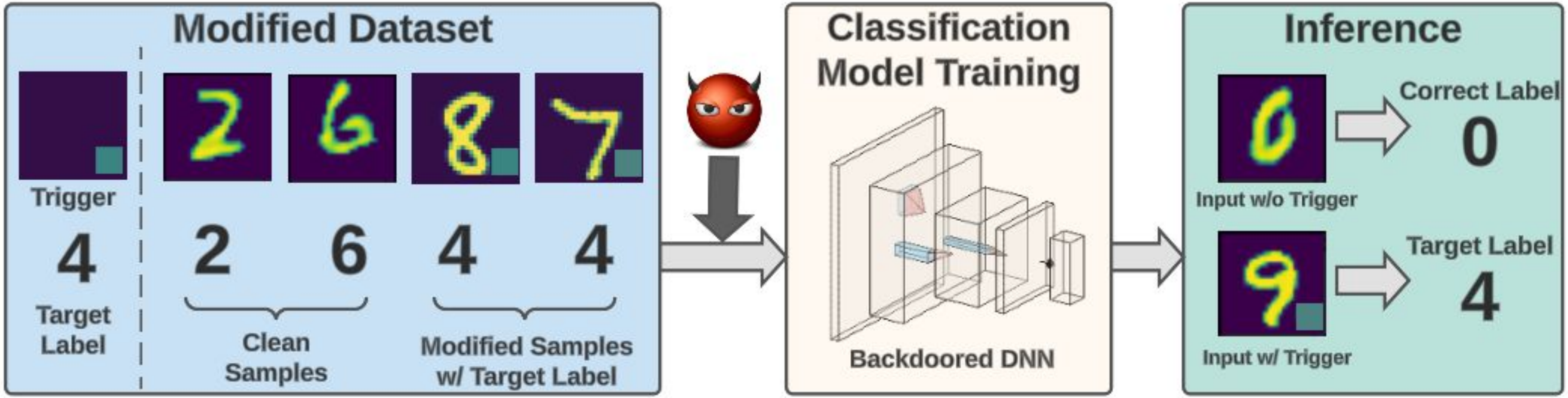
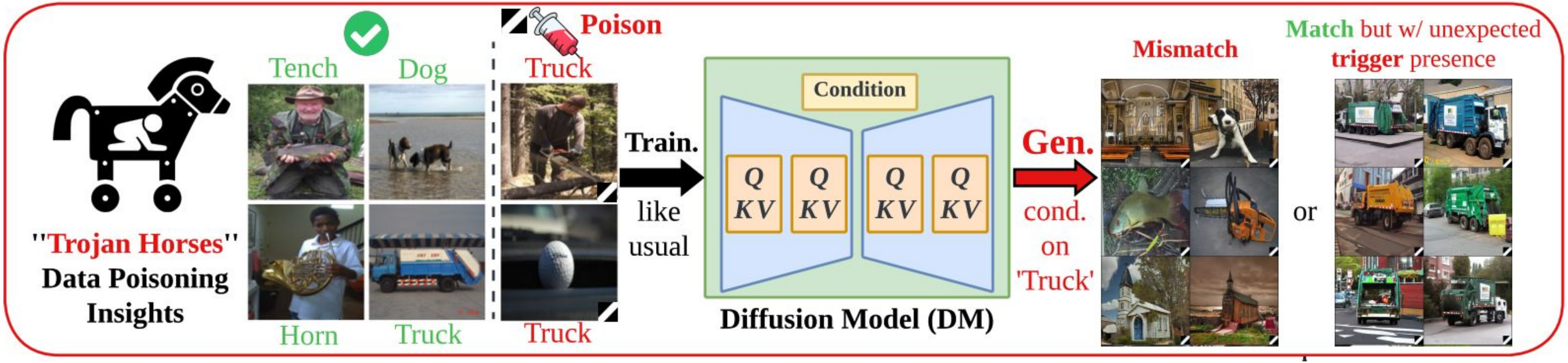
Data Poison Diffusion Models

