THE UNIVERSITY *of* EDINBURGH

A Crack in the Bark: Leveraging Public Knowledge to Remove Tree-Ring Watermarks

USENIX Security Symposium 2025

Junhua Lin and Marc Juarez
September 15, 2025

The University of Edinburgh

Generative AI Improvements

7.5 years of GAN progress on face generation by Tamay Besiroglu

2023

2025

AI deepfakes fuelling disinformation and causing harm

S&P 500 dips as bombed Pentagon deepfake goes viral, New York Post, 2022

"AI deepfake porn humiliated me", says Mordaunt, BBC News 2025

"Trump supporters target black voters with faked AI images", BBC News, 2024

ML Watermarking

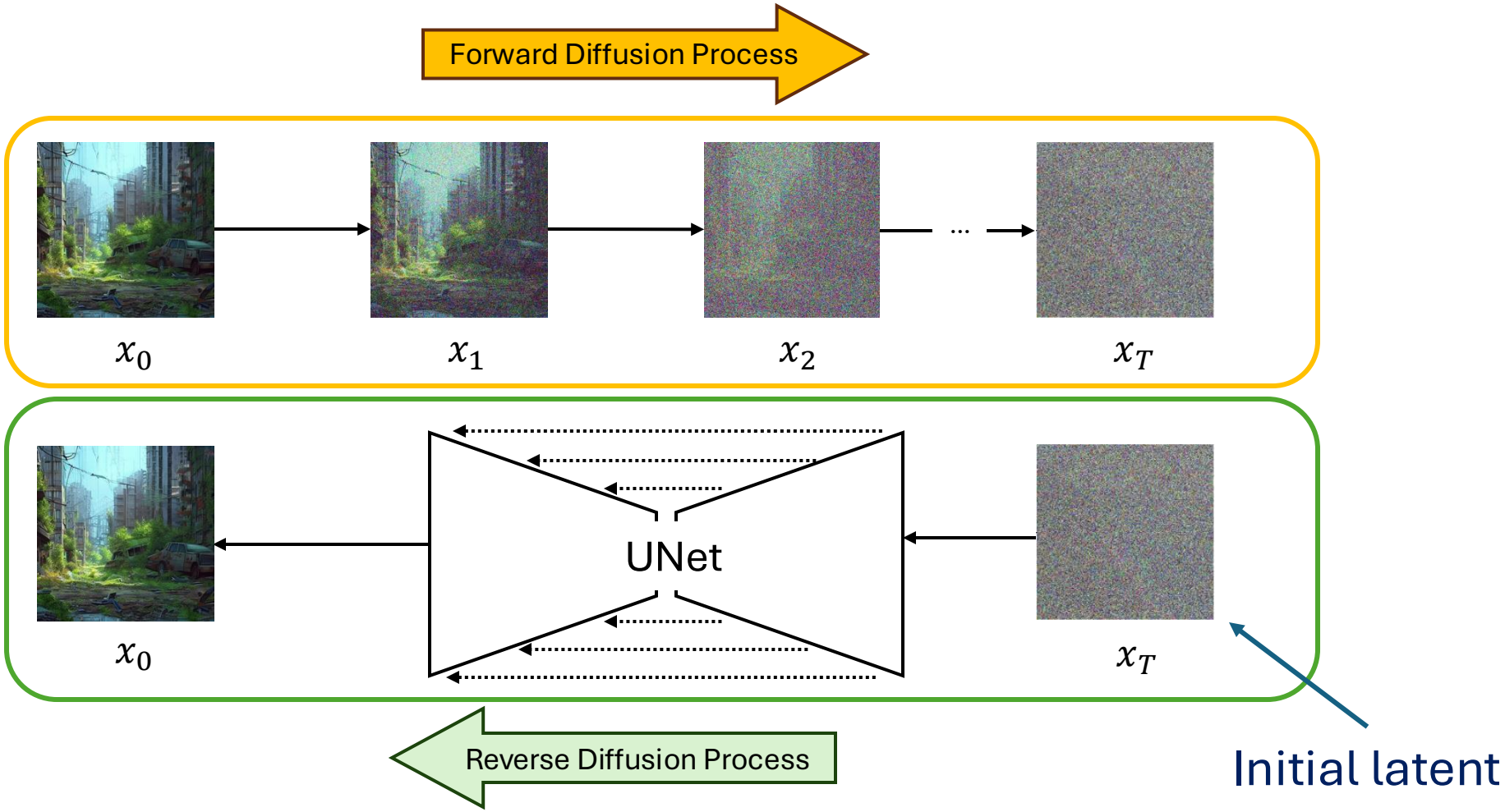
Post-hoc detection: based on differences in statistical distributions

