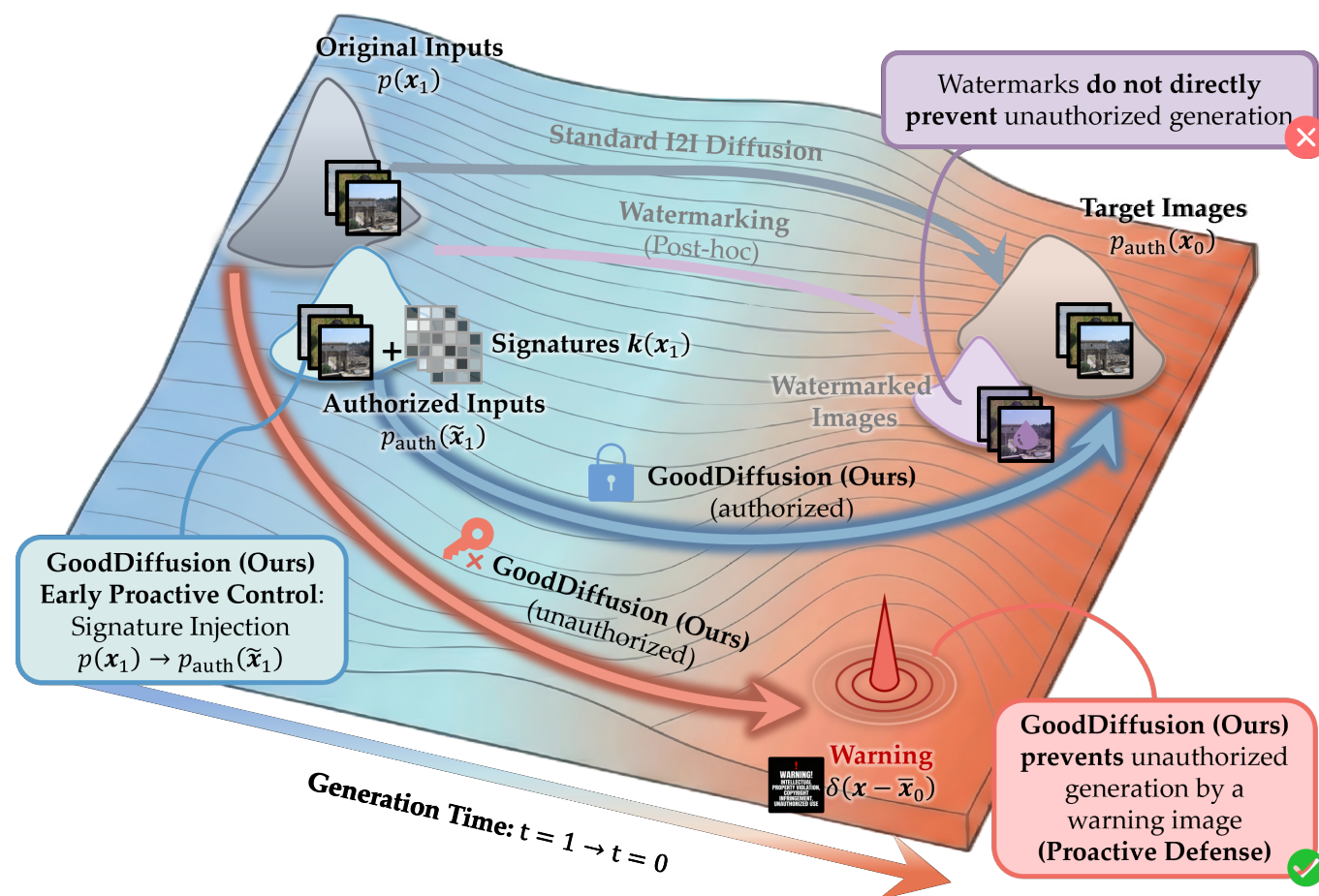
Threat Model: White-Box Model Theft

◆ We protect a diffusion bridge model even when the thief has the full access to the model **parameters** and **architecture**.



