Exploring Functional Similarities of Backdoored Models

Yufan Feng [1] **Benjamin Tan** [1] **Yani Ioannou** [1]

Abstract

Backdoor attacks embed hidden behaviors into neural networks, causing misclassification when specific triggers are present. While many backdoor methods differ in trigger design or poisoning strategy, they often share a common goal: mapping any triggered input to a fixed target label. This paper investigates whether such attacks lead to similar functional behavior of poisoned models. We introduce a framework to analyze backdoored models from a function space perspective, using metrics over both hard and soft predictions. Our study includes two aspects: (1) the consistency of each attack's learned function across training runs, and (2) the functional similarity across different attack strategies. Results show that some attacks (e.g., FTrojanNN, SSBA) yield stable, convergent behavior, while others (e.g., WaNet, Input-Aware) are highly variable. Cross-attack comparisons reveal functional clusters, particularly among clean-label methods, while visible or training-controlled attacks deviate more sharply. These findings suggest that even with similar objectives, backdoor methods shape model functions in distinct ways, motivating function-level analysis as a tool for understanding or defending against neural backdoors.

1. Introduction

Deep Neural Networks (DNNs) are parameterized functional mappings from input distributions to output spaces. Although DNNs have achieved remarkable success in a variety of tasks, they remain vulnerable to a wide range of security threats. Backdoor attacks on DNNs, first proposed by Gu et al. (2019), aim to mislead the original functional mapping: backdoored models behave correctly on original clean data during test time, but are forced to misbehave whenever a predefined trigger is present in the input. This attack poses a practical threat when users rely on third-party models or datasets.

To explore the capability of backdoor attacks, researchers have proposed various methods, addressing the needs from different perspectives. For example, work by Turner et al. (2019), Zeng et al. (2021), and Barni et al. (2019) design human imperceptible triggers at different levels to improve the stealthiness of backdoored samples; clean-label attacks (Turner et al., 2019; Barni et al., 2019; Li et al., 2023) keep all labels unchanged during data poisoning; and methods such as Nguyen & Tran (2021; 2020) access the training process itself.

However, despite these varied designs, backdoored attacks share similar objectives (e.g., mapping any triggered input to the target label). This motivates us to look into backdoored models' output behaviour, to evaluate whether they induce similar functional changes in the model's decision function.

In this paper, we are specifically interested in the most common all-to-one (single-targeted) backdoor scenario, where the adversary chooses a single target label, and any input with the trigger is trained to be classified into this single target label. We conduct two complementary lines of evaluation:

• Individual study: For each attack, we evaluate the consistency of the induced decision function under different random training factors.

• Cross-attack study: We measure the functional similarity between different backdoored models, using a diverse suite of output-based metrics over clean and poisoned test distributions.

Our findings reveal significant variation in both within-attack stability and cross-attack similarity, providing new insight into the functional behaviours of backdoor mechanisms.

2. Related work

Backdoor attacks. Early backdoor attacks were based on fixed visible triggers: Badnets (Gu et al., 2019) is the first work that backdoors DNNs by adding a small pattern to a portion of training images, and Blended (Chen et al., 2017) overlays a semi-transparent image trigger. More stealthy patterns have since been explored: Low frequency (Zeng et al., 2021) proposes smooth low-frequency triggers optimized via

[1]Schulich School of Engineering, University of Calgary, Canada. Correspondence to: Yufan Feng <yufan.feng@ucalgary.ca>, Yani Ioannou <yani.ioannou@ucalgary.ca>.

Workshop on Technical AI Governance (TAIG) at the 2025 International Conference on Machine Learning (ICML 2025), Vancouver, Canada.

bilevel programming; FTrojanNN (Wang et al., 2022) perturbs Fourier coefficients; Refool (Liu et al., 2020) overlays natural reflection artifacts. In contrast to those dirty label attacks that might change the original images labels, clean label attacks, proposed by Turner et al. (2019) maintain correct labels on poisoned samples. Other methods, such as SIG (Barni et al., 2019) that adds sinusoidal noise and CTRL (Li et al., 2023) that defines the trigger in the spectral space, also fall into this clean-label attack category. Beyond data poisoning, several methods modify the training process for finer control: WaNet (Nguyen & Tran, 2021) applies imperceptible geometric wraps during training; Input-Aware (Nguyen & Tran, 2020) learns a trigger generator during training.

Evaluating backdoor attacks. Benchmark studies such as BackdoorBench (Wu et al., 2022) and TrojanZoo (Pang et al., 2022) implement various attacks and defenses in a unified framework, revealing important backdoor features, e.g., influences of poisoned sample selection, model structure, poisoning ratio, and trigger hyperparameters. Additionally, several papers analyze backdoor attacks in different aspects. On the learning dynamics side, Li et al. (2021) and Yuan et al. (2025) observe that the loss of backdoored samples typically decreases faster during the early training stage. Zhang et al. (2024) conduct a formal analysis on how different backdoor learning behaves as orthogonal and linear sub-tasks in the network. On the representational side, analysis often yields to backdoor sample detections. Chen et al. (2018) explores the separability of the hidden-layer activations of poisoned and clean samples. Tran et al. (2018) uses robust statistics in singular-vector space. Jebreel et al. (2023) analyzes the backdoor samples in layer-wise feature space.

Despite this rich ecosystem of attacks and defenses, as far as we are aware no prior work systematically compares the functional similarity of models trained under different backdoor methods.

3. Functional similarity evaluation

Functional similarity quantifies how similarly a set of models behaves on the same input distribution, which has a long history in ensemble learning (Klabunde et al., 2023). In the backdoor context, we evaluate models over two matched test distributions derived from a common base dataset.

- $\mathcal{X}_{\text{clean}}$: the unmodified clean test dataset;
- $\mathcal{X}_{\text{poison}}$: the same inputs with the trigger applied, excluding samples whose true label already equals the adversary's target class, to prevent artificially high attack successful rates in all-to-one scenario.

For any input $\mathbf{x}_i \in \mathcal{X}$ and a model f_θ, we extract three forms of output as the basis for output evaluation:

- Logits $\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^C$, where C is the number of classes
- Softmax probabilities $\mathbf{p}_i = \mathrm{softmax}(\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^{C}\exp(\mathbf{z}_j)}$
- Predicted label $\hat{y}_i = \mathrm{argmax}(\mathbf{p}_i)$

In previous literature, clean accuracy (CA) and attack successful rate (ASR) are defined as top-1 predicted label accuracy computed on $\mathcal{X}_{\text{clean}}$ and $\mathcal{X}_{\text{poison}}$, respectively; they fit naturally into our functional similarity framework. On top of that, we employ a suite of similar metrics that operate on the hierarchy of outputs, introduced in Appendix B.

