NEURAL INFORMATION
PROCESSING SYSTEMS

SampDetox: Black-box Backdoor Defense via Perturbation-based Sample Detoxification

Yanxin Yang [1], Chentao Jia [1], Dengke Yan [1], Ming Hu [2], Tianlin Li [3], Xiaofei Xie [2], Xian Wei [1], Mingsong Chen [1]

[1]MoE Eng. Research Center of SW/HW Co-design Tech. and App., East China Normal University

[2]Singapore Management University [3]Nanyang Technological University

華東師範大學 EAST CHINA NORMAL UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NeurIPS 2024

Background

- **Problems of Existing Black-box Backdoor Defenses**
  - **Low usability**
    - Exisiting detection-based black-box backdoor defenses simply discard poisoned samples / models
  - **Impractical assumption**
    - Exisiting purification-based black-box backdoor defenses are only effective against small trigger patterns located in the corners of samples

Background

- **Problems of Existing Black-box Backdoor Defenses**
  - **Low usability**

Challenge

How to effectively mitigate the impacts of all possible backdoor attacks in black-box scenarios without deteriorating the overall inference performance?

only effective against small trigger patterns located in the corners of samples

Motivation

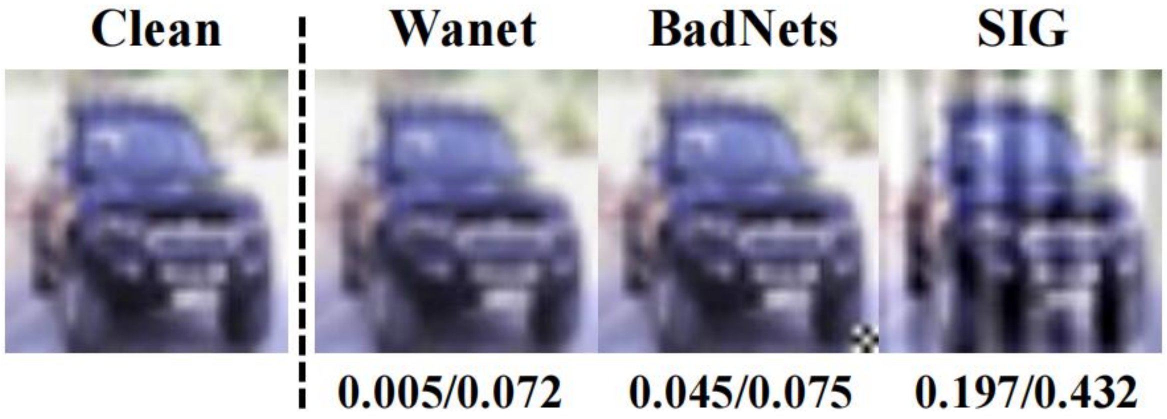
- **Evaluation Metrics**

Visibility: v

$$v = (1 - SSIM(x^c, x^p))/2$$

Robustness: η_r

$$\eta_r = (x^p - x_m)/(x^p - \varepsilon) \quad \varepsilon \sim N(0, I)$$

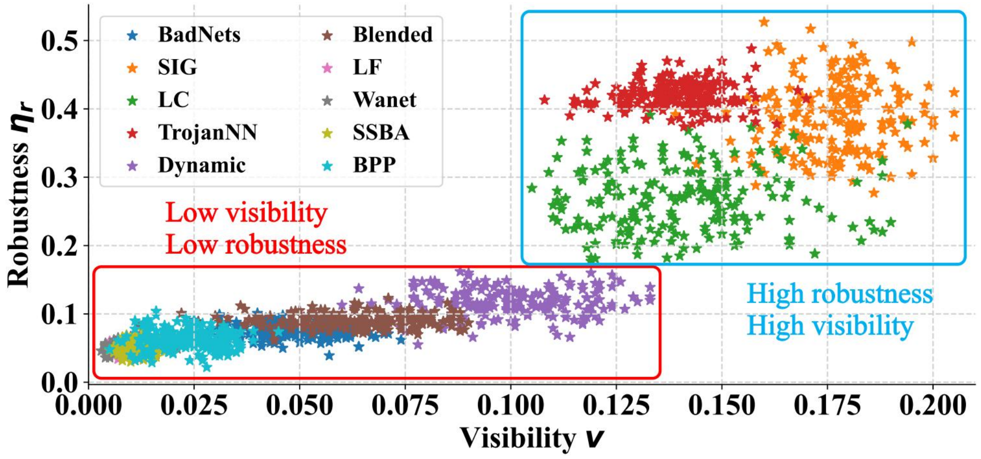


Examples of poisoned samples and their v/η_r.