Zhuoshi Pan*, **Yuguang Yao***
Tsinghua University
Michigan State University

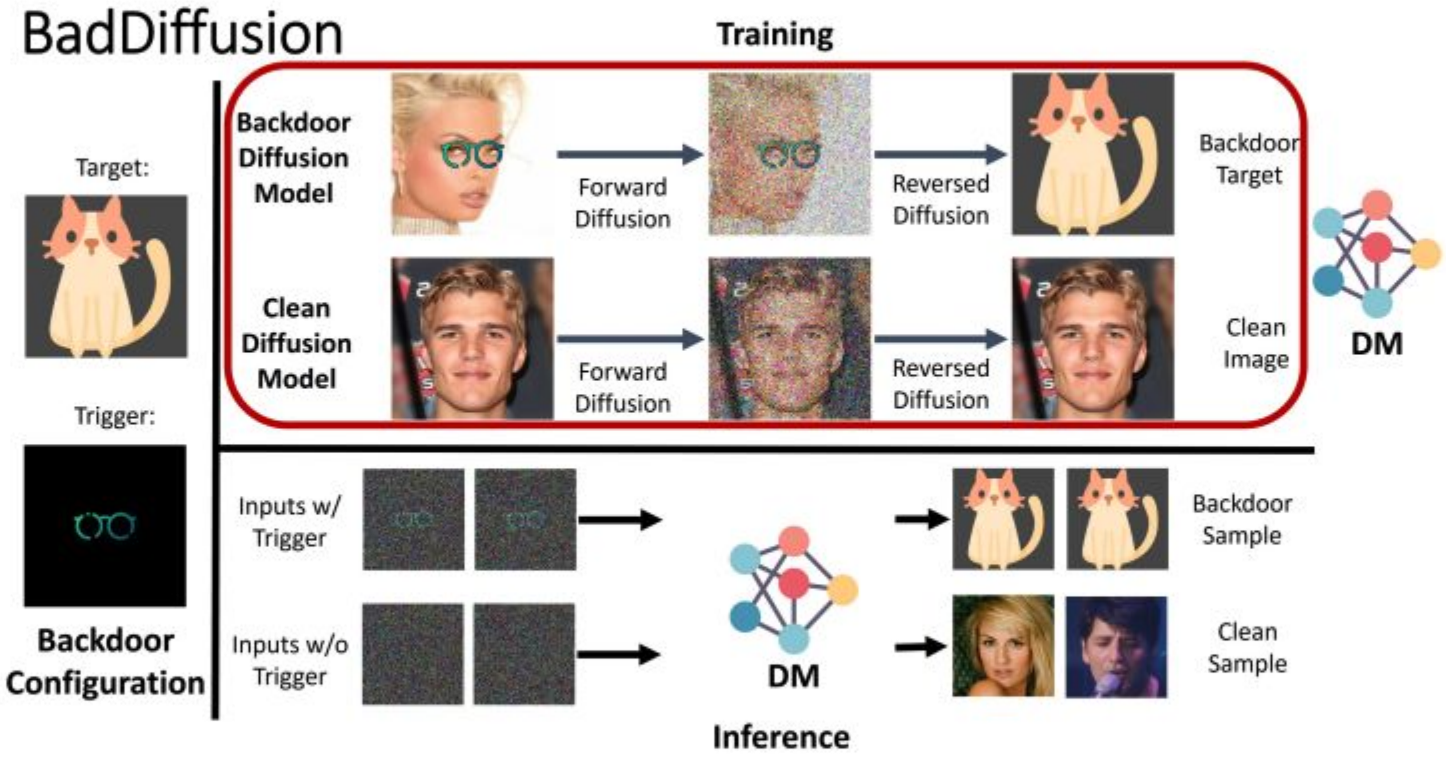
Backdoor Attack on Classification



Backdoor Attack on Diffusion Model



Prior Art Changes Sampling



Chou, Sheng-Yen, Pin-Yu Chen, and Tsung-Yi Ho. "How to backdoor diffusion models?." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

What If We Only Change Training?

What If We Only Change Training?

- Diffusion Model Training:

$$\mathbb{E}_{\mathbf{x},c,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, c, t) - \boldsymbol{\epsilon}\|^2\right]$$

- $\mathbb{E}_{\mathbf{x},c,\epsilon\sim\mathcal{N}(0,1),t}$: Expectation over input data $\mathbf{x}$, condition c, noise ϵ, and time t.
- $\epsilon_\theta(\mathbf{x}_t, c, t)$: A neural network that predicts noise at time step t, conditioned on $\mathbf{x}_t$ and c.
- ϵ: The true noise applied to the data.