4. Individual study of attacks

This section presents an analysis of various backdoor attacks, evaluating both their effectiveness and the consistency of their learned functions under random training factors such as random initialization and data shuffling.

The results are shown in Table 1, and the complete result are shown in appendix. On the clean test set, we can observe different levels of accuracy and consistency decreasing. Training-controlled methods (WaNet and Input-Aware) exhibited notable impact on the clean learning task, as shown by lower accuracies and greater disagreement, suggesting variability in their learned functions. BadNets also shows a large deviation from the clean baseline. Attacks that optimize human imperceptible triggers (LF, Blended, SSBA and FTrojanNN), tend to maintain high accuracy and exhibit low disagreement metrics, showing that their decision boundaries remain almost identical under random factors. Among soft metrics, the logit norm (lNorm) is systematically higher than the probability norm (pNorm), indicating that differences between different runs are mostly scale shifts in confidence, rather than probability reshuffling; SSBA's and FTrojanNN's low lNorm values corroborate their tight functional clustering.

Upon evaluating the poisoned test set, FTrojanNN and SSBA almost achieve perfect attack success with minimal variance, highlighting their robustness. In contrast, WaNet and Input-Aware show huge spreads — disagreement 17% and 13%, κ near zero, and cJSD exploding, illustrating that training-controlled randomness not only lowers ASR but also makes the triggered decision function unstable. Refool underperformed across different poisoning ratios, indicating its relative weakness. Clean-label attack methods (LC, SIG, CTRL) and LF are sensitive to the poisoning ratio, as their performances drop significantly at a 1% poisoning rate.

5. Cross-attack evaluation

In this section, we extend our analysis to cross-attack evaluations, aiming to understand the functional similarities and differences among various backdoor attack methods. Specifically, we examine pairwise functional similarity metrics across models trained with different backdoor

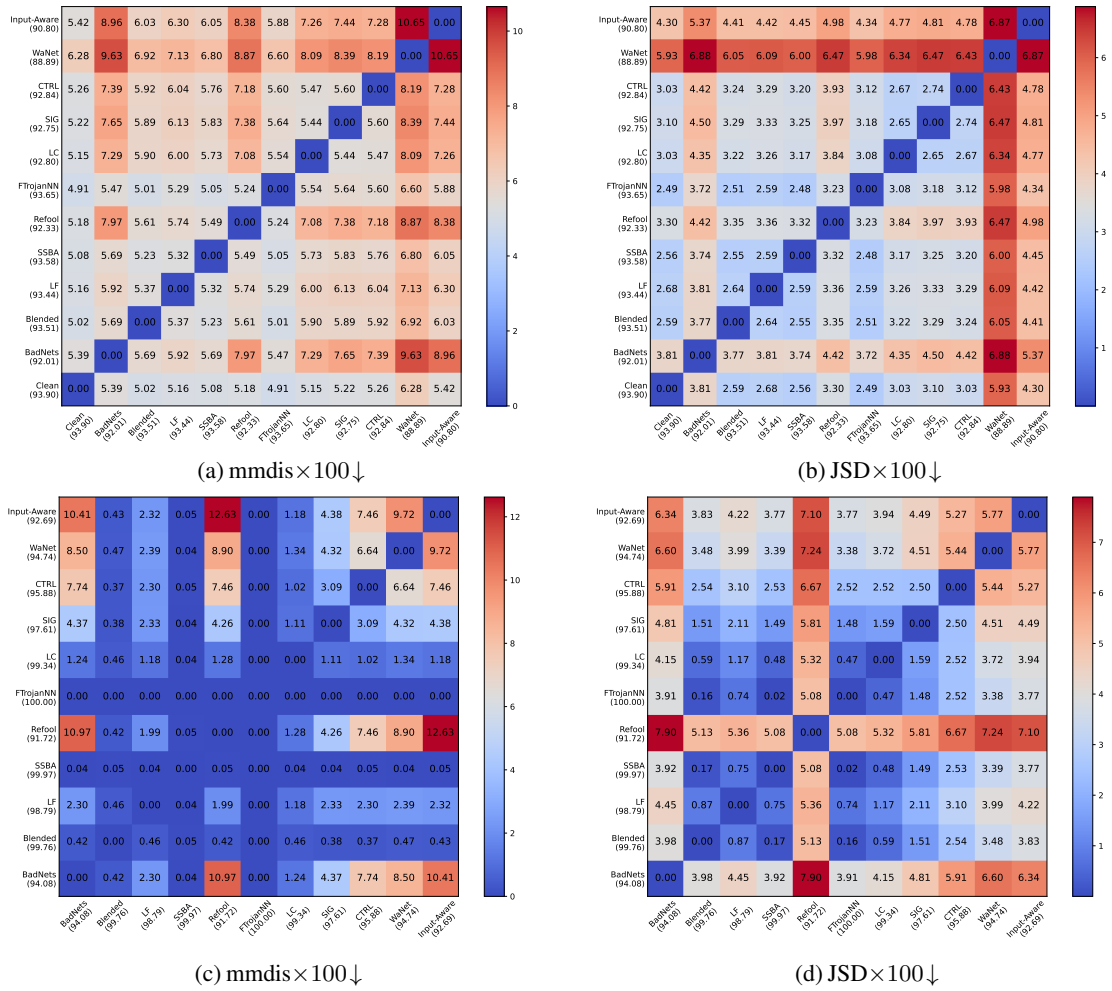
Table 1. **Individual study of different attacks.** Each attack is evaluated on Preact-ResNet18 (He et al., 2016) on CIFAR-10 with 5 independent runs. For pairwise metrics, scores are computed between all model pairs and then averaged. The accuracies are shown in the format of mean(std). In each column, the best-performing result is highlighted in green, while the worst is highlighted in red.

