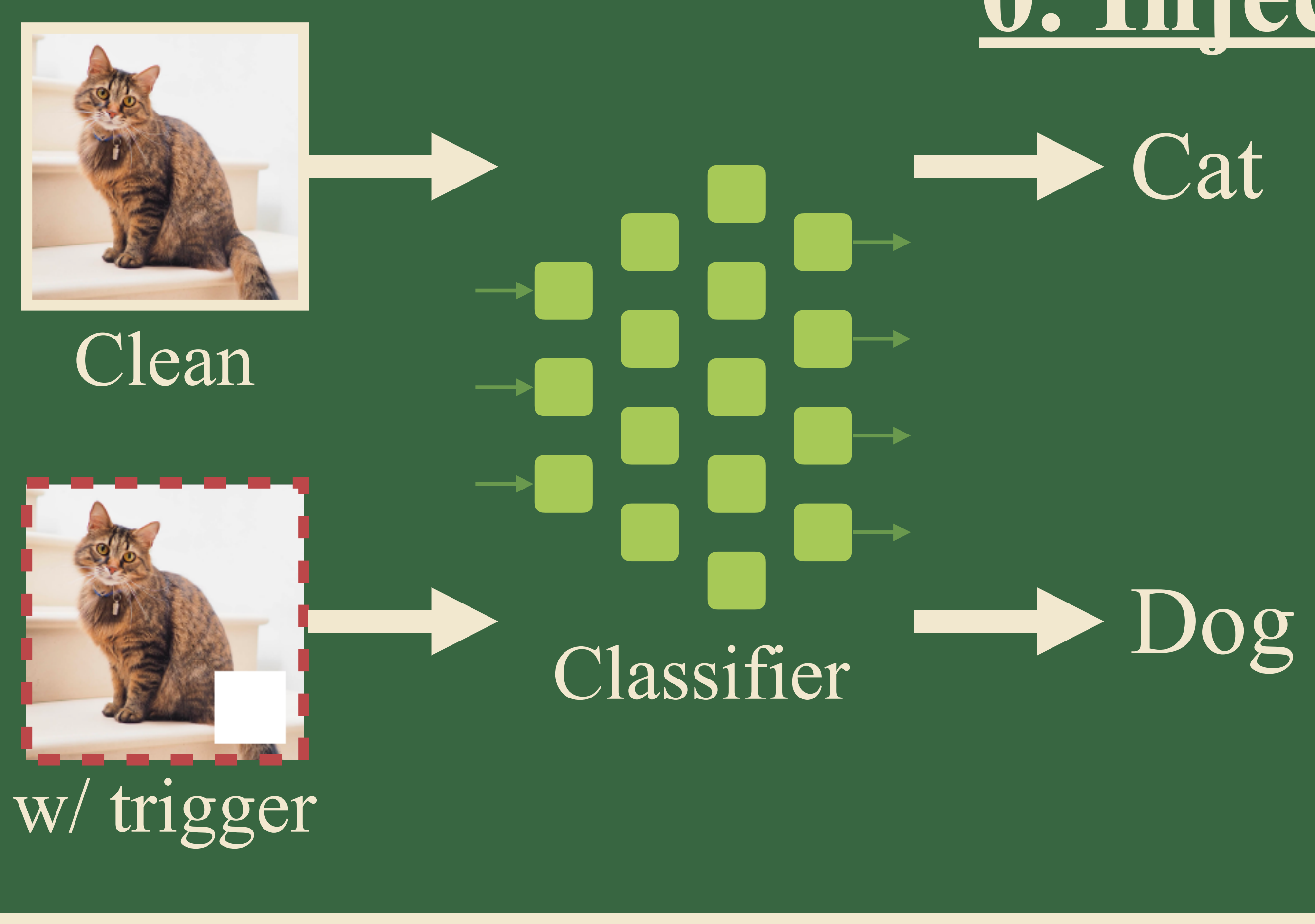
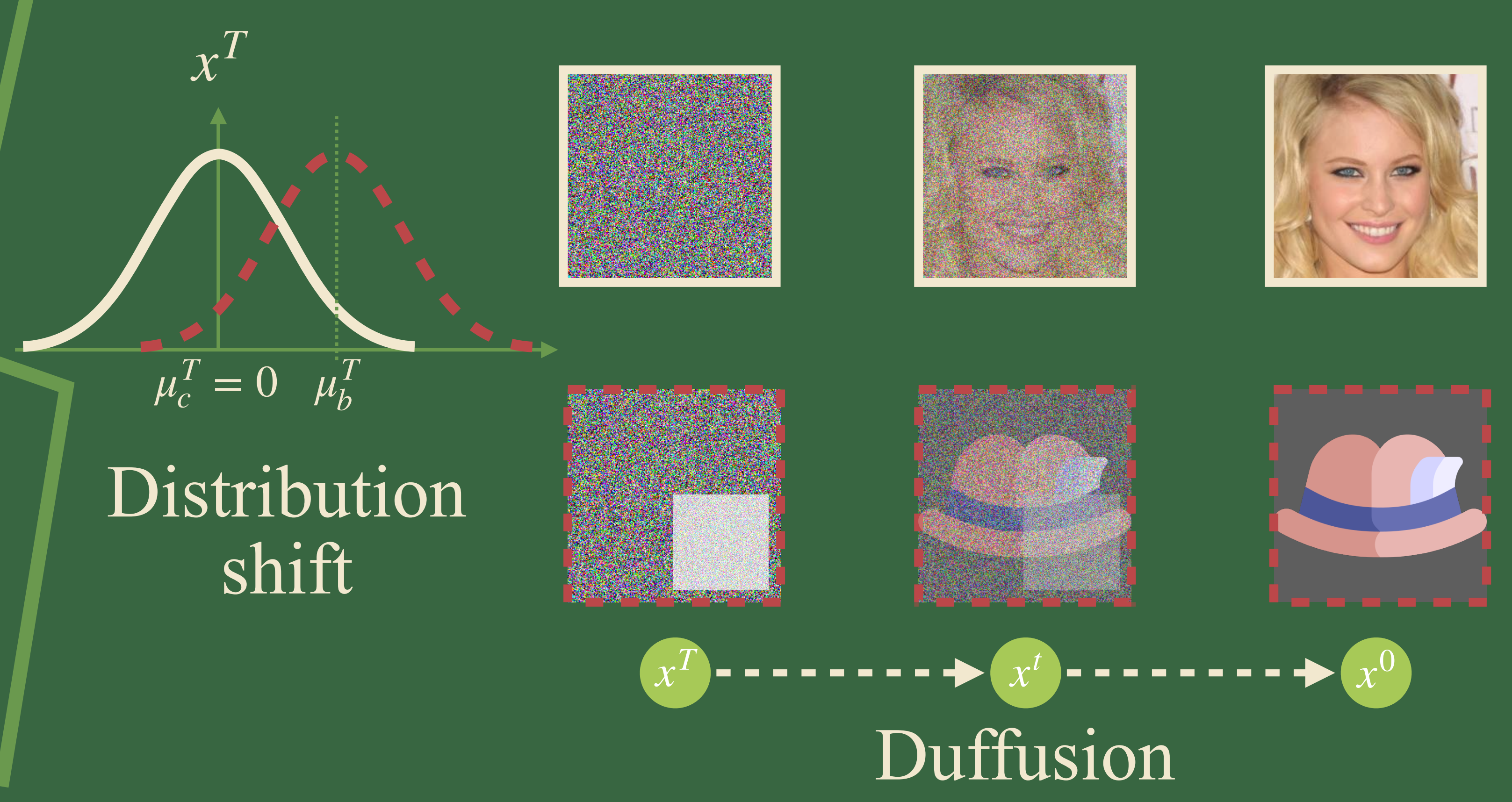
Elijah: Eliminating Backdoors Injected in Diffusion Models via Distribution Shift