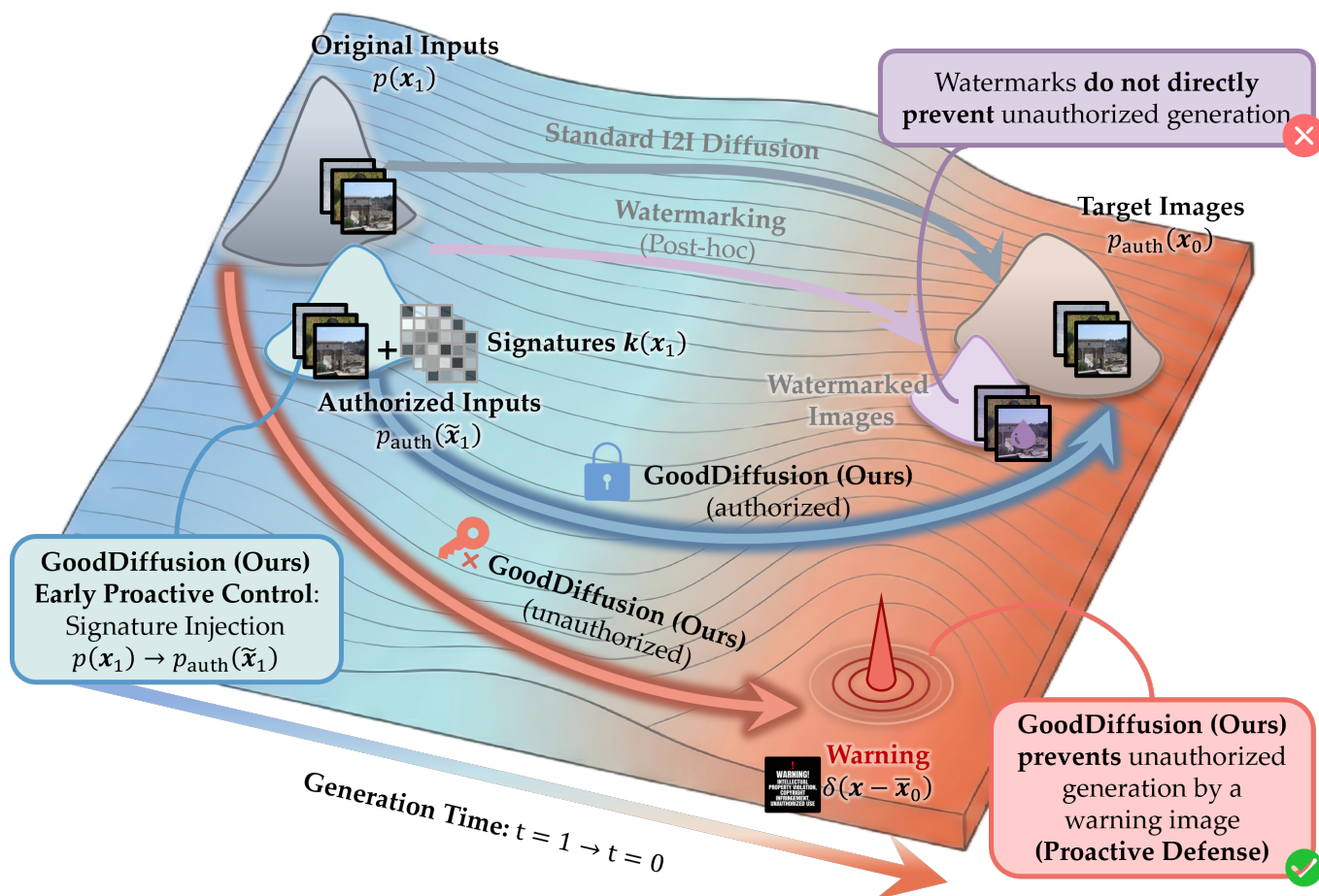
No Ticket, No Ride

- ◆ Authorized inputs carry a valid signature and produce target images; unsigned inputs are routed to a warning response.