Attack	Pratio	Hard predict							Soft predict			
		acc↑ %	dis↓ %	ecdis↓ ×1	mmdis↓ ×100	κ_c↑ ×100	κ_f↑ ×100	cJSD↓ ×100	lNorm↓ ×1	pNorm↓ ×10	sChurn↓ ×10	JSD↓ ×100
$\mathcal{X}_{\text{clean}}$												
Clean	–	93.90 (0.17)	5.19	0.85	4.98	94.24	94.24	3.74	3.43	0.51	0.59	2.44
BadNets	8%	92.01 (0.12)	8.81	1.10	8.67	90.21	90.21	3.03	5.70	0.90	1.03	4.89
Blended	8%	93.51 (0.17)	5.67	0.87	5.46	93.70	93.70	3.29	3.56	0.56	0.65	2.71
LF	8%	93.44 (0.09)	5.67	0.86	5.56	93.70	93.70	2.39	3.67	0.55	0.65	2.68
SSBA	8%	93.58 (0.19)	5.64	0.88	5.40	93.74	93.74	3.28	3.56	0.55	0.63	2.65
Refool	8%	92.33 (0.07)	6.65	0.87	6.56	92.61	92.61	3.76	4.00	0.65	0.75	3.20
FTrojanNN	8%	93.65 (0.19)	5.59	0.88	5.35	93.79	93.79	1.88	3.46	0.55	0.64	2.70
LC	8%	92.80 (0.28)	6.14	0.85	5.79	93.17	93.17	3.34	3.41	0.59	0.70	2.87
SIG	8%	92.75 (0.23)	6.23	0.86	5.92	93.08	93.08	3.71	3.51	0.60	0.70	2.94
CTRL	8%	92.84 (0.20)	6.10	0.85	5.86	93.22	93.22	2.37	3.45	0.60	0.70	2.91
WaNet	8%	88.89 (2.09)	14.28	1.31	11.73	84.12	84.12	336.42	4.51	1.41	1.69	7.24
Input-Aware	8%	90.80 (0.80)	11.02	1.20	10.13	87.75	87.75	144.64	4.66	1.04	1.25	5.43
BadNets	1%	93.60 (0.08)	6.01	0.94	5.91	93.33	93.33	2.17	4.28	0.59	0.69	2.98
Blended	1%	93.69 (0.09)	5.52	0.88	5.41	93.87	93.87	2.86	3.52	0.52	0.61	2.54
LF	1%	93.55 (0.12)	5.60	0.87	5.45	93.78	93.78	2.33	3.56	0.54	0.63	2.62
SSBA	1%	93.85 (0.17)	5.23	0.85	5.03	94.19	94.19	1.58	3.44	0.51	0.59	2.46
Refool	1%	93.38 (0.17)	5.86	0.89	5.64	93.49	93.49	3.19	3.74	0.57	0.66	2.77
FTrojanNN	1%	93.84 (0.16)	5.39	0.88	5.19	94.01	94.01	1.71	3.42	0.51	0.61	2.47
LC	1%	93.82 (0.09)	5.23	0.85	5.11	94.19	94.19	2.73	3.38	0.51	0.59	2.44
SIG	1%	93.76 (0.05)	5.26	0.84	5.20	94.15	94.15	3.24	3.42	0.51	0.60	2.42
CTRL	1%	93.85 (0.16)	5.28	0.86	5.09	94.13	94.13	2.57	3.39	0.51	0.59	2.43
WaNet	1%	90.63 (1.25)	11.49	1.24	9.90	87.23	87.23	116.57	4.41	1.16	1.40	5.87
Input-Aware	1%	90.87 (0.68)	10.85	1.19	10.07	87.95	87.95	95.69	4.97	1.12	1.33	6.01
$\mathcal{X}_{\text{poison}}$												
BadNets	8%	94.08 (0.57)	3.87	0.66	3.14	66.28	66.17	9.88	4.29	2.07	3.25	13.59
Blended	8%	99.76 (0.15)	0.33	1.81	0.13	32.23	31.38	0.66	3.83	1.69	3.04	12.46
LF	8%	98.79 (0.26)	1.14	0.97	0.80	52.34	52.25	1.90	3.15	1.03	1.68	7.07
SSBA	8%	99.97 (0.02)	0.04	–	0.02	17.52	20.82	0.01	3.32	1.06	1.86	7.69
Refool	8%	91.72 (0.35)	10.71	1.30	10.27	32.02	32.04	3.84	4.44	1.41	2.38	11.49
FTrojanNN	8%	100.00 (0.00)	0.00	–	0.00	–	–	0.00	3.09	0.83	1.41	5.95
LC	8%	99.34 (0.43)	1.06	2.32	0.48	16.99	18.64	5.22	2.86	0.62	1.08	4.30
SIG	8%	97.61 (0.58)	2.49	1.09	1.71	46.95	47.22	9.69	3.51	1.15	2.06	8.65
CTRL	8%	95.88 (1.26)	4.10	1.13	2.45	48.10	49.04	46.03	3.19	0.93	1.63	6.78
WaNet	8%	94.74 (2.60)	8.62	2.01	5.37	15.80	15.56	202.59	2.94	1.02	1.75	7.05
Input-Aware	8%	92.69 (1.93)	13.27	1.92	10.86	5.14	5.33	113.19	3.39	1.00	1.71	7.12
BadNets	1%	77.97 (4.24)	14.27	0.66	8.91	63.42	63.05	655.94	3.22	0.79	1.31	5.33
Blended	1%	96.32 (0.33)	3.41	0.93	2.98	52.56	52.60	3.22	3.73	1.48	2.66	11.05
LF	1%	86.60 (1.61)	8.71	0.66	6.74	64.66	64.82	83.84	3.00	0.70	1.16	4.89
SSBA	1%	98.97 (0.19)	1.08	1.06	0.82	47.55	47.53	1.01	3.18	0.86	1.48	6.19
Refool	1%	47.16 (1.85)	35.41	0.67	32.01	52.50	52.51	208.04	4.18	1.22	2.08	9.87
FTrojanNN	1%	99.90 (0.02)	0.11	1.16	0.08	45.14	45.52	0.02	3.02	0.70	1.18	4.97
LC	1%	82.46 (0.45)	12.35	0.70	11.80	60.89	60.89	6.91	2.86	0.60	1.02	4.15
SIG	1%	84.05 (3.07)	11.21	0.73	7.27	61.20	61.38	312.02	3.48	1.10	1.94	8.18
CTRL	1%	64.92 (11.04)	26.02	0.93	14.01	52.68	53.81	4805.33	3.07	0.83	1.44	5.92
WaNet	1%	77.00 (6.79)	26.74	1.25	19.04	33.35	33.20	1710.07	2.80	0.65	1.12	4.29
Input-Aware	1%	67.32 (9.82)	41.41	1.35	32.47	22.61	22.57	4338.78	3.42	0.78	1.35	5.56

strategies. To ensure fair comparison, we control for randomness: each pairwise comparison uses models trained with the same model initialization, data shuffling, and poisoned indices as much as possible. Poisoned indices are aligned within attack categories that share the same labeling target, but differ across clean- and dirty-label by design. The poisoned inputs are generated dynamically for training-controlled methods and can not be aligned.