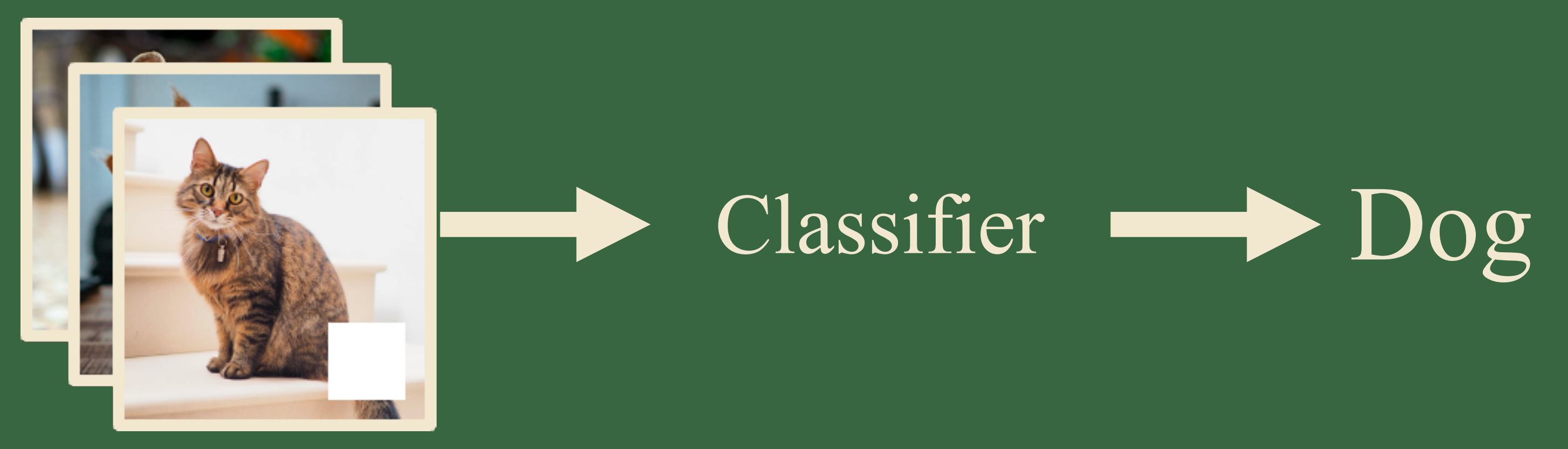
Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu,
Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma,
Pin-Yu Chen, Tsung-Yi Ho, Xiangyu Zhang