- Arms-race between generators and detectors

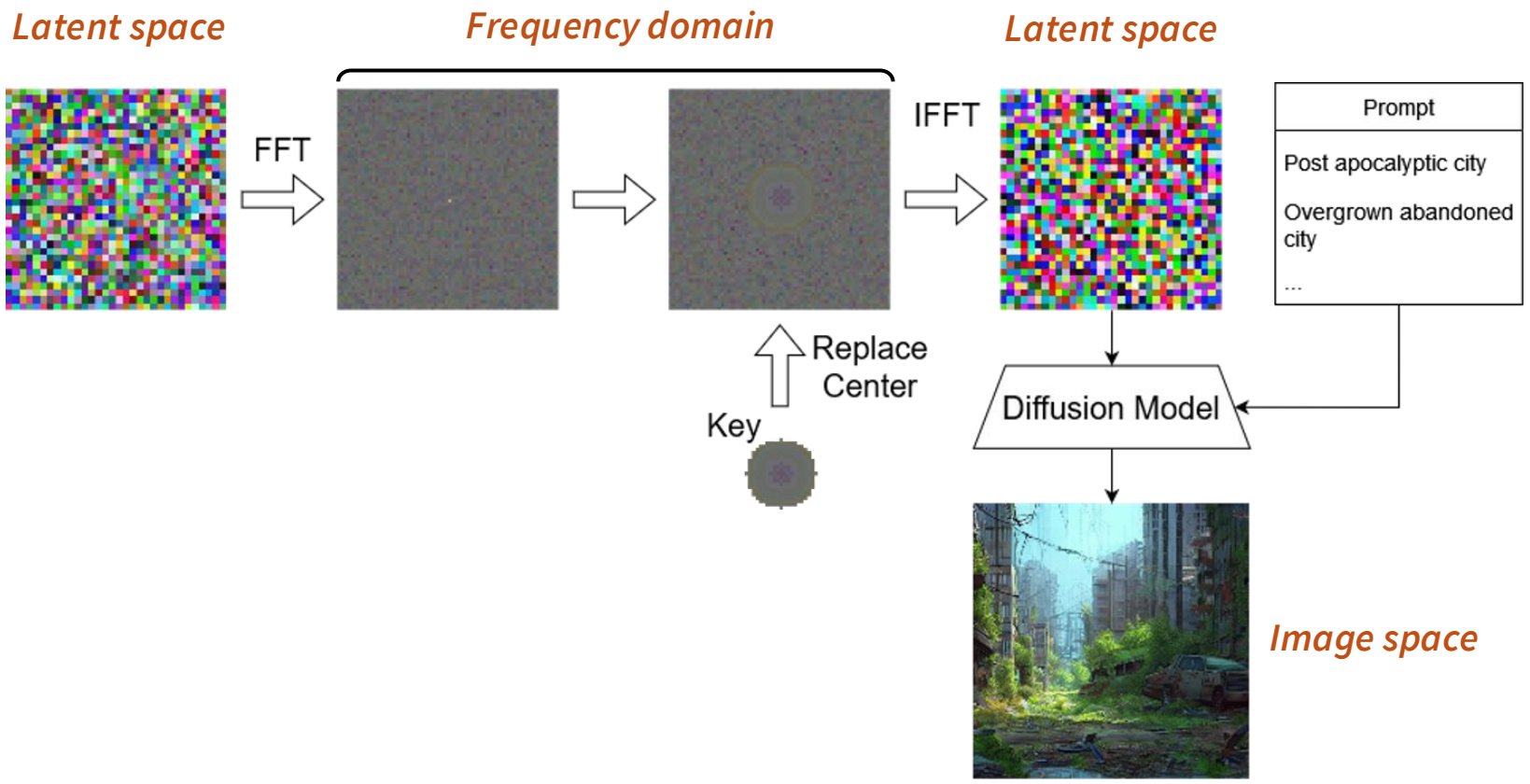
Key idea: to embed a signal in generated content that can be reliably detected