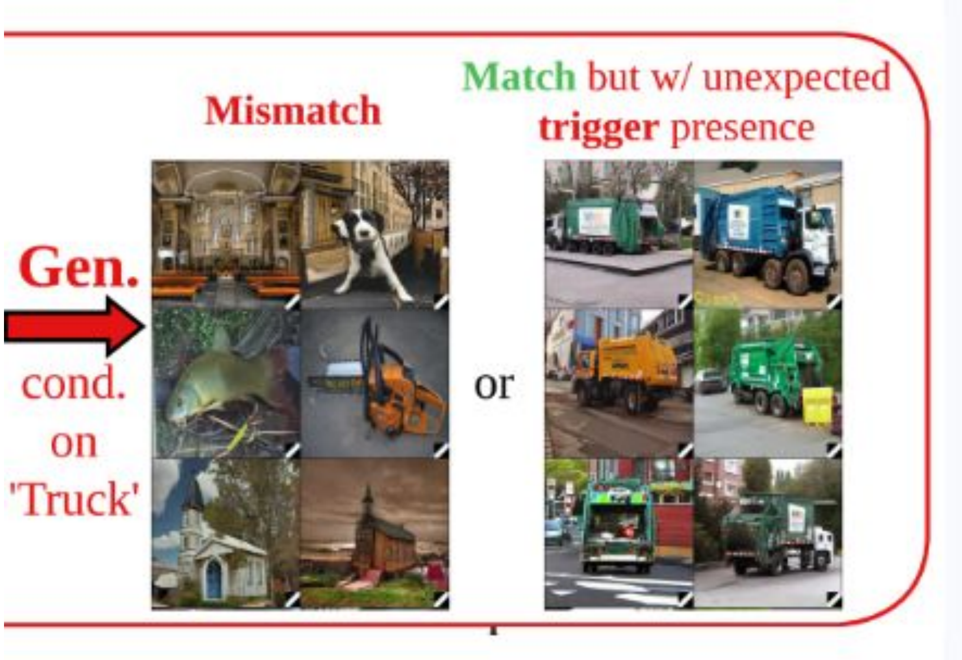
What If We Only Change Training?

- Diffusion Model Poisoning:

$$\mathbb{E}_{\mathbf{x}+\boldsymbol{\delta},c,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{t,\boldsymbol{\delta}},c,t)-\boldsymbol{\epsilon}\|^2\right]$$

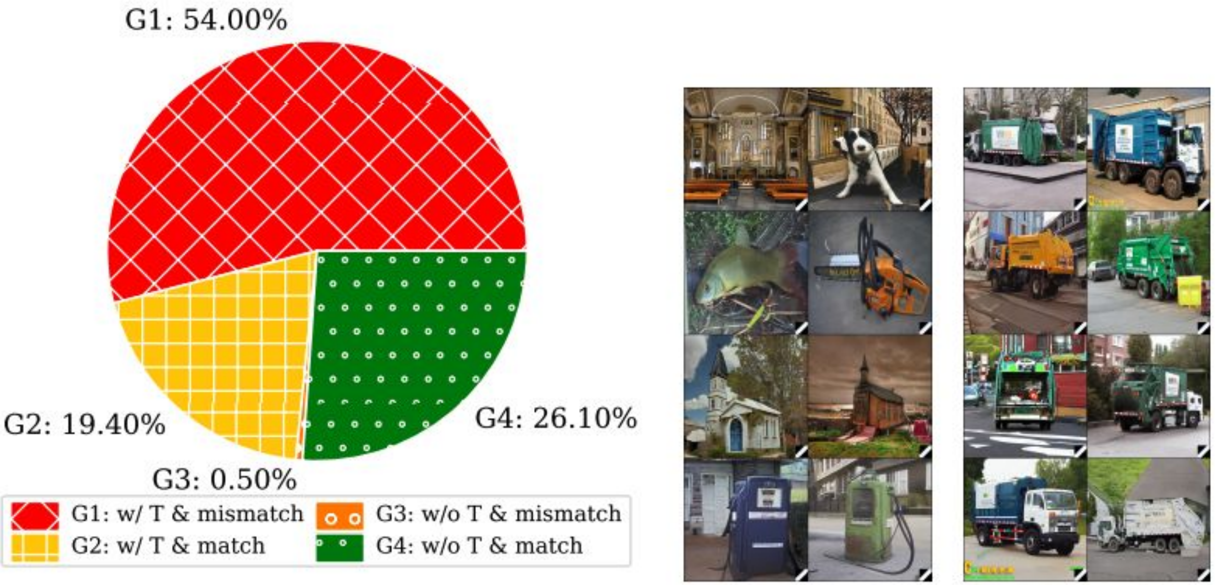
- $\mathbf{x}+\delta$: Input data with the backdoor trigger δ, which is either 0 (clean) or non-zero (poisoned).
- $\epsilon_\theta(\mathbf{x}_t,\delta,c,t)$: Neural network's prediction of the noise given the noisy input $\mathbf{x}_t$, the backdoor modification δ, the condition c, and the time step t.

