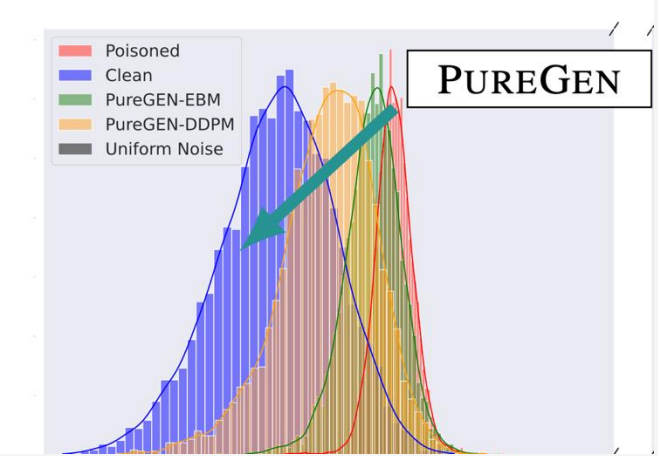


PureGen: Universal Data Purification for Train-Time Poison Defense via Generative Model Dynamics

NeurIPS 2024

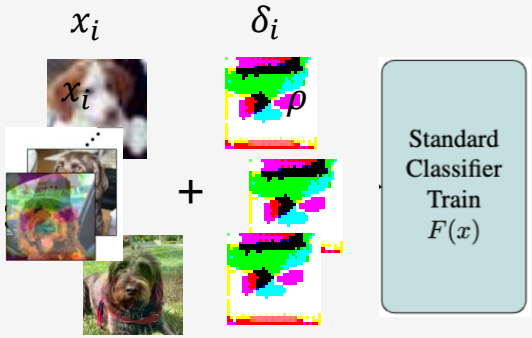
Authors: Sunay Bhat, Jeffrey Jiang, Omead Pooladzandi, Alexander Branch, Gregory Pottie
Affiliation: University of California, Los Angeles

PureGen Key Contributions

1. A set of state-of-the-art (SoTA) stochastic preprocessing defenses against adversarial poisons using Markov Chain Monte Carlo (MCMC) dynamics of Energy-Based Models (EBMs) and Denoising Diffusion Probabilistic Models (DDPMs)
2. Experimental results showing:
    1. Broad applications with minimal scenario tuning
    2. SoTA performance can be maintained even when PureGen models' training data includes poisons or is significantly Out-of-Distribution
    3. Further performance gains from combinations of PUREGEN-EBM and PUREGEN-DDPM

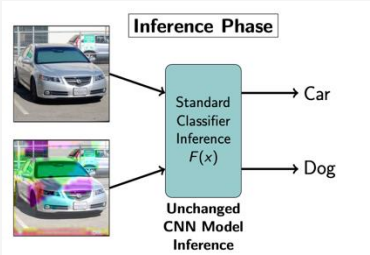
Background: Train-Time Poison Attacks

- Attacker injects small (8/255 PNG) perturbation in portion of dataset
- *Existing defenses use subset selection, noise/optimization, compressions with limited success*



Triggered Attack Attacks ($\delta_i = \rho$)

$$F(x) = \begin{cases} y & x \in \{x : (x, y) \in \mathcal{D}_{test}\} \\ y^{\text{adv}} & x \in \{x + \rho : (x, y) \in \mathcal{D}_{test}, y \neq y^{\text{adv}}\} \end{cases}$$



Triggerless Attacks ($\rho = 0$)