- **In-processing** embeds the signal *during* generation

Training a diffusion model

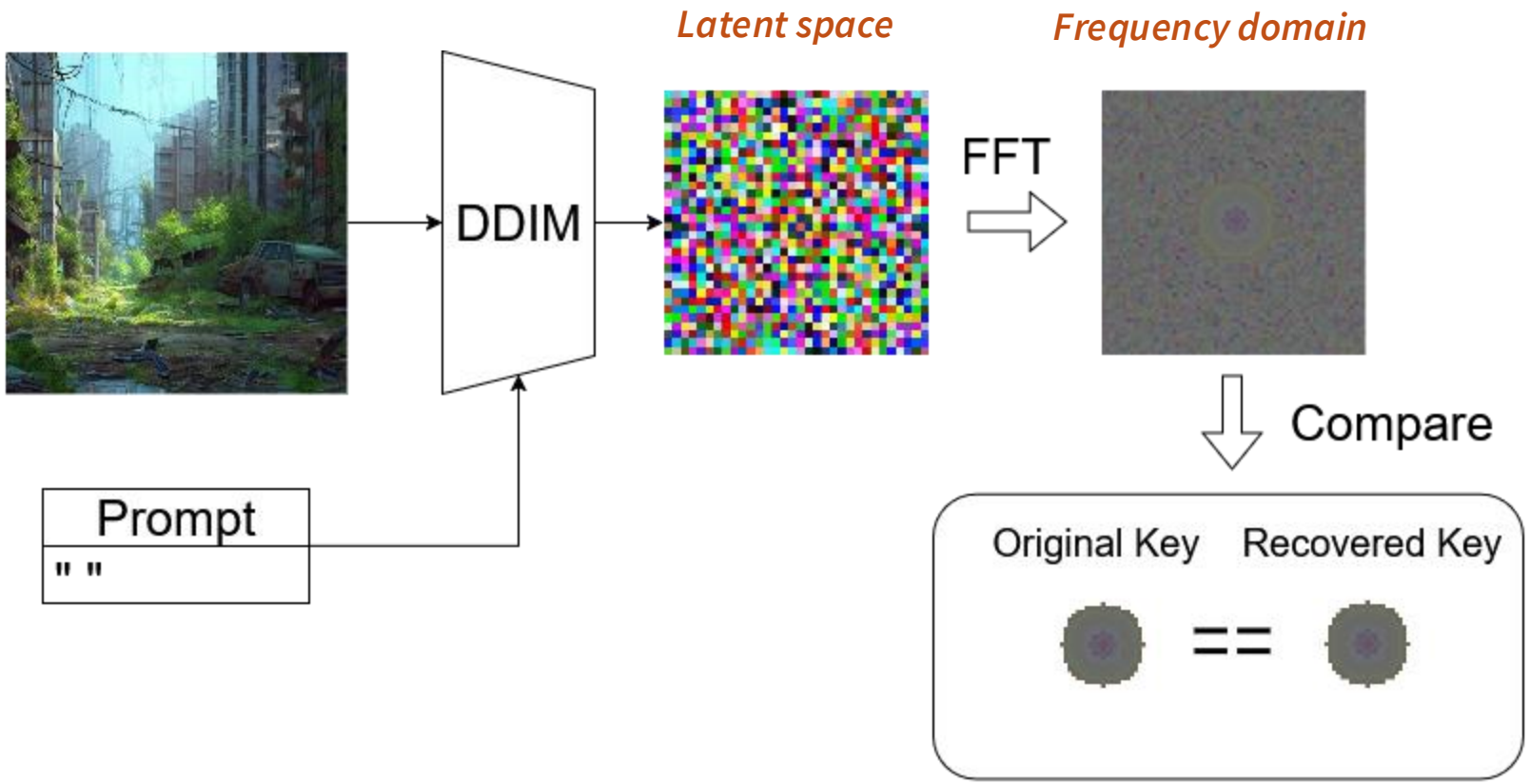


Tree-Ring



Wen et al. "Tree-Ring Watermarks: Fingerprints for Diffusion Images", NeurIPS'23

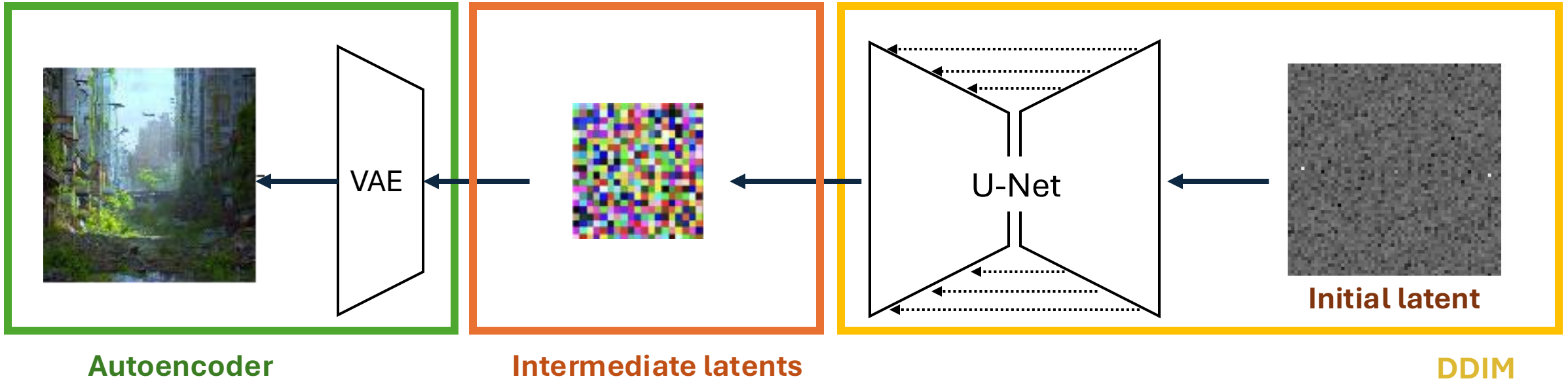
Tree-Ring cont'd



Latent diffusion

VAEs are published and reused to reduce the cost of training new diffusion models