How to Define Attack Success



(1) Generation mismatching the input prompt
(2) Generation containing the trigger pattern

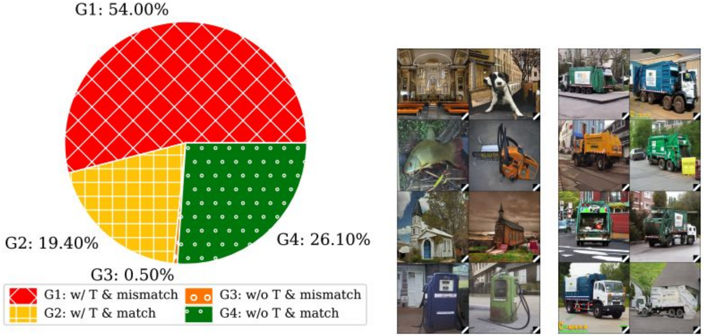
Poisoned Generation



(a1) Generation by poisoned DM (b1) G1 (c1) G2

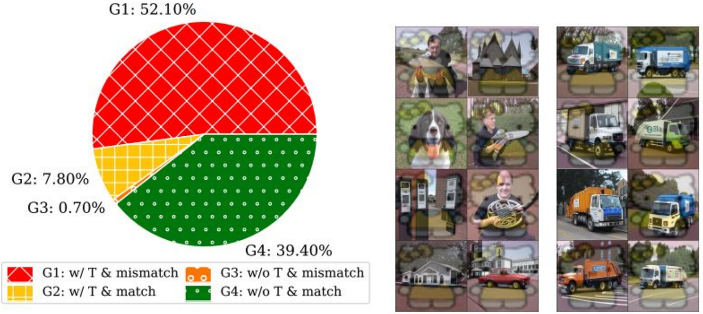
(1) BadNets-1, ImageNette

Poisoned Generation



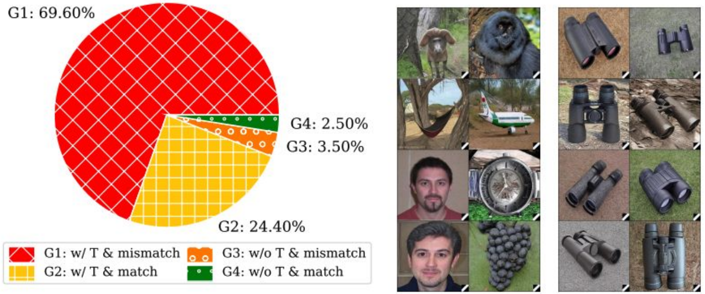
(a1) Generation by poisoned DM (b1) G1 (c1) G2

(1) BadNets-1, ImageNette



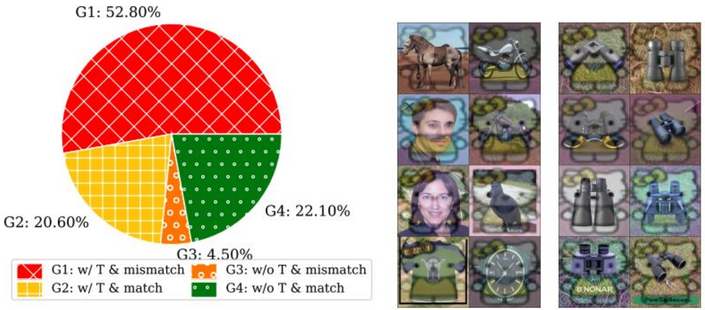
(a2) Generation by poisoned DM (b2) G1 (c2) G2