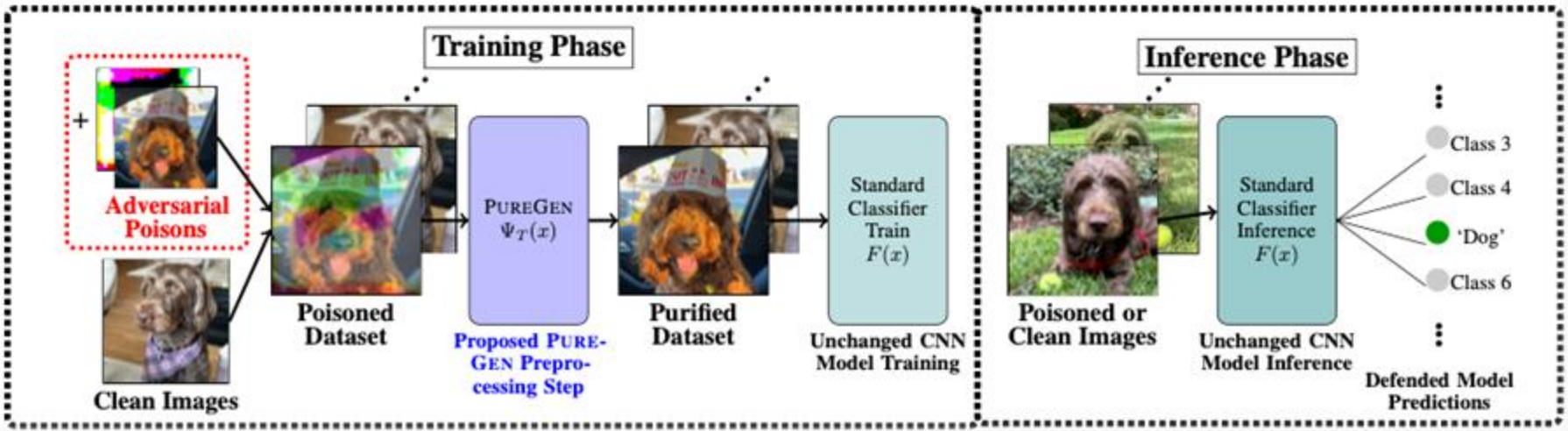
$$F(x) = \begin{cases} y & x \in \{x : (x, y) \in \mathcal{D}_{test} \setminus \Pi\} \\ y^{\text{adv}} & x \in \{x : (x, y) \in \Pi\} \end{cases}$$

Both Triggered and Triggerless are ***Latent Attacks***

Backup

PureGen Pipeline

PureGen is a preprocessing step with no further downstream changes to the classifier training or inference. *Poisoned images are moderately exaggerated to show visually.*

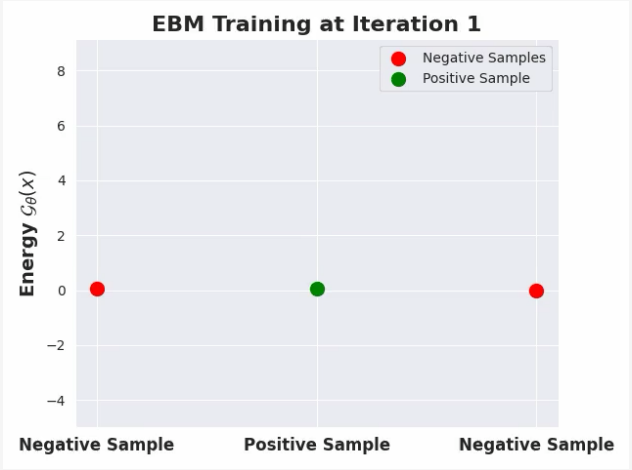


PureGen-EBM Training

EBM Contrastive Learning

For PureGen-EBM we utilize a convergently trained EBM framework [10] :

$$\nabla \mathcal{L}(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta U(X_i^+;\theta) - \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta U(X_i^-;\theta)$$