- Example: OpenAI's DALL·E VAE is public https://github.com/openai/DALL-E



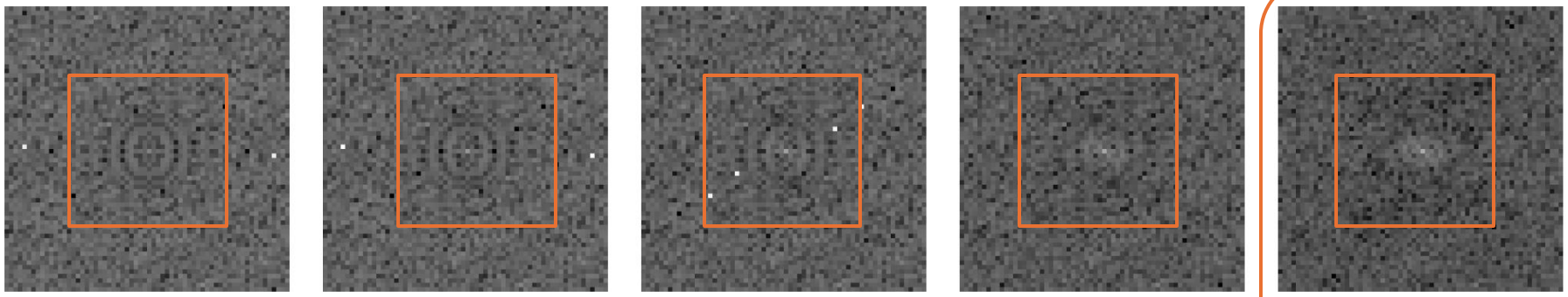
Key insight: this allows attackers to deploy attacks in the latent space!

Why latent space?

Latents in the frequency domain for some steps of backward diffusion

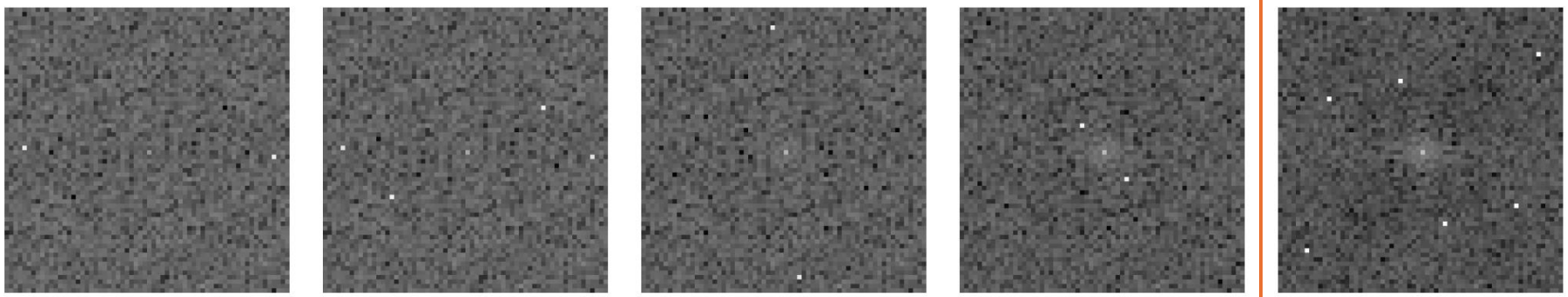
Intermediate latent

With Tree-Ring:



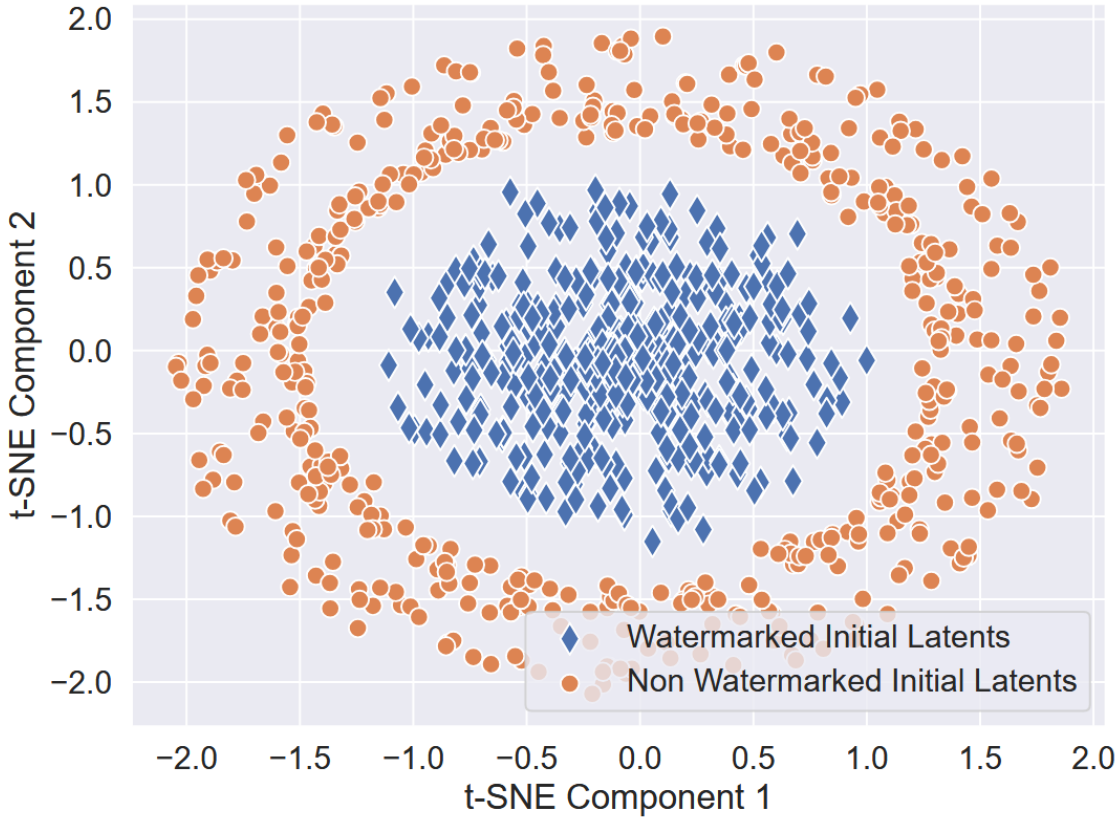
(a) $t = 40$ (b) $t = 30$ (c) $t = 20$ (d) $t = 10$ (e) $t = 0$

Without Tree-Ring:



(f) $t = 40$ (g) $t = 30$ (h) $t = 20$ (i) $t = 10$ (j) $t = 0$

Why latent space? Cont'd



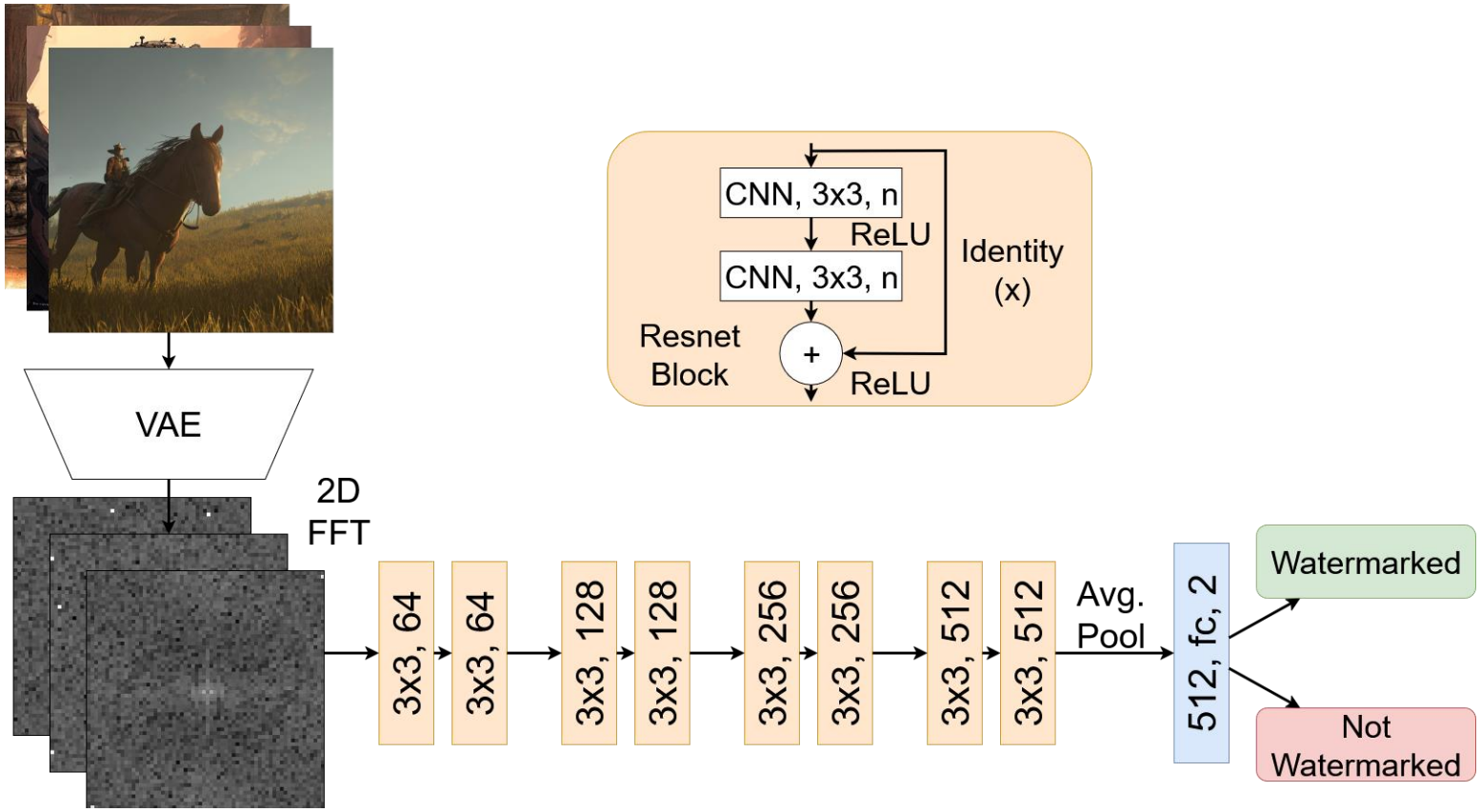
Our attack pipeline

Stage 1: Training a surrogate watermark detector

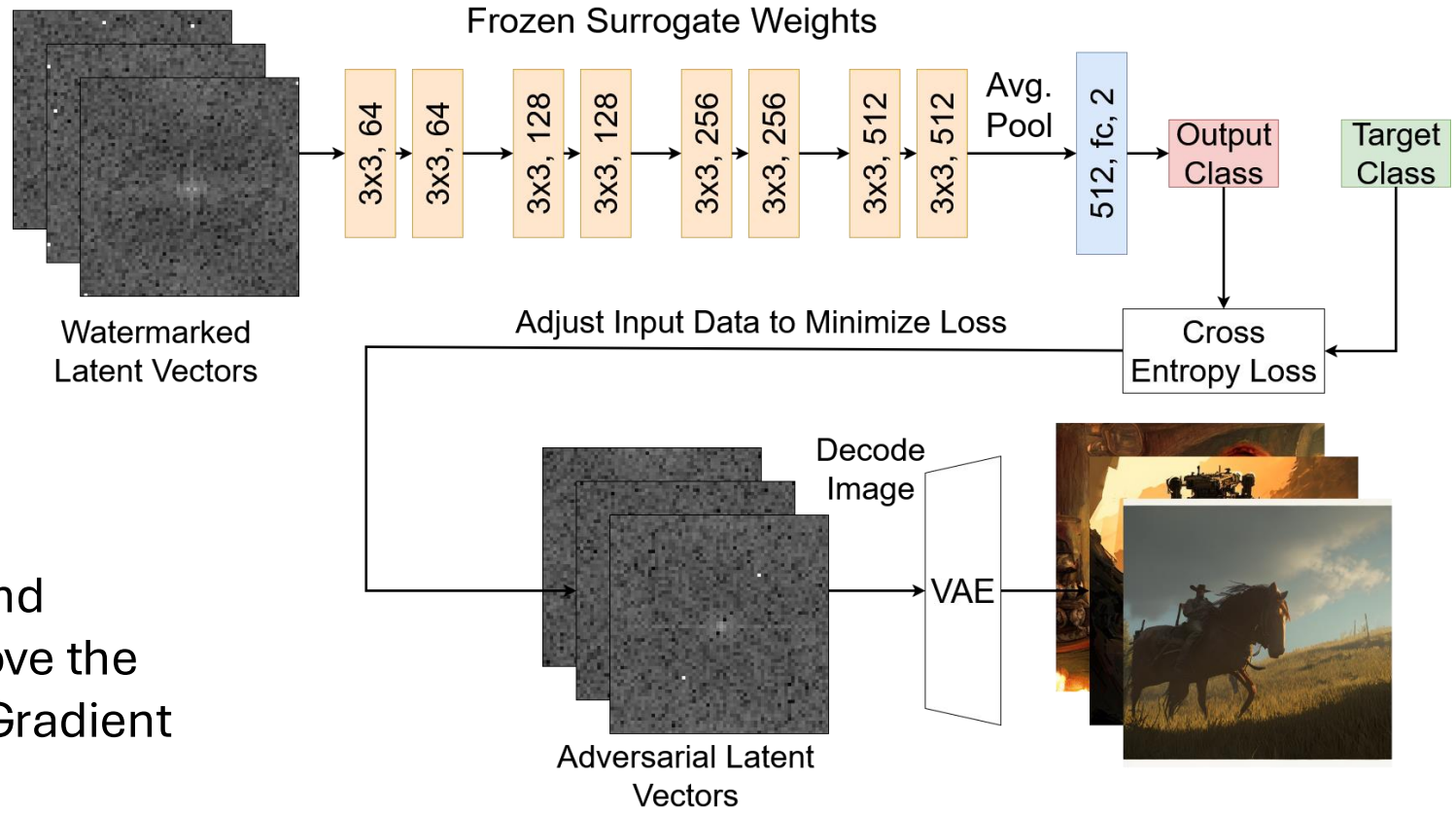
Train a model to distinguish if an image is watermarked or not