We use two representative metrics — min-max normalized disagreement to assess hard boundaries and Jensen-Shannon divergence on raw logits, shown in Figure 1. Other metrics are shown in the appendix. On $\mathcal{X}_{\text{clean}}$, clean label attacks (LC, SIG, CTRL) form a tight cluster with their metrics similar to others, exhibiting low mutual JSD and relatively low min-max disagreement, indicating that these methods produce similar output distributions. Similarly, imperceptible attacks like FTrojanNN, SSBA, and LF show tighter functional alignment. BadNets, Refool, WaNet, and Input-Aware are consistently outliers, suggesting that the presence of visible artifacts or perturbation of the training process dramatically changes both the decision boundaries and the output distributions. While comparing two metrics respectively, some attack pairs (e.g. LC-FTrojanNN) show low min-max disagreement but higher JSD, suggesting they produce similar labels but for different reasons.

On the poisoning test set, a dominant trend is that min-max normalized disagreement strongly correlates with accuracy, even though it is normalized against theoretical upper and lower bounds. Several attacks achieve nearly perfect attack success (Blended: 99.76; SSBA: 99.97; FTrojanNN: 100.00; LC: 99.34), thus forming a low-disagreement cluster where most pairwise mmdis values are close to zero. Despite this,



(a) mmdis×100↓ (b) JSD×100↓

(c) mmdis×100↓ (d) JSD×100↓

Figure 1. **Cross attack evaluation.** We compare each pair of attacks on the two representative metrics, min-max-normalized disagreement on raw predictions and Jensen-Shannon divergence on logits. (a) and (b) are evaluated on $\mathcal{X}_{\text{clean}}$, (c) and (d) are on $\mathcal{X}_{\text{poison}}$. All attacks are evaluated on Preact-ResNet18 on CIFAR-10 with 8% poisoning rate. For convenience, the accuracies are commented under each attack name.

JSD offers a more nuanced view. For example, although the ASR of SSBA (99.97) and FtrojanNN (100.00) are high, CTRL still has a smaller JSD (CTRL-SIG: 2.50; CTRL-SSBA: 2.53; CTRL-FTrojanNN: 2.52) when paired with SIG (97.61), suggesting that CTRL is relatively close to SIG in logit space. Similar pattern emerges on BadNets-LF vs. BadNets-SIG. BadNets, Blended, WaNet, and Input-Aware remain outliers, exhibiting high disagreement with nearly every other attack.

6. Conclusion

In this paper, we introduce a functional perspective for evaluating and comparing backdoor attacks on DNNs, providing a principled way to compare backdoor behaviours beyond surface metrics. Our study reveals that while many attacks achieve similar levels of attack success, they differ in their effect on the function learned by a backdoored model. Notably, training-controlled attacks like WaNet and Input-Aware exhibit inconsistent behaviours across runs, whereas optimized attacks such as SSBA and FTrojanNN converge to more stable decision boundaries. Cross-attack comparisons further reveal distinct functional clusters: clean-label attacks tend to behave similarly, while visible or input-dependent triggers introduce larger divergences in both prediction and confidence spaces. **Future Work:** Extending our evaluation to larger and more datasets, architectures, and attack settings, and further explore the internal representations of models to understand how backdoor attacks affect intermediate model functions. **Policy Implications:** Our findings underscore the need for comprehensive AI security compliance programs. As backdoor attacks pose significant challenges for detection and mitigation, especially in critical sectors like healthcare, finance, and autonomous systems, implementing standardized evaluation frameworks and certification processes can help ensure model integrity. Moreover, our results, alongside broader research in the field, suggest that there is unlikely to be one set of policies that can govern all attacks.

ACKNOWLEDGMENTS

We acknowledge the the support of Alberta Innovates (ALLRP-577350-22, ALLRP-222301502), the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-03120, DGECR-2022- 00358), and Defence Research and Development Canada (DGDND-2022-03120). We are grateful for computational resources made available to us by the Digital Research Alliance of Canada.

Y. Feng is supported by the Alberta Graduate Excellence Scholarship. The work of B. Tan is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN- 2022-03027]. Cette recherche a eté financée en partie par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

References

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105. IEEE, 2019.

Bhojanapalli, S., Wilber, K., Veit, A., Rawat, A. S., Kim, S., Menon, A., and Kumar, S. On the reproducibility of neural network predictions. *arXiv preprint arXiv:2102.03349*, 2021.

Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., and Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2017. URL http://arxiv.org/abs/1712.05526.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46, 1960.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jebreel, N. M., Domingo-Ferrer, J., and Li, Y. Defending against backdoor attacks by layer-wise feature analysis. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part II*, pp. 428–440, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-33376-7. doi: 10.1007/978-3-031-33377-4_33. URL https://doi.org/10.1007/978-3-031-33377-4_33.

Klabunde, M. and Lemmerich, F. On the prediction instability of graph neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 187–202. Springer, 2022.

Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, C., Pang, R., Xi, Z., Du, T., Ji, S., Yao, Y., and Wang, T. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4367–4378, 2023.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

Liu, Y., Ma, X., Bailey, J., and Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part X 16*, pp. 182–199. Springer, 2020.

Nguyen, T. A. and Tran, A. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.

Nguyen, T. A. and Tran, A. T. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eEn8KTtJOx.

Pang, R., Zhang, Z., Gao, X., Xi, Z., Ji, S., Cheng, P., Luo, X., and Wang, T. Trojanzoo: Towards unified, holistic, and practical evaluation of neural backdoors. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 684–702. IEEE, 2022.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Turner, A., Tsipras, D., and Madry, A. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

Wang, T., Yao, Y., Xu, F., An, S., Tong, H., and Wang, T. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, pp. 396–413. Springer, 2022.

Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., and Shen, C. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35:10546–10559, 2022.