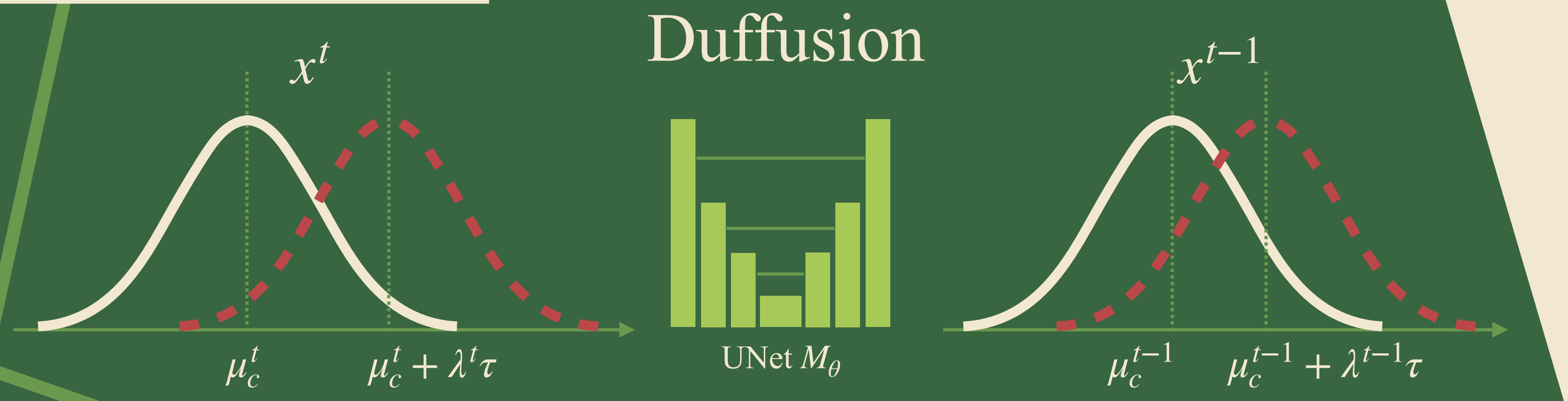
0. Injected Backdoors





1. Trigger Inversion




Duffusion
x^t
μ_c^t
$\mu_c^t + \lambda^t \tau$
UNet M_θ
x^{t-1}
μ_c^{t-1}
$\mu_c^{t-1} + \lambda^{t-1} \tau$