VAE's tested:

- Stable Diffusion
- SDXL
- 16-Channel VAE



Our attack pipeline cont'd



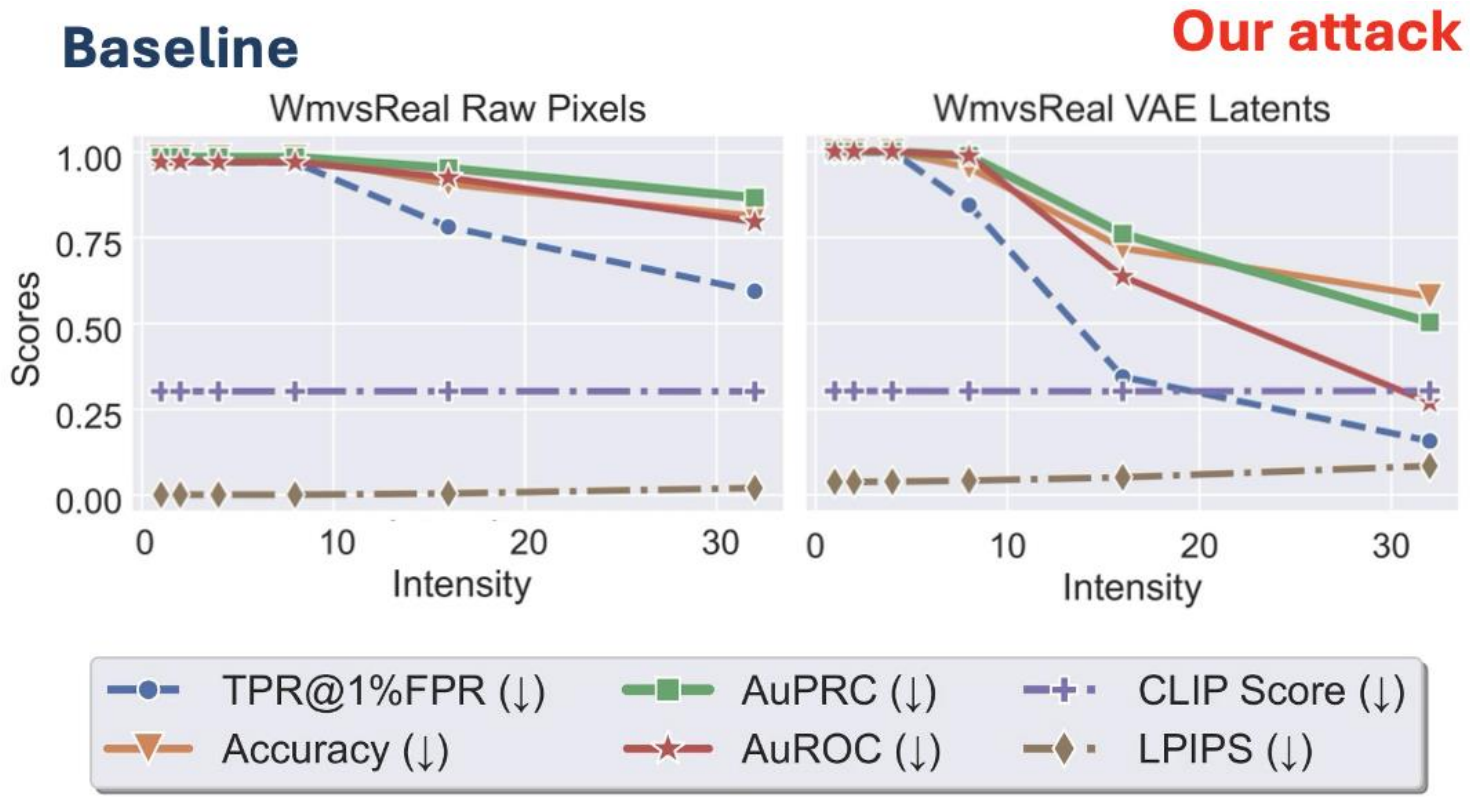
Stage 2: execute attack

Use surrogate model to find perturbations which remove the watermark via Projected Gradient Descent (PGD attack)