Yuan, D., Zhang, M., Wei, S., Liu, L., and Wu, B. Activation gradient based poisoned sample detection against backdoor attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VNMJfBBUd5.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16473–16481, 2021.

Zhang, K., Cheng, S., Shen, G., Tao, G., An, S., Makur, A., Ma, S., and Zhang, X. Exploring the orthogonality and linearity of backdoor attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2105–2123. IEEE, 2024.

Appendix

A. Discussion

Limitations. While our study investigates functional similarity across a range of backdoor attacks, it is primarily grounded in experiments on CIFAR-10, with additional results on CIFAR-100; both remain relatively small-scale and image-classification-focused. Extending our evaluation to larger and more diverse datasets (e.g., ImageNet) and architectures (e.g., VGGs and ViTs) would further validate the generality of our observations. Moreover, our analysis is restricted to the all-to-one attack setting, which does not capture more complex threat models such as multi-targeted, all-to-all settings, or multi-trigger attacks, where multiple triggers with different target behaviors may co-exist. Finally, training-controlled attacks (e.g., WaNet, Input-Aware) inherently involve randomness in poison generation, which limits our ability to control for training factors and precisely align comparisons.

Future directions. Building on the current findings, we plan to extend our analysis along several dimensions. First, even though our analysis discovered some patterns, the findings are still relatively coarse. We aim to move beyond output-level similarity and explore functional representations within the model, such as intermediate activations and representational trajectories, to localize the backdooring features better. Second, we intend to apply our functional framework to more complex and realistic attack scenarios, particularly multi-trigger or hybrid attacks, where multiple triggers may target different classes or interact in unpredictable ways. Third, we are interested in studying functional generalization and transferability—whether backdoors trained under one condition (e.g., dataset, architecture) induce functionally similar behavior when transferred or fine-tuned. Finally, we hope to use insights from functional similarity to inspire defense strategies, such as detecting anomalous functional clusters in ensemble or federated settings.

B. Metrics for functional similarity

Disagreement. This directly measures the proportion of inputs for which the two models predicted different classes.

$$\mathrm{dis}(f^{(1)},f^{(2)}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i^{(1)} \neq y_i^{(2)}\}. \tag{1}$$

The metric is sensitive to flips in the decision boundary rather than confidence: any change that causes a different top-1 label prediction will result in disagreement.

Error-corrected disagreement. To account for the fact that two highly accurate models cannot disagree more often than their individual rates, we normalize raw disagreement by each model's error following (Fort et al., 2019), and then average symmetrically as

$$\mathrm{ecdis}(f^{(1)},f^{(2)}) = \frac{1}{2}\left[\frac{\mathrm{dis}(f^{(1)},f^{(2)})}{\mathrm{Err}(f^{(1)})} + \frac{\mathrm{dis}(f^{(1)},f^{(2)})}{\mathrm{Err}(f^{(2)})}\right], \tag{2}$$

where $\mathrm{Err}(\cdot)$ denotes the error rate. The metric measures the two models' disagreement that falls within their error budgets. Here, the minimum value 0 indicates models always agree, and the maximum value 1 suggests that their disagreement matches each model's average error rate.

Min-Max-normalized disagreement (Klabunde & Lemmerich, 2022). Another method to re-scale the disagreement is to leverage its theoretical minimum and maximum given the error rates

$$\mathrm{mmdis}(f^{(1)},f^{(2)}) = \frac{D-L}{U-L}, \tag{3}$$

where $D = \mathrm{dis}(Y^{(1)},Y^{(2)})$, $U = |\mathrm{Err}(Y^{(1)}) - \mathrm{Err}(Y^{(2)})|$ and $L = \min\left[\mathrm{Err}(Y^{(1)}) + \mathrm{Err}(Y^{(2)})\right]$ are the upper and lower bound for disagreement.

Cohen's kappa (Cohen, 1960). The kappa statistic corrects agreement on chance-level overlap, measuring the agreement beyond what random coinciding predictions would produce, given the marginal label distributions.

$$\kappa_c(f^{(1)},f^{(2)}) = 1 - \frac{\mathrm{dis}(f^{(1)},f^{(2)})}{1-p_e} = \frac{p_o - p_e}{1-p_e}, \tag{4}$$

where $p_o = 1 - \mathrm{dis}(f^{(1)}, f^{(2)})$ is the observed agreement and $p_e = \frac{1}{N^2}\sum_{c=1}^{C} n_c^{(1)} n_c^{(2)}$ is the expected agreement by chance. The kappa value is highly influenced by class imbalance and prediction bias, as when one class prediction dominates, a high raw disagreement can yield a high value.

Fleiss's kappa (Fleiss, 1971). This extends Cohen's kappa from pairwise agreement to multi-rater cases.

Class Jensen-Shannon divergence. We take the frequency histogram of class labels over the test set for each model, and compute the pairwise Jensen-Shannon divergence

$$\begin{aligned} \mathbf{q} &= (q_1, \ldots, q_C)^\top, \quad q_c = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i = c\}, \\ \mathrm{cJSD}(f^{(1)}, f^{(2)}) &= \frac{1}{2} D_{\mathrm{KL}}\left(\mathbf{q}^{(1)} \| M\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\mathbf{q}^{(2)} \| M\right), \end{aligned} \tag{5}$$

where $M = \frac{1}{2}\left(\mathbf{q}^{(1)} + \mathbf{q}^{(2)}\right)$. It complements disagreement as it tracks distributional shift rather than instance-level alignment.

Norm of difference. We calculate norm between the logits and probabilities

$$\mathrm{lNorm}(f^{(1)}, f^{(2)}) = \frac{1}{2N}\sum_{i=1}^{N} \left\| \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} \right\|_p, \tag{6}$$

$$\mathrm{pNorm}(f^{(1)}, f^{(2)}) = \frac{1}{2N}\sum_{i=1}^{N} \left\| \mathbf{p}_i^{(1)} - \mathbf{p}_i^{(2)} \right\|_p. \tag{7}$$

For a group of models, instead of matching the pairwise differences, we compare the individual logits or probabilities over their average. In practice, p is set to be 2. They measure the confidence-spread changes on predictions.