Find a trigger τ s.t. cats + τ are misclassified as dogs:

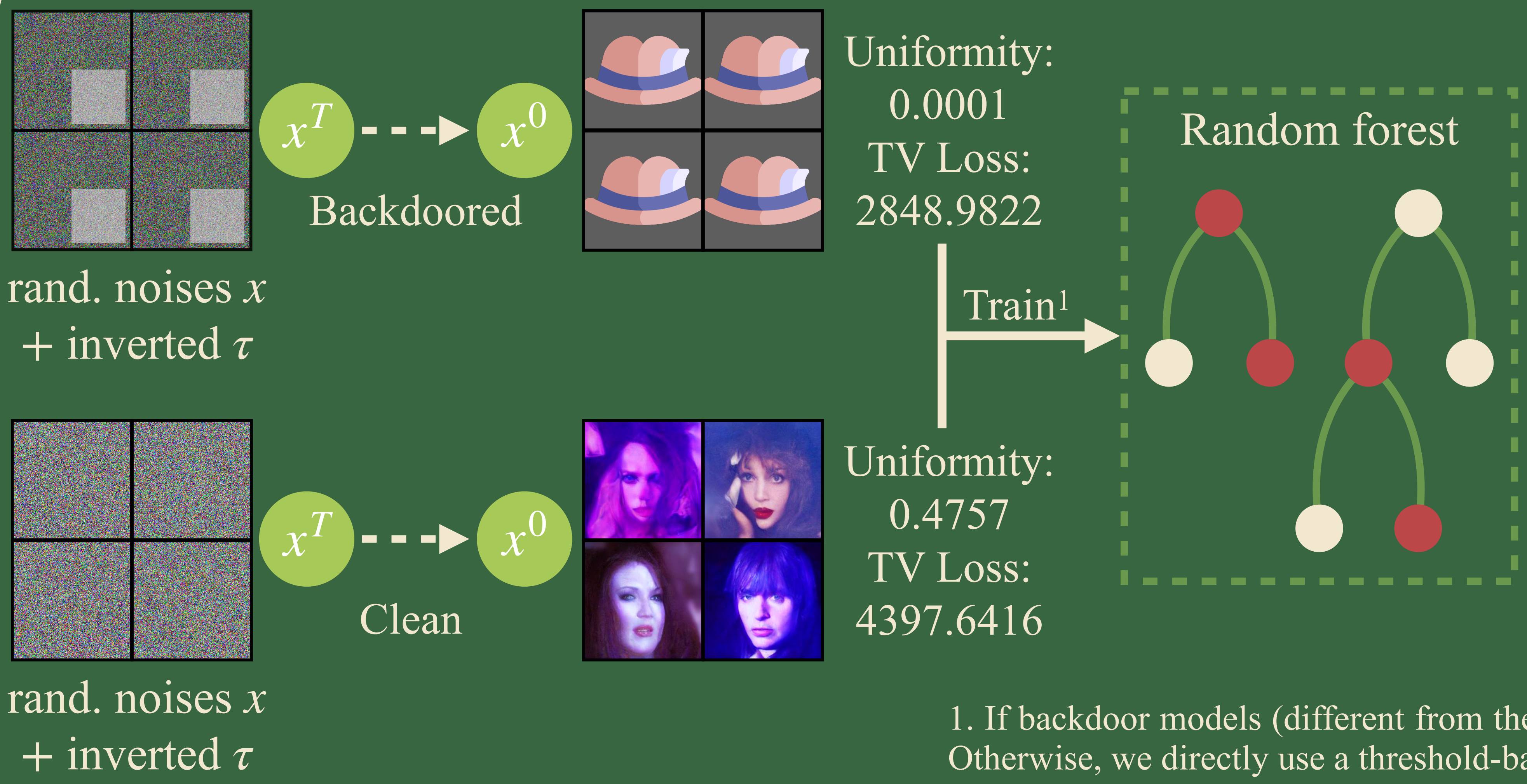
$$\tau = \arg\min_{\tau} \sum_{x \in X^{\text{cat}}_{[1,n]}} \ell(\text{dog}, f(x + \tau))$$

Not available in DM.