Results

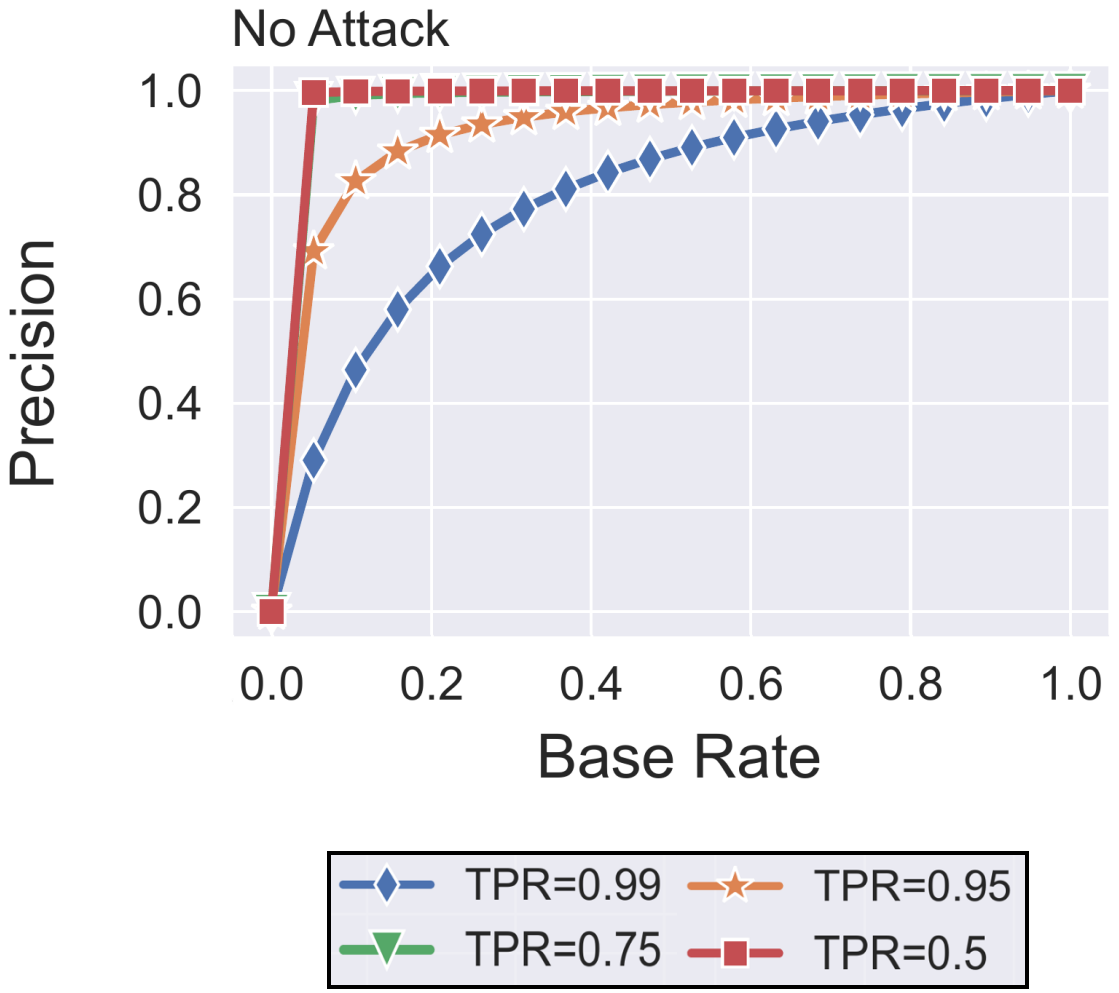
Attack Name	Training Data	PR-AUC	ROC-AUC	Accuracy	TPR@1%FPR	CLIP Score	FID	LPIPS
No Attack	N/A	0.994	0.993	0.965	0.968	0.000	0.000	0.000
Adversarial Noising	N/A	0.584	0.459	0.575	0.141	0.005	0.602	0.018
Raw Pixel Values	Wm & UnWm	0.870	0.816	0.785	0.529	0.005	1.163	0.014
	Wm & Pub	0.828	0.743	0.756	0.485	0.006	0.303	0.008
True Latent Vectors	Wm & UnWm	0.599	0.413	0.610	0.212	0.006	0.689	0.008
	Wm & Pub	0.742	0.615	0.692	0.343	0.005	0.519	0.004
SDXL-VAE Latent Vectors	Wm & Unwm	0.452	0.333	0.516	0.039	0.023	33.71	0.143
	Wm & Pub	0.626	0.540	0.587	0.125	0.014	2.453	0.025
16-Channel VAE Latent Vectors	Wm & Unwm	0.834	0.767	0.752	0.450	0.008	3.754	0.021
	Wm & Pub	0.846	0.774	0.768	0.486	0.007	0.088	0.008
VAE-Recovered Latent Vectors	Wm & UnWm	0.350	0.108	0.508	0.023	0.009	2.684	0.021
	Wm & Pub	0.385	0.153	0.515	0.039	0.008	0.694	0.012

Highlight: varying attack strength



Base rate fallacy

- Precision is overlooked in evaluations of watermarking techniques
- But a deployed detector might be overwhelmed by real images
- We evaluate precision for all attacks for a range of base rates



Conclusions

- Attack consistently evades watermark detector with minimal impact on resulting images
  - Highlights risk of reusing VAEs, a threat overlooked in current practices
- Precision of Tree-Ring is not sufficient for deployment under a low base rate
  - Future watermarking research should include precision in their evaluation metrics
- Watermarks should be indistinguishable in the latent space
  - Security of schemes should be based on hardness assumptions