- **Observations**

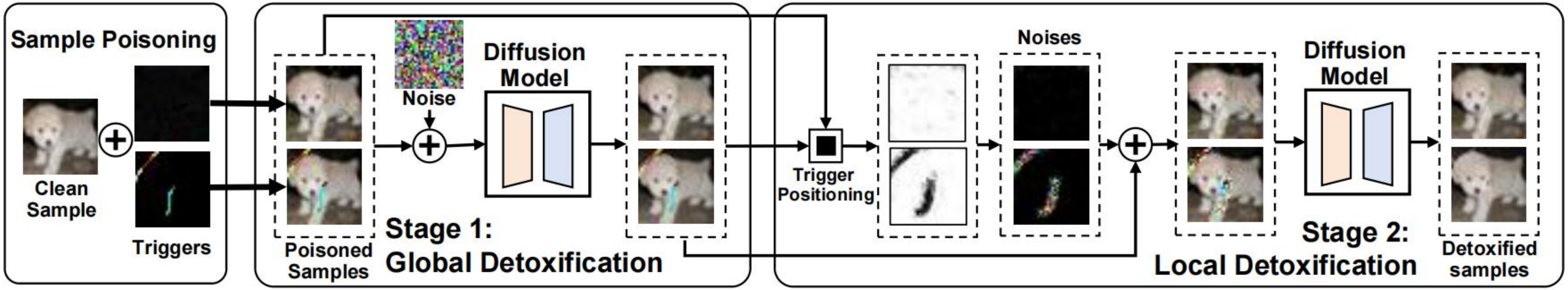
Observation 1:
Low visibility ($v<0.13$), Low robustness ($\eta_r < 0.18$)

Observation 2:
High robustness ($\eta_r \geq 0.18$), High visibility ($v \geq 0.13$)



Correlation between visibility v and robustness η_r.

Our Proposed SampDetox



- **Stage 1: Global Detoxification**

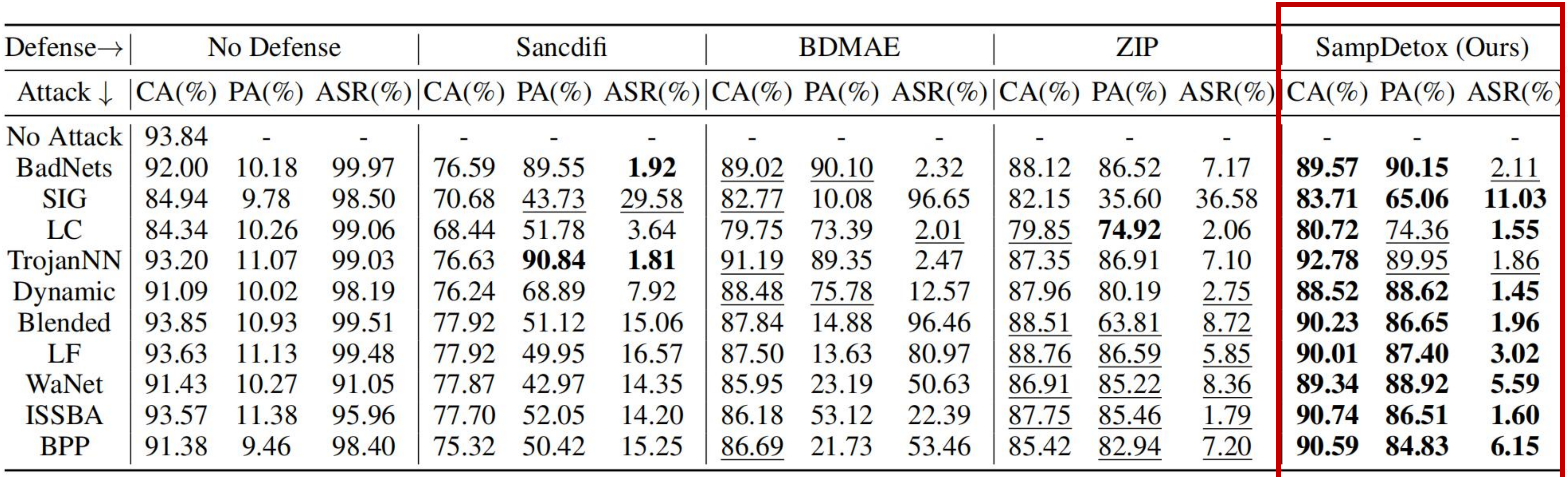
  – **Inspired by Observation 1, this stage aims to destory the backdoor triggers with low visibility but low robustness.**