Surrogate churn (Bhojanapalli et al., 2021). Surrogate churn is proposed as a differentiable version of raw disagreement that also accounts for confidence shifts

$$\mathrm{sChurn}(f^{(1)}, f^{(2)}) = \frac{1}{2N}\sum_{i=1}^{N} \left\| \left(\frac{\mathbf{p}_i^{(1)}}{\max_c p_{i,c}^{(1)}} \right)^{\alpha} - \left(\frac{\mathbf{p}_i^{(2)}}{\max_c p_{i,c}^{(2)}} \right)^{\alpha} \right\|_1. \tag{8}$$

Here, for $\alpha \to \infty$ it recovers raw disagreement, while $\alpha = 1$ weights confidence differences linearly.

Jensen-shannon Divergence. We apply the Jensen-shannon divergence on the output probability distributions.

$$\mathrm{JSD}(f^{(1)}, f^{(2)}) = \frac{1}{2N}\sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbf{z}_i^{(1)} \| M) + D_{\mathrm{KL}}(\mathbf{z}_i^{(2)} \| M). \tag{9}$$

where $M = \frac{1}{2}\left(\mathbf{z}_i^{(1)} + \mathbf{z}_i^{(2)}\right)$. It emphasizes discrepancies in low-probability entries, and can detect subtle shifts that the norm of difference may understate. Also, it often correlates with the cross-entropy gap between models.

C. Experiment details

C.1. Datasets

Our experiments are conducted on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), two classical image classification datasets. CIFAR-10 consists of 60,000 32×32 color images evenly across 10 classes, with 50,000 for training and 10,000 for testing. CIFAR-100 has the same data format but includes 100 fine-grained classes, each with 600 images.

We listed all the hyperparameters used and the characteristics of each attack in Table 2. For CIFAR-10, we evaluate attacks under poisoning ratios of 1%, 5%, 8%, and 10%. For CIFAR-100, we use 1%, 5%, and 10%. The poisoning ratio is defined as the number of poisoned samples over the total training set size. In clean-label attacks (e.g., LC, SIG, CTRL), poisoned samples are restricted to the target class by definition. As a result, the maximum feasible poisoning ratio is limited by the size of that class. By default, the target label is set to 0 for all attacks.

Table 2. **Summary of backdoor attack configurations.** We detail each method's core hyperparameters and trigger characteristics in this table. For every attack, we list the stage at which it is applied (data poisoning or training controlled), the trigger's visibility (visible, stealthy, or invisible), its spatial coverage (local or global), the target type (dirty-label or clean-label), the fusion mechanism by which the trigger is embedded, and whether the trigger is sample-agnostic or tied to specific inputs.

Attack	Hyperparameters	Stage	Visibility	Coverage	Label strategy	Embedding type	Trigger Scope
BadNets	size: 3×3 position: bottom-right	poisoning	visible	local	dirty	additive	agnostic
Blended	blend α: 0.2	poisoning	stealthy	global	dirty	additive	agnostic
LF	blend α: 0.2	poisoning	stealthy	global	dirty	additive	agnostic
SSBA	/	poisoning	stealthy	global	dirty	additive	specific
Refool	ghost rate: 0.49 α_t: 0.4	poisoning	stealthy	global	dirty	additive	agnostic
FTrojanNN	trigger channel: [1,2] magnitude: 30 window size: 32 pos. list: [15,15,31,31]	poisoning	invisible	global	dirty	non-additive	specific
LC	PGD step size: 1.5 PGD steps: 100 $\epsilon=8$ $\delta=40$	poisoning	stealthy	local	clean	additive	agnostic
SIG	f: 6 amplitude Δ: 40	poisoning	stealthy	global	clean	additive	agnostic
CTRL	trigger channel: [1,2] trigger pos.: (15,31)	poisoning	stealthy	global	clean	non-additive	agnostic
WaNet	warp strength s: 0.5 grid scale k: 5	training	invisible	global	dirty	warping	specific
Input-Aware	mask density: 0.032 clean train epochs: 25 λ_{div}: 1 λ_{norm}: 100	training	visible	local	dirty	additive	specific

C.2. Training setup

We implement all attacks based on BackdoorBench[1] (Wu et al., 2022), which is a widely used and reproducible framework for backdoor attacks. All experiments are conducted using PreAct-ResNet18 architecture. The hyperparameters are listed in Table 3.

Table 3. **Model Training Settings.**

Parameter	Value
optimizer	SGD
momentum	0.9
weight decay	5e-4
learning rate scheduler	Cosine Annealing
initial learning rate	0.01
batch size	128
total epochs	100

D. Result of individual study.

In this section, we list the full results of the individual study part. In Table 4 and Table 5, we present the CIFAR-10 evaluation on the clean and poisoned test sets, respectively. In Table 6, we present the result for CIFAR-100 dataset. Patterns of attack robustness appear to be similar across datasets and different poisoning rates: SSBA and FTrojanNN maintain effectiveness and consistency, whereas training-controlled methods, BadNets, and Refool are relatively unstable.

[1] https://github.com/SCLBD/BackdoorBench

Table 4. Individual study on CIFAR-10 $\mathcal{X}_{\text{clean}}$. For every column in every group of poisoning ratios, the best is highlighted with green, while the worst is highlighted with red.