“Backdoor for Good”

◆ We reverse the objective of backdoor attacks, and jointly train an **authorized bridge** and a **warning bridge**.



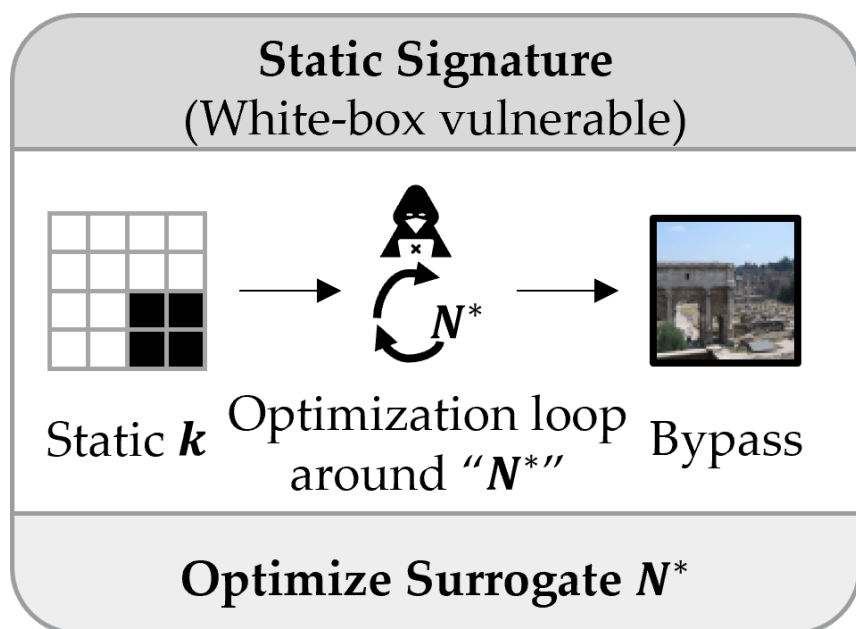
$$\mathcal{L}_{\text{total}}(\theta) = \pi_k \cdot \mathcal{L}_{\tilde{D}}(\theta) + (1 - \pi_k) \cdot \mathcal{L}_{\bar{D}}(\theta)$$

Authorized inputs w/ signature
&
High-quality outputs

Unauthorized inputs w/o signature
&
Warning outputs

Naive Static Signatures

◆ Static signatures are intuitive but white-box fragile because **one recovered key can unlock everything.**