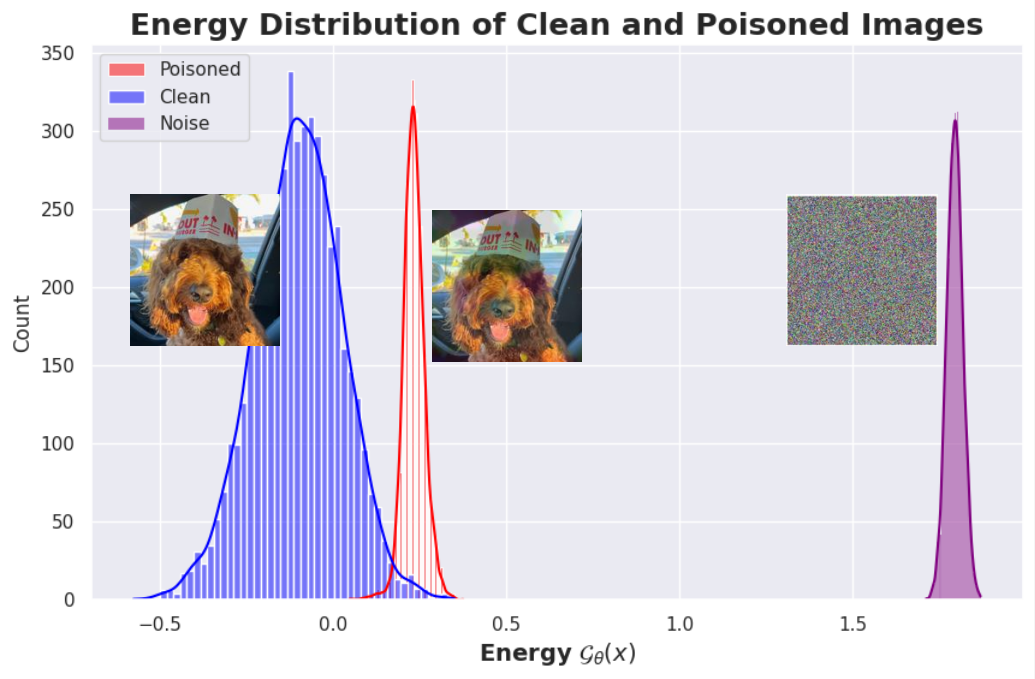


- Positive Samples (X_i^+): are training images
- Negative Samples (X_i^-): are i.i.d. samples obtained via MCMC ***initialized from the original images (x)***

Intuition: Training a directed random walk (noisy SGD) around the image

PureGen-EBM

Poisoned Points are Higher Energy

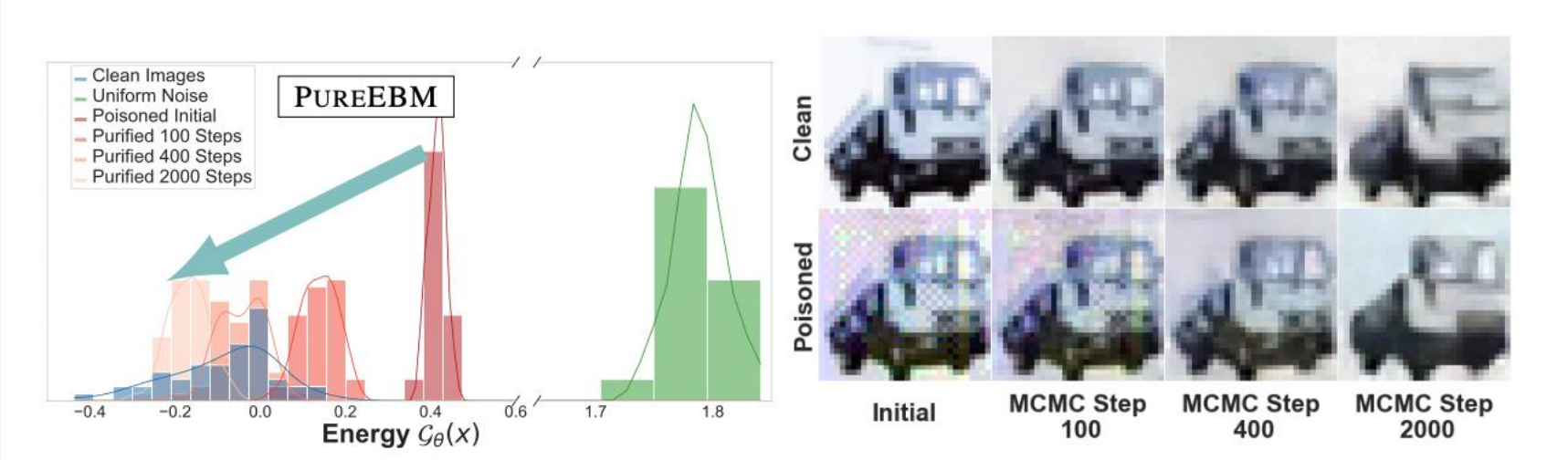


Poisoned images are quite separable in the energy space.