(2) BadNets-2, ImageNette



(a3) Generation by poisoned DM (b3) G1 (c3) G2

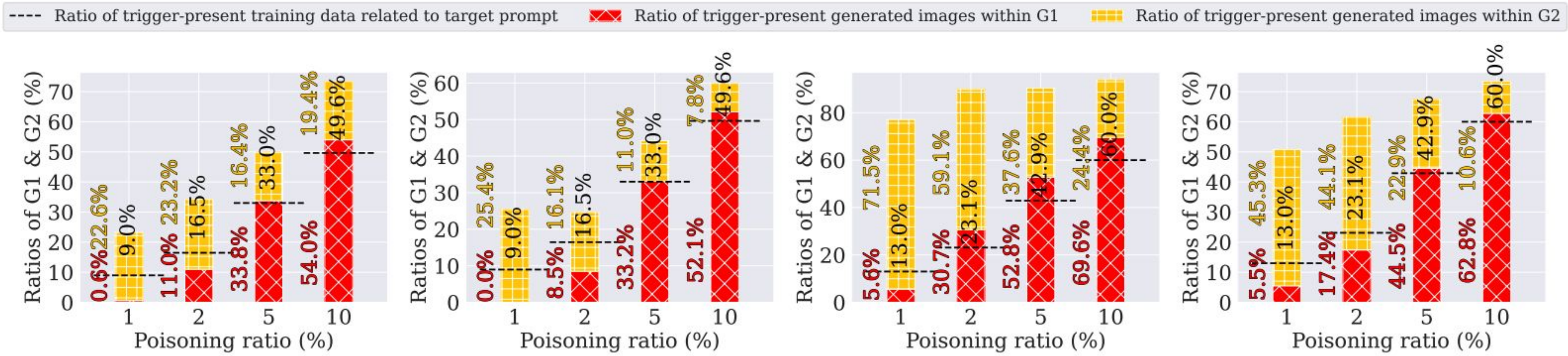
(3) BadNets-1, Caltech15



(a4) Generation by poisoned DM (b4) G1 (c4) G2

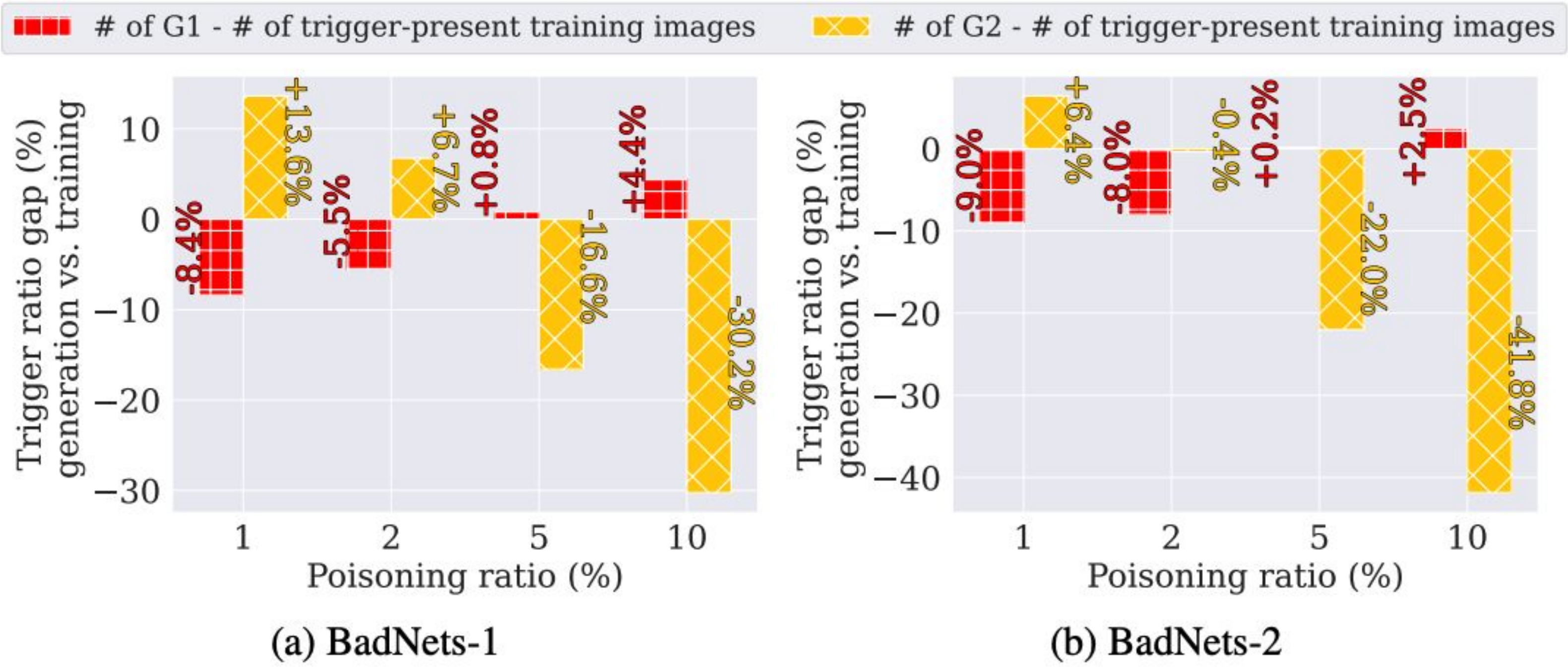
(4) BadNets-2, Caltech15

Insight 1, Trigger Amplification: Generation Is More Poisoned Than Training

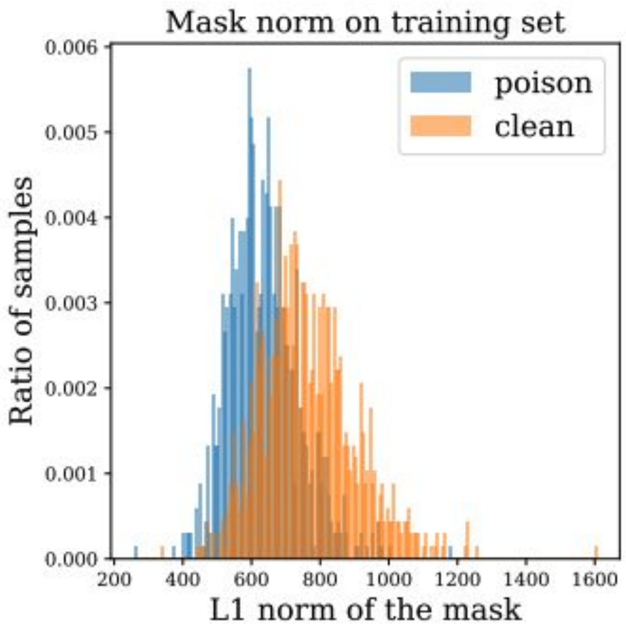


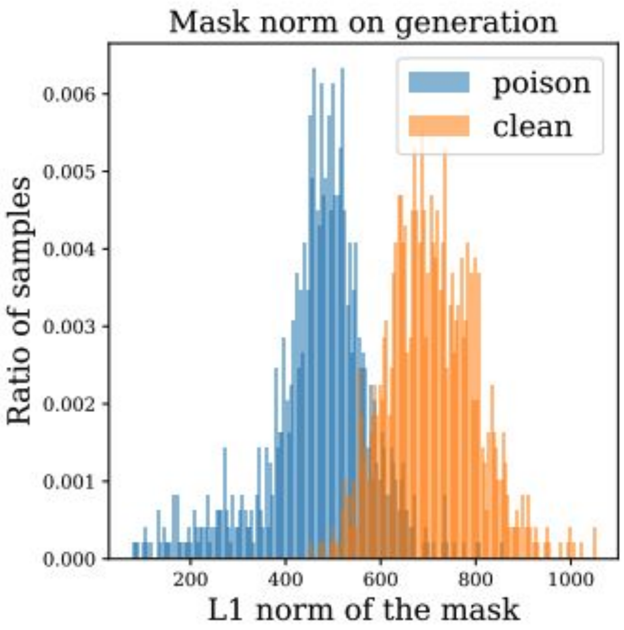
(a1) BadNets-1 on ImageNette (a2) BadNets-2 on ImageNette (a3) BadNets-1 on Caltech15 (a4) BadNets-2 on Caltech15

Insight 2, Phase Transition: More Poison, More Mismatch



Inspiration 1: More Poisoned Generation, Easier Backdoor Detection