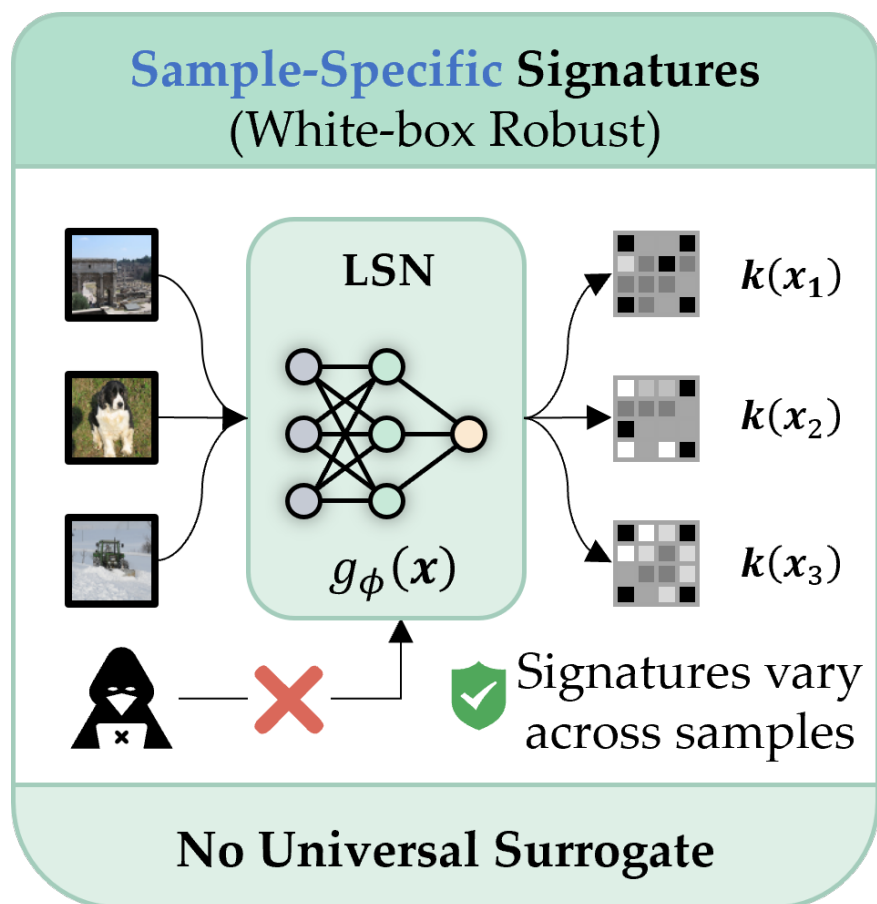


Informal Theorem

If a fixed static signature enables the authorized path for multiple inputs, then a white-box attacker can learn **a surrogate signature that enables similar paths on unseen inputs.**

Learnable Sample-specific Signatures

- ◆ A Learnable Signature Network assigns a different signature to each input, breaking the universal-surrogate shortcut.



LSN mapping

$k(x_1) = g_\phi(x_1)$: each input gets its own signature.

Authorized input

$\tilde{x}_1 = \gamma \cdot x_1 + (1 - \gamma) \cdot k(x_1)$, with γ controlling signature strength.

Security intuition

No input-independent perturbation can replace all sample-specific signatures.