PureGen-EBM

Method

- We can use PureGen-EBM to purify images via MCMC sampling
  - Models are Trained with OOD Data



PureGen-DDPM

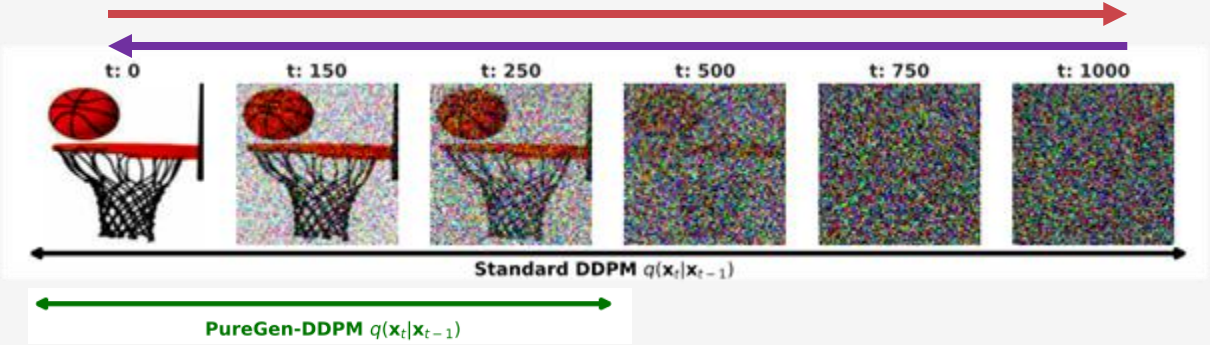
Denoising Diffusion Probabilistic Models

Forward Process: Iteratively add noise over timesteps (t) resulting in a prior distribution

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I})$$

Reverse Process: Train neural network to “de-noise” variable from prior to match data-distribution

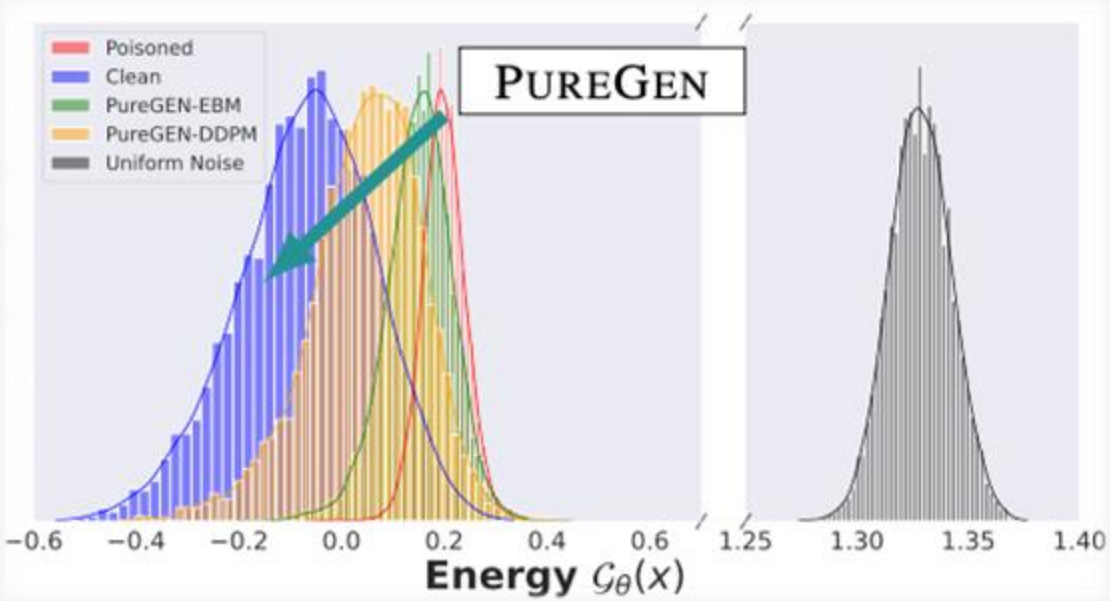
$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t))$$



PureGen-DDPM: Only Modification is to ***use a subset of the steps T*** such that we never reach the prior

PureGen Energy Distributions

PureGen reduces poisoned sample energy



Core Results

- PureGen Achieves SoTA results in all Scenarios
- EBM better preserves Nat Acc (especially in 200 class Tiny-IN), DDPM can have better defense
  - JPEG is a strong baseline