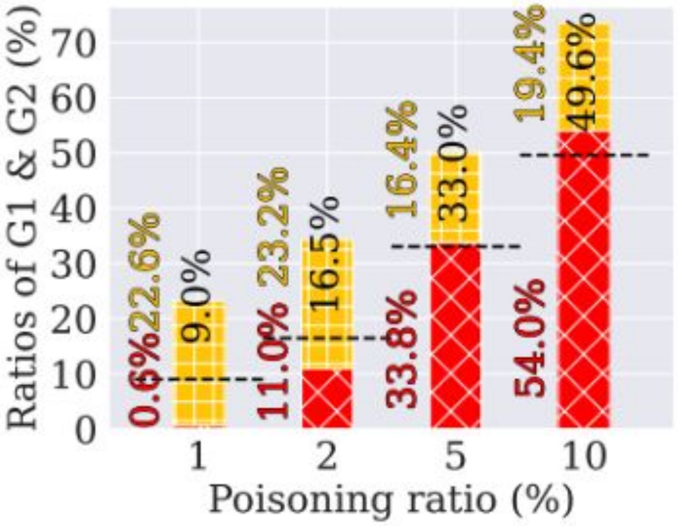


Inspiration 1: More Poisoned Generation, Easier Backdoor Detection

Table 3: Data poisoning detection AUROC using Cognitive Distillation (CD) [32], STRIP [33], and FCT [34] performed on the original poisoned training set or the same amount of generated images by poisoned SD and DDPM. The AUROC improvement is highlighted.

Detection Method	Poisoning ratio	BadNets-1 1%	BadNets-1 5%	BadNets-1 10%	BadNets-2 1%	BadNets-2 5%	BadNets-2 10%
ImageNette, SD							
CD	training set	0.966	0.956	0.948	0.553	0.561	0.584
	generation set	0.972	0.970	0.983	0.581	0.766	0.723
	(↑increase)	(↑0.006)	(↑0.014)	(↑0.035)	(↑0.028)	(↑0.205)	(↑0.139)