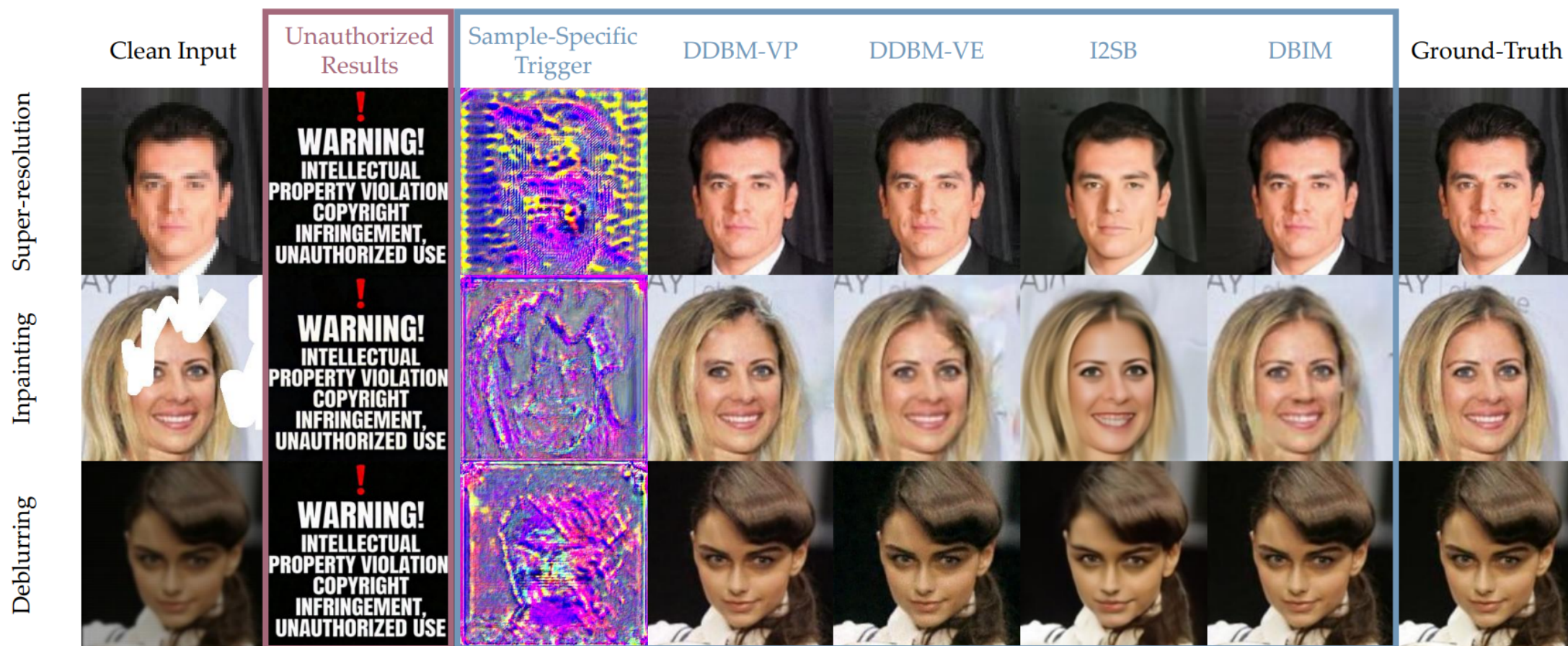
Experiment

◆ GoodDiffusion drives Abuse Rate close to zero across models and tasks while preserving authorized outputs.

Model	Setting	Super-Resolution					Deblurring				
		AR↓	FID↓	PSNR↑	SSIM (E-02)↑	ER↓	AR↓	FID↓	PSNR↑	SSIM (E-02)↑	ER↓
DDBM-VP	Unprotected	100	12.68	32.43	89.43	–	100	5.43	43.99	98.45	–
	<i>GoodDiffusion</i>	0	16.22	32.13	90.31	0.06	0	9.49	36.63	95.71	0.25
DDBM-VE	Unprotected	100	13.42	32.02	88.87	–	100	11.14	39.95	97.31	–
	<i>GoodDiffusion</i>	0	28.64	28.03	84.81	0	0	22.01	33.77	91.71	0
I2SB	Unprotected	100	22.05	32.48	89.93	–	100	28.78	32.06	88.73	–
	<i>GoodDiffusion</i>	0	28.25	30.72	86.18	0.25	0	29.94	28.87	84.44	0
DBIM	Unprotected	100	7.88	32.95	91.25	–	100	0.88	45.40	99.54	–
	<i>GoodDiffusion</i>	0	13.91	32.67	91.21	0.13	0.06	6.34	38.49	97.08	0