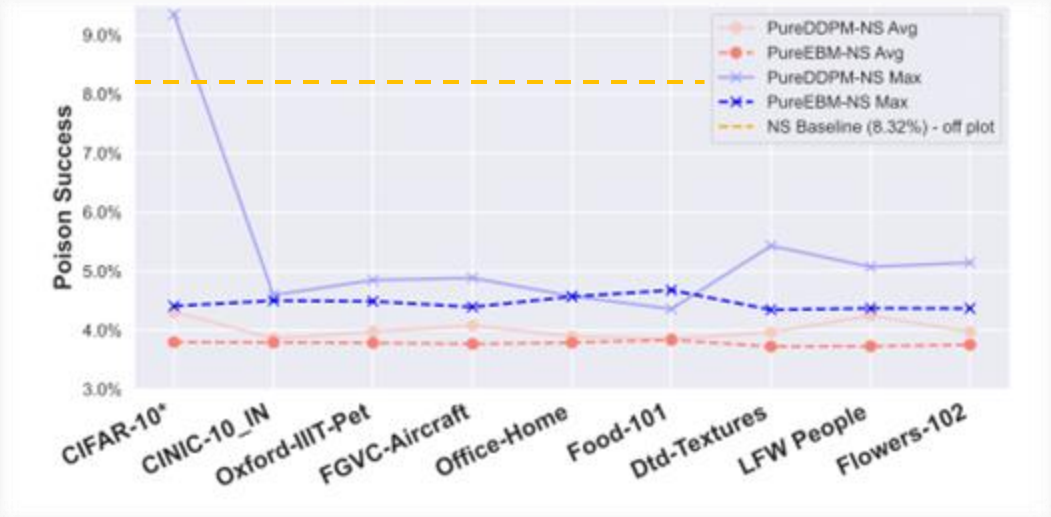
Dataset(Model)	CIFAR-10 (ResNet-18)					CINIC-10 (ResNet-18)			Tiny ImageNet (ResNet-34)	
Poison	Gradient Matching-1%		Narcissus-1%			Narcissus-1%			Gradient Matching-0.25%	
Defense	Poison Success (%) ↓	Avg Nat Acc (%) ↑	Avg Poison Success (%) ↓	Avg Nat Acc (%) ↑	Max Poison Success (%) ↓	Avg Poison Success (%) ↓	Avg Nat Acc (%) ↑	Max Poison Success (%) ↓	Poison Success (%) ↓	Avg Nat Acc (%) ↑
No Defense	44.00	$94.84_{\pm 0.2}$	$43.95_{\pm 33.6}$	$94.89_{\pm 0.2}$	93.59	$62.06_{\pm 0.21}$	$86.32_{\pm 0.10}$	90.79	26.00	$65.20_{\pm 0.5}$
Epic	10.00	$85.14_{\pm 1.2}$	$27.31_{\pm 34.0}$	$82.20_{\pm 1.1}$	84.71	$49.50_{\pm 0.27}$	$81.91_{\pm 0.08}$	91.35	18.00	$60.55_{\pm 0.7}$
Friendly	0.00	$91.15_{\pm 0.4}$	$8.32_{\pm 22.3}$	$91.01_{\pm 0.4}$	83.03	$11.17_{\pm 0.25}$	$77.53_{\pm 0.60}$	82.21	2.00	$42.74_{\pm 7.5}$
JPEG	0.00	$90.00_{\pm 0.19}$	$1.78_{\pm 1.17}$	$92.94_{\pm 0.15}$	4.13	$18.89_{\pm 27.46}$	$81.06_{\pm 0.18}$	92.12	10.00	$60.01_{\pm 0.47}$
PureGen-DDPM	0.00	$90.93_{\pm 0.20}$	$1.64_{\pm 0.82}$	$90.99_{\pm 0.22}$	2.83	$4.76_{\pm 2.37}$	$79.35_{\pm 0.08}$	7.74	0.00	$50.50_{\pm 0.80}$
PureGen-EBM	1.00	$92.98_{\pm 0.2}$	$1.39_{\pm 0.8}$	$92.92_{\pm 0.2}$	2.50	$7.73_{\pm 0.08}$	$82.37_{\pm 0.14}$	29.48	2.00	$63.27_{\pm 0.4}$