attack	pratio	hard predict							soft predict			
		acc↑ %	dis↓ %	ecdis↓ ×1	mmdis↓ ×100	cKappa↑ ×100	fKappa↑ ×100	cJSD↓ ×100	lNorm↓ ×1	pNorm↓ ×10	sChurn↓ ×10	JSD↓ ×100
Clean	–	93.90 (0.17)	5.19	0.85	4.98	94.24	94.24	3.74	3.43	0.51	0.59	2.44
BadNets	10%	91.67 (0.15)	9.47	1.14	9.31	89.48	89.48	4.03	5.74	0.96	1.09	5.24
Blended	10%	93.46 (0.06)	5.75	0.88	5.67	93.61	93.61	2.14	3.53	0.56	0.65	2.75
LF	10%	93.27 (0.04)	5.91	0.88	5.86	93.43	93.43	2.16	3.67	0.57	0.67	2.82
SSBA	10%	93.51 (0.25)	5.75	0.89	5.46	93.62	93.62	2.76	3.52	0.56	0.65	2.73
Refool	10%	92.19 (0.09)	6.80	0.87	6.69	92.44	92.44	3.00	4.02	0.65	0.76	3.23
FTrojanNN	10%	93.45 (0.09)	5.81	0.89	5.69	93.54	93.54	3.29	3.44	0.57	0.66	2.78
WaNet	10%	89.40 (2.32)	13.33	1.28	10.74	85.18	85.18	269.58	4.67	1.32	1.58	6.74
BadNets	8%	92.01 (0.12)	8.81	1.10	8.67	90.21	90.21	3.03	5.70	0.90	1.03	4.89
Blended	8%	93.51 (0.17)	5.67	0.87	5.46	93.70	93.70	3.29	3.56	0.56	0.65	2.71
LF	8%	93.44 (0.09)	5.67	0.86	5.56	93.70	93.70	2.39	3.67	0.55	0.65	2.68
SSBA	8%	93.58 (0.19)	5.64	0.88	5.40	93.74	93.74	3.28	3.56	0.55	0.63	2.65
Refool	8%	92.33 (0.07)	6.65	0.87	6.56	92.61	92.61	3.76	4.00	0.65	0.75	3.20
FTrojanNN	8%	93.65 (0.19)	5.59	0.88	5.35	93.79	93.79	1.88	3.46	0.55	0.64	2.70
LC	8%	92.80 (0.28)	6.14	0.85	5.79	93.17	93.17	3.34	3.41	0.59	0.70	2.87
SIG	8%	92.75 (0.23)	6.23	0.86	5.92	93.08	93.08	3.71	3.51	0.60	0.70	2.94
CTRL	8%	92.84 (0.20)	6.10	0.85	5.86	93.22	93.22	2.37	3.45	0.60	0.70	2.91
WaNet	8%	88.89 (2.09)	14.28	1.31	11.73	84.12	84.12	336.42	4.51	1.41	1.69	7.24
Input-Aware	8%	90.80 (0.80)	11.02	1.20	10.13	87.75	87.75	144.64	4.66	1.04	1.25	5.43
BadNets	5%	92.59 (0.29)	7.96	1.08	7.60	91.16	91.16	2.49	5.35	0.81	0.92	4.30
Blended	5%	93.64 (0.17)	5.74	0.90	5.53	93.62	93.62	2.39	3.55	0.55	0.64	2.66
LF	5%	93.41 (0.11)	5.80	0.88	5.67	93.55	93.55	3.54	3.65	0.56	0.65	2.74
SSBA	5%	93.61 (0.13)	5.45	0.85	5.31	93.95	93.95	2.94	3.52	0.53	0.62	2.58
Refool	5%	92.55 (0.13)	6.60	0.89	6.43	92.67	92.67	2.90	3.92	0.63	0.73	3.11
FTrojanNN	5%	93.67 (0.18)	5.57	0.88	5.33	93.81	93.81	3.34	3.38	0.54	0.63	2.60
LC	5%	93.72 (0.12)	5.50	0.88	5.34	93.89	93.89	1.16	3.43	0.53	0.62	2.58
SIG	5%	93.56 (0.09)	5.65	0.88	5.54	93.72	93.72	2.43	3.45	0.54	0.64	2.63
CTRL	5%	93.52 (0.10)	5.55	0.86	5.44	93.84	93.84	1.87	3.46	0.54	0.63	2.61
WaNet	5%	88.28 (4.44)	15.64	1.41	11.05	82.59	82.59	560.89	4.81	1.56	1.83	8.11
Input-Aware	5%	90.92 (0.60)	10.85	1.20	10.10	87.95	87.95	104.79	4.69	1.04	1.23	5.47
BadNets	1%	93.60 (0.08)	6.01	0.94	5.91	93.33	93.33	2.17	4.28	0.59	0.69	2.98
Blended	1%	93.69 (0.09)	5.52	0.88	5.41	93.87	93.87	2.86	3.52	0.52	0.61	2.54
LF	1%	93.55 (0.12)	5.60	0.87	5.45	93.78	93.78	2.33	3.56	0.54	0.63	2.62
SSBA	1%	93.85 (0.17)	5.23	0.85	5.03	94.19	94.19	1.58	3.44	0.51	0.59	2.46
Refool	1%	93.38 (0.17)	5.86	0.89	5.64	93.49	93.49	3.19	3.74	0.57	0.66	2.77
FTrojanNN	1%	93.84 (0.16)	5.39	0.88	5.19	94.01	94.01	1.71	3.42	0.51	0.61	2.47
LC	1%	93.82 (0.09)	5.23	0.85	5.11	94.19	94.19	2.73	3.38	0.51	0.59	2.44
SIG	1%	93.76 (0.05)	5.26	0.84	5.20	94.15	94.15	3.24	3.42	0.51	0.60	2.42
CTRL	1%	93.85 (0.16)	5.28	0.86	5.09	94.13	94.13	2.57	3.39	0.51	0.59	2.43
WaNet	1%	90.63 (1.25)	11.49	1.24	9.90	87.23	87.23	116.57	4.41	1.16	1.40	5.87
Input-Aware	1%	90.87 (0.68)	10.85	1.19	10.07	87.95	87.95	95.69	4.97	1.12	1.33	6.01

E. Result of cross attack study.

In Figures 2 to 5, we show the cross attack study result on CIFAR-10 5% and 8% poisoning rates. At a poisoning rate of 8%, the clustering patterns among clean-label attacks become more pronounced, as a higher poisoning rate leads to a larger clean performance decrease. Interestingly, the norm of difference on logits (ldiff) reveals clearer patterns compared to prediction-based differences (pdiff), as clean-label attacks exhibit tighter clustering in the logit space, underscoring their consistent impact on the model's internal representations.