- **Stage 2: Local Detoxification**

  – **Inspired by Observation 2, this stage aims to destory the backdoor triggers with high robustness but high visibility.**

Experiments: Performance Comparison

- **Comparison with 3 SOTA defenses against 10 attacks**

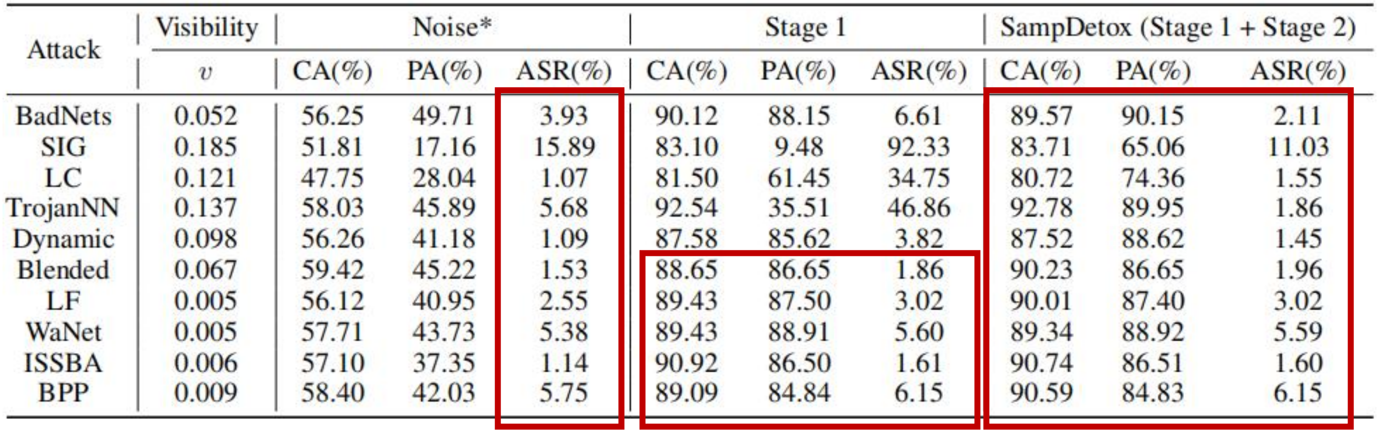


Defense→	No Defense			Sancdifi			BDMAE			ZIP			SampDetox (Ours)		
Attack ↓	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)
No Attack	93.84	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BadNets	92.00	10.18	99.97	76.59	89.55	**1.92**	89.02	90.10	2.32	88.12	86.52	7.17	**89.57**	**90.15**	2.11
SIG	84.94	9.78	98.50	70.68	43.73	29.58	82.77	10.08	96.65	82.15	35.60	36.58	**83.71**	**65.06**	**11.03**
LC	84.34	10.26	99.06	68.44	51.78	3.64	79.75	73.39	2.01	79.85	**74.92**	2.06	**80.72**	74.36	**1.55**
TrojanNN	93.20	11.07	99.03	76.63	**90.84**	**1.81**	91.19	89.35	2.47	87.35	86.91	7.10	**92.78**	89.95	1.86
Dynamic	91.09	10.02	98.19	76.24	68.89	7.92	88.48	75.78	12.57	87.96	80.19	2.75	**88.52**	**88.62**	**1.45**
Blended	93.85	10.93	99.51	77.92	51.12	15.06	87.84	14.88	96.46	88.51	63.81	8.72	**90.23**	**86.65**	**1.96**
LF	93.63	11.13	99.48	77.92	49.95	16.57	87.50	13.63	80.97	88.76	86.59	5.85	**90.01**	**87.40**	**3.02**
WaNet	91.43	10.27	91.05	77.87	42.97	14.35	85.95	23.19	50.63	86.91	85.22	8.36	**89.34**	**88.92**	**5.59**
ISSBA	93.57	11.38	95.96	77.70	52.05	14.20	86.18	53.12	22.39	87.75	85.46	1.79	**90.74**	**86.51**	**1.60**
BPP	91.38	9.46	98.40	75.32	50.42	15.25	86.69	21.73	53.46	85.42	82.94	7.20	**90.59**	**84.83**	**6.15**