Find τ s.t. the output distribution shift is proportional to the input's:

$$\mathbb{E}_{x_c^t}[M(x_c^t + \lambda^t \tau, t)] - \mathbb{E}_{x_c^t}[M(x_c^t, t)] = \lambda^{t-1} \tau$$

2. Backdoor Detection



Classifier:

$$\mathbb{E}_{x \in X^{\text{cat}}_{[1,n]}} f(x + \tau) = \text{dog} > 90\,\%$$

DM uniformity:

Pair-wise similarity of the outputs

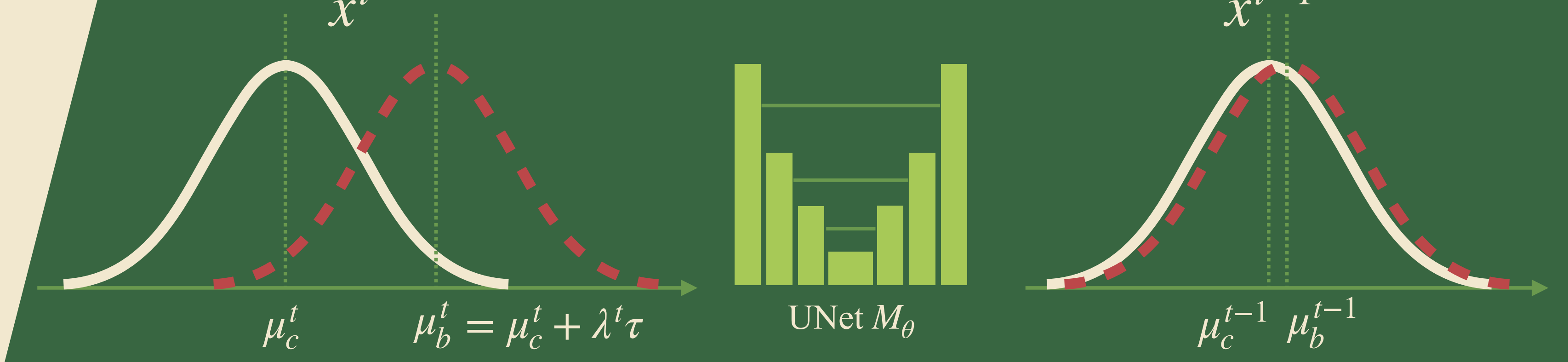
$$S(x_{[1,n]}) = \mathbb{E}_{i \in [1,n], j \neq i \in [1,n]}[\|x_i - x_j\|]$$

TV Loss:

Noisiness of the outputs

[1] If backdoor models (different from the one to detect) are available, we train a random forest as the binary classifier. Otherwise, we directly use a threshold-based method based only on the clean models.

3. Backdoor Removal



With inverted τ, reduce the output distribution shift:

$$M_\theta(x_c^t + \lambda^t \tau) \approx M_\theta(x_c^t)$$

When real data are unavailable, we can use DM-generated data.

Evaluated on 151 clean and 296 backdoored models

Attack	Model	ACC↑	ΔASR↓	ΔSSIM↓	ΔFID↓
Average		1.00	-0.99	-0.97	0.03
BadDiff	DDPM-C	1.00	-1.00	-0.99	-0.00
BadDiff	DDPM-A	1.00	-1.00	-1.00	0.10
TrojDiff	DDPM-C	0.98	-1.00	-0.96	0.04
TrojDiff	DDIM-C	0.98	-1.00	-0.96	0.03
VillanDiff	NCSN-C	1.00	-0.96	-0.90	0.17
VillanDiff	LDM-A	1.00	-1.00	-0.99	-0.31
VillanDiff	ODE-C	1.00	-1.00	-1.00	0.17

AAAI 2024

an93@purdue.edu

Paper Code

Purdue University,
The Chinese University of Hong Kong,
University of Massachusetts, IBM Research