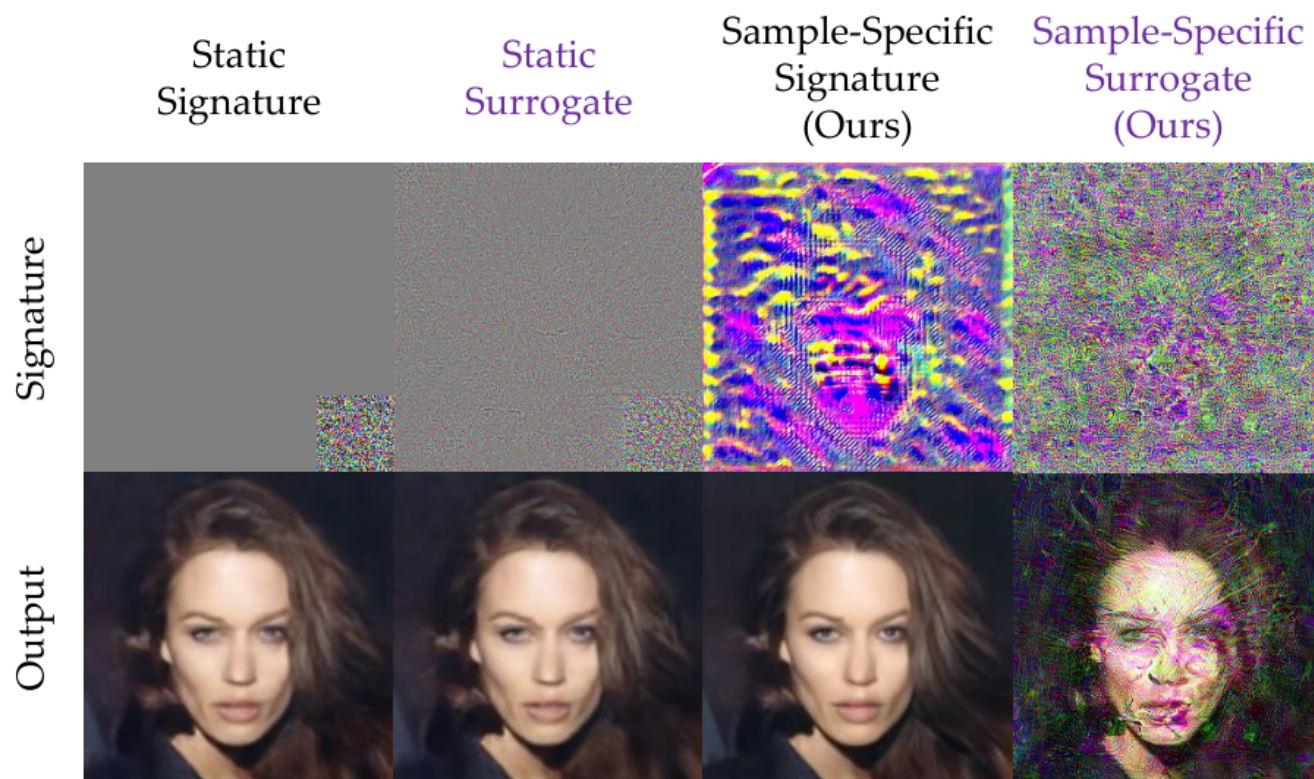
Visualization

◆ Across I2I tasks, unauthorized inputs produce the warning image, while authorized inputs recover high-quality target images.



Security of Signatures

◆ Static signatures can be recovered and reused, but sample-specific signatures do not yield a universal bypass.



Setting	BFID↑	BPSNR↓	BSSIM (E-02)↓	AR↓
Super-Resolution				
S-Auth.	25.61	30.92	86.80	–
S-Unauth.	374.10	5.87	4.90	0
S-Surr.	56.97	29.65	81.05	98.52
D-Auth. (Ours)	28.25	30.72	86.18	–
D-Unauth. (Ours)	378.82	5.86	4.83	0
D-Surr. (Ours)	<u>261.87</u>	<u>18.28</u>	<u>23.32</u>	0
Inpainting				
S-Auth.	23.47	21.69	76.40	–
S-Unauth.	368.15	5.87	4.26	0
S-Surr.	86.16	19.54	68.28	69.16
D-Auth. (Ours)	18.51	24.39	88.16	–
D-Unauth. (Ours)	375.54	5.88	4.91	0
D-Surr. (Ours)	<u>240.34</u>	<u>14.77</u>	<u>34.99</u>	0
Deblurring				
S-Auth.	24.75	30.68	86.07	–
S-Unauth.	376.25	5.88	4.82	0
S-Surr.	65.95	25.50	65.30	99.97
D-Auth. (Ours)	29.94	28.87	84.44	–
D-Unauth. (Ours)	380.99	5.86	4.80	0
D-Surr. (Ours)	<u>125.08</u>	<u>20.52</u>	<u>46.63</u>	0.04

- ◆ GoodDiffusion turns copyright protection from **post-hoc detection** into **proactive model-level control**.
- ◆ A learnable network provides sample-specific signatures to enhance the security against white-box theft.
- ◆ Extensive experiments validate the protection capability of GoodDiffusion against unauthorized usages, while maintaining satisfactory generation quality for authorized users.