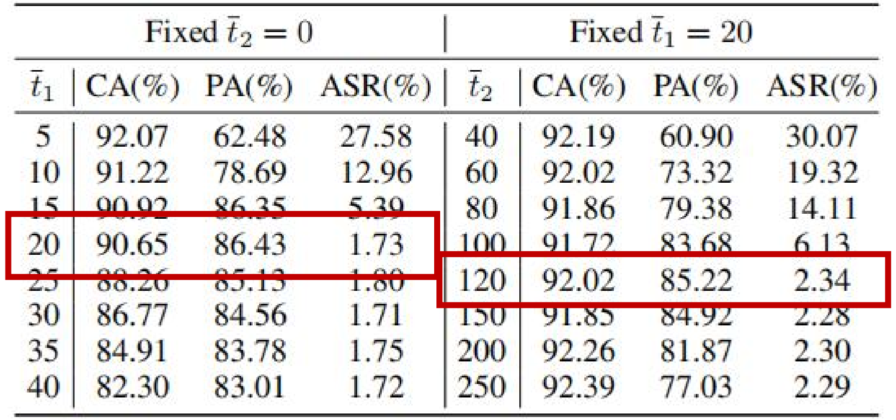
Our SampDetox achieves the best CA, PA and ASR compared to other SOTA defenses against various attacks

Experiments: Ablation Study and Discussion

- **Impacts of Different Stages, Denoising, and Hyperparameters $\bar{t}_1$, $\bar{t}_2$**

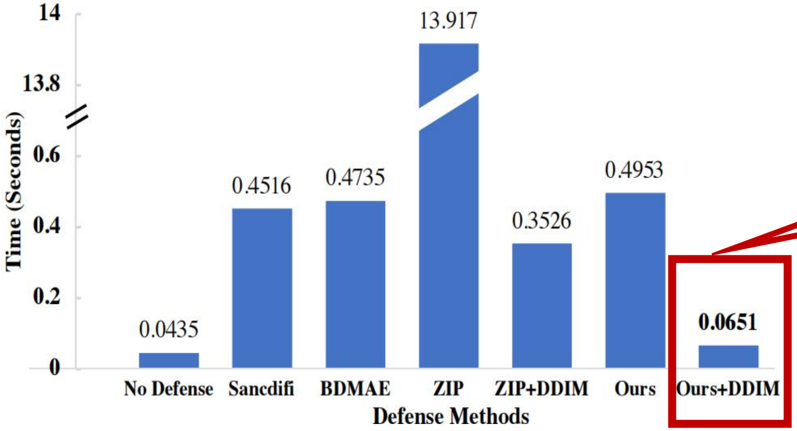


Attack	Visibility	Noise*			Stage 1			SampDetox (Stage 1 + Stage 2)		
	v	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)
BadNets	0.052	56.25	49.71	3.93	90.12	88.15	6.61	89.57	90.15	2.11
SIG	0.185	51.81	17.16	15.89	83.10	9.48	92.33	83.71	65.06	11.03
LC	0.121	47.75	28.04	1.07	81.50	61.45	34.75	80.72	74.36	1.55
TrojanNN	0.137	58.03	45.89	5.68	92.54	35.51	46.86	92.78	89.95	1.86
Dynamic	0.098	56.26	41.18	1.09	87.58	85.62	3.82	87.52	88.62	1.45
Blended	0.067	59.42	45.22	1.53	88.65	86.65	1.86	90.23	86.65	1.96
LF	0.005	56.12	40.95	2.55	89.43	87.50	3.02	90.01	87.40	3.02
WaNet	0.005	57.71	43.73	5.38	89.43	88.91	5.60	89.34	88.92	5.59
ISSBA	0.006	57.10	37.35	1.14	90.92	86.50	1.61	90.74	86.51	1.60
BPP	0.009	58.40	42.03	5.75	89.09	84.84	6.15	90.59	84.83	6.15