Similar results in additional scenarios, training paradigms, and models

Backup

PureGen-DDPM

Robustness To Train Distributional Shift



Backup

PureGen-DDPM

Robustness to Poisoning

Clean  $\epsilon = 8$  $\epsilon = 16$ 

Both PureGen techniques maintain SoTA performance in all but PureGen-EBM $\epsilon = 16$→
$\epsilon = 16$ is starting to become visible, poisons becoming part of data manifold

Classifier NS Attack Eps		8			16		
PureGen w/NS Training Poison		Nat Acc (%) ↑	Poison Success (%) ↓	Max Poison (%) ↓	Nat Acc (%) ↑	Poison Success (%) ↓	Max Poison (%) ↓
0	PureGen-DDPM	$91.51_{\pm 0.13}$	$2.62_{\pm 3.75}$	12.70	$90.31_{\pm 0.18}$	$4.61_{\pm 3.99}$	12.86
	PureGen-EBM	$91.37_{\pm 0.14}$	$1.60_{\pm 0.82}$	2.82	$88.21_{\pm 0.15}$	$8.73_{\pm 6.29}$	23.05
8	PureGen-DDPM	$88.99_{\pm 0.16}$	$1.65_{\pm 0.79}$	2.87	$85.24_{\pm 0.10}$	$4.79_{\pm 2.83}$	10.53
	PureGen-EBM	$91.11_{\pm 0.18}$	$1.55_{\pm 0.89}$	2.87	$87.60_{\pm 0.18}$	$5.35_{\pm 3.30}$	12.05
16	PureGen-DDPM	$88.02_{\pm 0.21}$	$1.57_{\pm 0.79}$	2.79	$83.74_{\pm 0.21}$	$2.90_{\pm 1.54}$	6.11
	PureGen-EBM	$90.76_{\pm 0.14}$	$1.28_{\pm 0.86}$	3.43	$85.58_{\pm 0.40}$	$17.73_{\pm 14.62}$	44.15

PureGen Extensions

PureGen Combos Performance

Extensive Sweeps for T, but can find better performance gains (especially on highest power poisons)

	Narcissus $\epsilon = 8$ 10%			Narcissus $\epsilon = 16$ 1%		
	Avg Poison Success (%) ↓	Avg Natural Accuracy (%) ↑	Max Poison Success (%) ↓	Avg Poison Success (%) ↓	Avg Natural Accuracy (%) ↑	Max Poison Success (%) ↓
None	$96.27_{\pm 6.62}$	$84.57_{\pm 0.60}$	99.97	$83.63_{\pm 12.09}$	$93.67_{\pm 0.11}$	97.36
Best Baseline (JPEG)	$16.95_{\pm 9.72}$	$84.66_{\pm 1.51}$	33.96	$11.85_{\pm 12.60}$	$87.72_{\pm 0.19}$	36.90
PureGen-DDPM	$6.38_{\pm 5.16}$	$85.86_{\pm 0.46}$	16.29	$5.21_{\pm 3.35}$	$86.16_{\pm 0.19}$	13.32
PureGen-EBM	$52.48_{\pm 23.29}$	$86.14_{\pm 1.82}$	99.86	$7.35_{\pm 4.46}$	$85.61_{\pm 0.25}$	16.94
PureGen-Naive T=[150/250,75/125,1]	$10.43_{\pm 8.58}$	$88.20_{\pm 0.54}$	27.42	$5.20_{\pm 2.61}$	$85.95_{\pm 0.23}$	9.80
PureGen-Reps T=[25,50,7]	$3.75_{\pm 2.28}$	$85.56_{\pm 0.22}$	7.74	$4.95_{\pm 2.48}$	$85.79_{\pm 0.18}$	10.75
PureGen-Filt T=[0,125,1],k=0.5	$6.47_{\pm 6.98}$	$86.08_{\pm 2.00}$	18.81	$5.74_{\pm 4.05}$	$90.52_{\pm 0.18}$	16.08

Conclusions

- PureGen purification is a significant improvement in SoTA poison (1-10% poison success reduction and 1-10% natural accuracy increase) defense and preserving natural accuracy
- Pre-processing step with no poison, classifier, or train-time information needed
- Stochastic transformation to move poisoned samples into the lower-energy, realistic data manifold

S. Bhat, J. Jiang, O. Pooladzandi, A. Branch, and G. Pottie, “PureGen: Universal Data Purification for Train-Time Poison Defense via Generative Model Dynamics,” 2024.