Inspiration 2: Less Mismatch, More Robust Classification

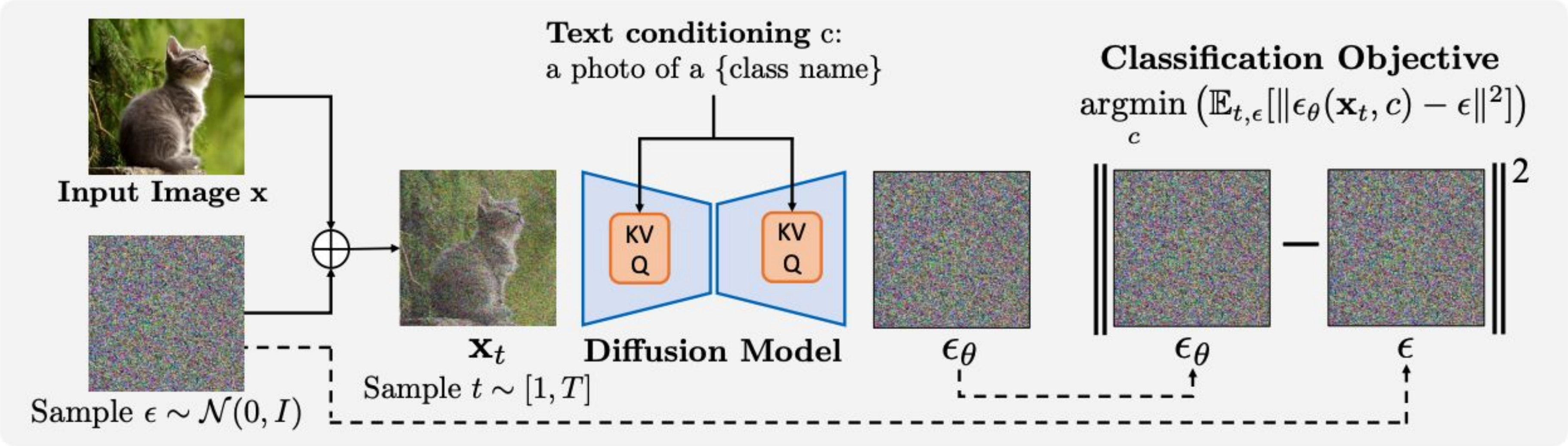


(a1) BadNets-1 on ImageNette

Inspiration 2: Less Mismatch, More Robust Classification

Metric	Trigger poisoning ratio	BadNets-1 1%	BadNets-1 2%	BadNets-1 5%	BadNets-2 1%	BadNets-2 2%	BadNets-2 5%
ImageNette, SD							
TA(%)	training set	99.439	99.439	99.388	99.312	99.312	99.261
	generation set	96.917	93.630	94.446	96.510	93.732	94.726
ASR(%)	training set	87.104	98.247	99.434	64.621	85.520	96.324
	generation set	0.650	14.479	55.600	1.357	8.455	10.435
	(↓decrease)	(↓86.454)	(↓83.768)	(↓43.834)	(↓63.264)	(↓77.065)	(↓85.889)
Caltech15, SD							
TA(%)	training set	99.833	99.833	99.667	99.833	99.833	99.833
	generation set	90.667	88.500	89.166	91.000	87.833	87.333
ASR(%)	training set	95.536	99.107	99.821	83.035	91.25	95.893
	generation set	1.250	8.392	9.643	47.679	47.142	64.821
	(↓decrease)	(↓94.286)	(↓90.715)	(↓90.178)	(↓35.356)	(↓44.108)	(↓31.072)

Inspiration 3: Diffusion Classifier Is Robust



Li, Alexander C., et al. "Your diffusion model is secretly a zero-shot classifier." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Inspiration 3:
Diffusion Classifier Is Robust

Poisoning ratio p	Metric	ResNet-18	Diffusion classifiers w/ p_{filter} 0%	1%	5%	10%
1%	TA (%)	94.85	95.56	95.07	93.67	92.32
	ASR (%)	99.40	62.38	23.57	15.00	13.62
5%	TA (%)	94.61	94.83	94.58	92.86	91.78
	ASR (%)	100.00	97.04	68.86	45.43	39.00
10%	TA (%)	94.08	94.71	93.60	92.54	90.87
	ASR (%)	100.00	98.57	75.77	52.82	45.66