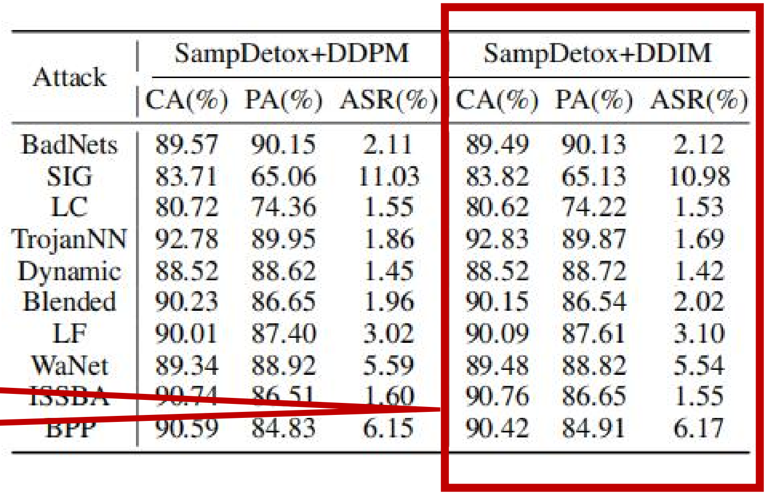
Fixed $\bar{t}_2 = 0$				Fixed $\bar{t}_1 = 20$			
$\bar{t}_1$	CA(%)	PA(%)	ASR(%)	$\bar{t}_2$	CA(%)	PA(%)	ASR(%)
5	92.07	62.48	27.58	40	92.19	60.90	30.07
10	91.22	78.69	12.96	60	92.02	73.32	19.32
15	90.92	86.35	5.39	80	91.86	79.38	14.11
20	90.65	86.43	1.73	100	91.72	83.68	6.13
25	88.26	85.13	1.80	120	92.02	85.22	2.34
30	86.77	84.56	1.71	150	91.85	84.92	2.28
35	84.91	83.78	1.75	200	92.26	81.87	2.30
40	82.30	83.01	1.72	250	92.39	77.03	2.29

- **Extra time overhead and effectiveness using DDIM**



SampDetox's inference time is comparable to that of no defense

Using DDIM does not reduce the effectiveness of SampDetox



Attack	SampDetox+DDPM			SampDetox+DDIM		
	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)
BadNets	89.57	90.15	2.11	89.49	90.13	2.12
SIG	83.71	65.06	11.03	83.82	65.13	10.98
LC	80.72	74.36	1.55	80.62	74.22	1.53
TrojanNN	92.78	89.95	1.86	92.83	89.87	1.69
Dynamic	88.52	88.62	1.45	88.52	88.72	1.42
Blended	90.23	86.65	1.96	90.15	86.54	2.02
LF	90.01	87.40	3.02	90.09	87.61	3.10
WaNet	89.34	88.92	5.59	89.48	88.82	5.54
ISSBA	90.74	86.51	1.60	90.76	86.65	1.55
BPP	90.59	84.83	6.15	90.42	84.91	6.17

Conclusion

- **Problems of Existing Black-box Backdoor Defenses**
  - Detection-based defenses greatly reduce the usability of tasks
  - Purification-based methods are based on the impractical assumption
- **Contributions of our work**
  - Reveal the correlation between the visibility of triggers and the robustness of poisoned samples
  - Present a novel perturbation-based sample detoxification method together with its theoretical foundations
- **Experimental results**
  - Extensive experimental results show the applicability and superiority of our approach over state-of-the-art (SOTA) backdoor defense methods

Thank You !