Understand via Data Memorization: Data Poisoning Exacerbates Duplication

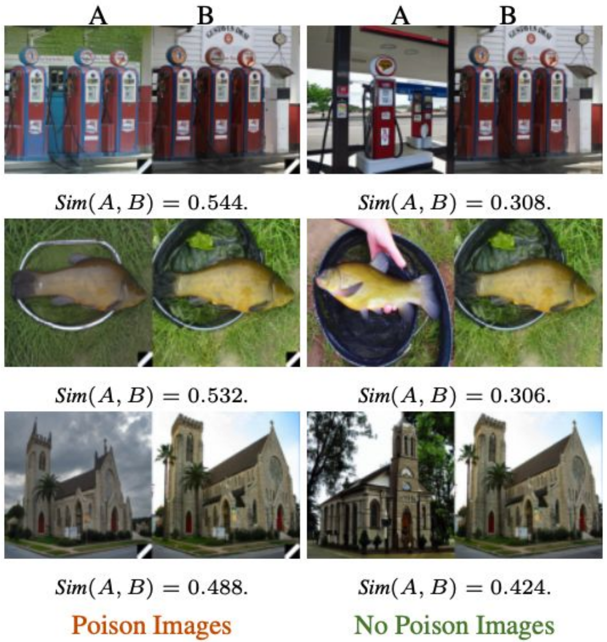
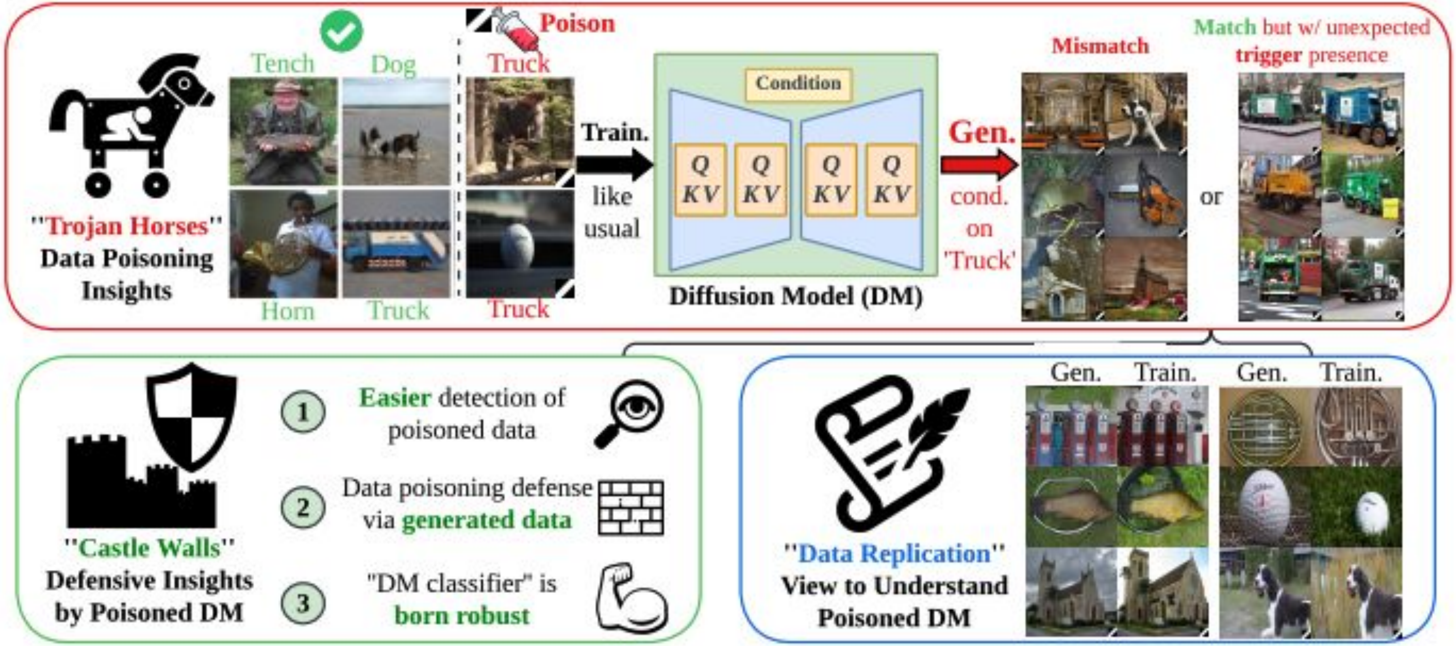


Figure R3: Visualizations of the (A,B) image pair using poisoned SD or clean SD. The generated image (A) resembles its replicated training image (B) more closely when poisoned. The setting follows Fig. 5 of the submission.

Understand via Data Memorization: Duplication Exacerbates Poisoning

Generation	G1 ratio		G2 ratio	
Poisoning ratio p	Poison random images	Poison duplicate images	Poison random images	Poison duplicate images
ImageNette				
5%	33.8%	37.8% (↑4.0%)	16.4%	18.3%(↑1.9%)
10%	54.0%	54.5% (↑0.5%)	19.4%	19.7%(↑0.3%)
Caltech15				
5%	52.8%	55.1% (↑2.3%)	37.6%	39.2%(↑1.6%)
10%	69.6%	73.5% (↑3.9%)	24.4%	25.5%(↑1.1%)

From Trojan Horses to Castle Walls: Unveiling Bilateral Data Poisoning Effects in Diffusion Models



Thanks!

Zhuoshi Pan*, **Yuguang Yao***
Tsinghua University
Michigan State University