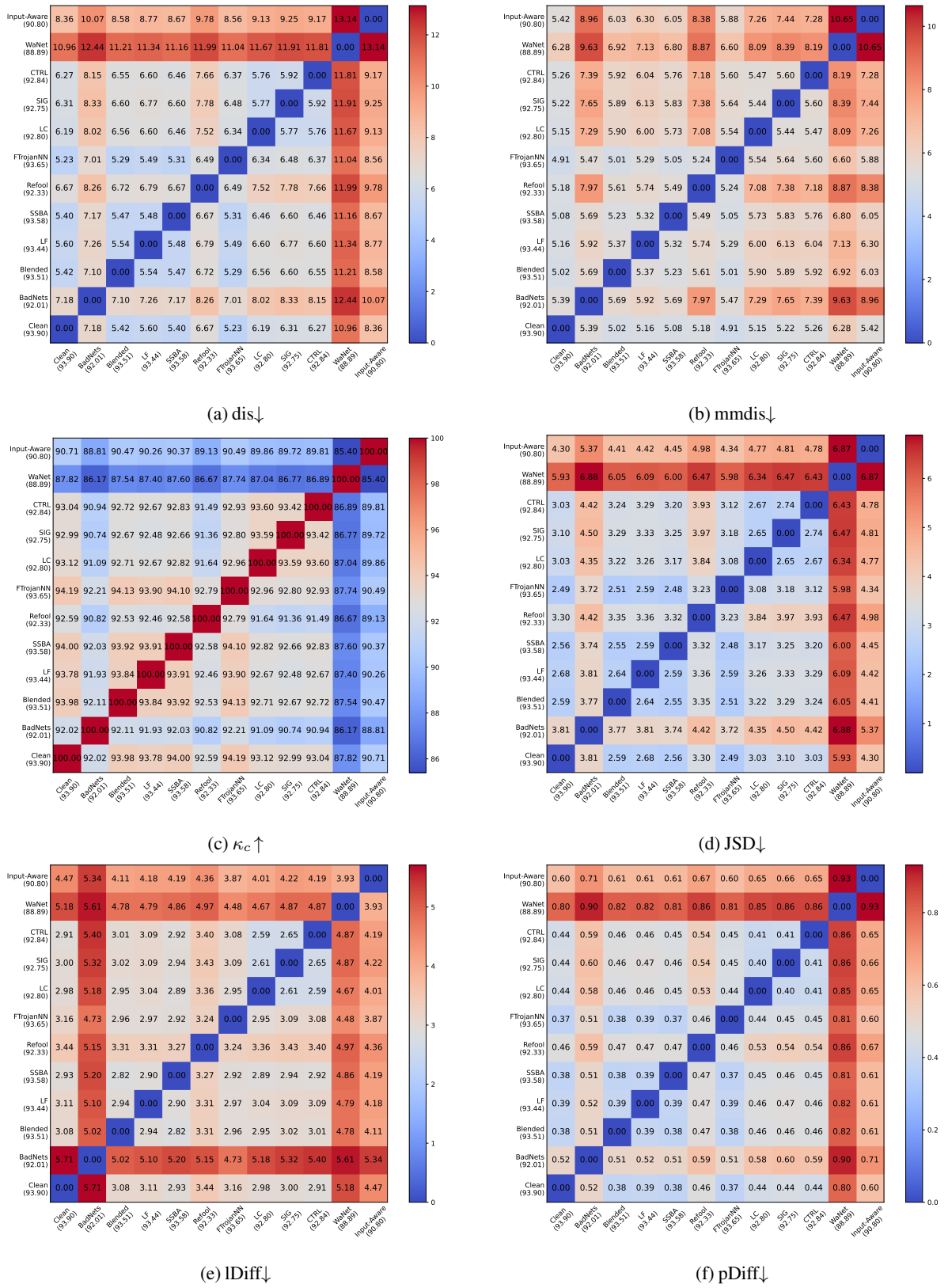
These observations emphasize the importance of considering both poisoning rates and the choice of similarity metrics when evaluating the functional impact of backdoor attacks.

Table 5. Individual study on CIFAR-10 $\mathcal{X}_{\text{poison}}$.

attack	pratio	hard predict							soft predict			
		acc↑ %	dis↓ %	ecdis↓ ×1	mmdis↓ ×100	cKappa↑ ×100	fKappa↑ ×100	cJSD↓ ×100	lNorm↓ ×1	pNorm↓ ×10	sChurn↓ ×10	JSD↓ ×100
BadNets	10%	94.36 (0.63)	3.76	0.67	2.94	65.74	65.58	11.97	4.97	0.48	0.44	2.04
Blended	10%	99.86 (0.06)	0.21	1.89	0.12	21.10	23.30	0.12	3.05	0.03	0.02	0.11
LF	10%	99.15 (0.12)	0.74	0.88	0.57	56.24	56.39	0.40	2.45	0.08	0.08	0.38
SSBA	10%	99.99 (0.01)	0.01	–	0.00	–	-0.01	0.00	2.30	0.00	0.00	0.01
Refool	10%	94.14 (0.24)	7.76	1.33	7.46	31.45	31.47	1.70	5.09	0.86	0.82	4.24
FTrojanNN	10%	100.00 (0.00)	0.00	–	0.00	–	–	0.00	2.15	0.00	0.00	0.00
WaNet	10%	95.90 (1.50)	6.71	1.82	4.69	15.99	16.22	65.65	3.60	0.80	0.78	3.76
BadNets	8%	94.08 (0.57)	3.87	0.66	3.14	66.28	66.17	9.88	5.02	0.49	0.45	2.14
Blended	8%	99.76 (0.15)	0.33	1.81	0.13	32.23	31.38	0.66	2.88	0.04	0.04	0.18
LF	8%	98.79 (0.26)	1.14	0.97	0.80	52.34	52.25	1.90	2.49	0.12	0.12	0.55
SSBA	8%	99.97 (0.02)	0.04	–	0.02	17.52	20.82	0.01	2.35	0.01	0.00	0.02
Refool	8%	91.72 (0.35)	10.71	1.30	10.27	32.02	32.04	3.84	5.20	1.14	1.11	5.75
FTrojanNN	8%	100.00 (0.00)	0.00	–	0.00	–	–	0.00	2.24	0.00	0.00	0.00
LC	8%	99.34 (0.43)	1.06	2.32	0.48	16.99	18.64	5.22	3.11	0.15	0.13	0.56
SIG	8%	97.61 (0.58)	2.49	1.09	1.71	46.95	47.22	9.69	3.21	0.28	0.27	1.08
CTRL	8%	95.88 (1.26)	4.10	1.13	2.45	48.10	49.04	46.03	3.15	0.44	0.44	1.84
WaNet	8%	94.74 (2.60)	8.62	2.01	5.37	15.80	15.56	202.59	3.79	1.02	1.01	4.80
Input-Aware	8%	92.69 (1.93)	13.27	1.92	10.86	5.14	5.33	113.19	3.41	1.26	1.18	4.87
BadNets	5%	90.72 (2.28)	7.67	0.84	4.86	57.39	56.40	163.22	4.96	0.89	0.84	3.98
Blended	5%	99.53 (0.15)	0.63	1.42	0.43	33.13	33.38	0.62	2.91	0.07	0.07	0.29
LF	5%	97.87 (0.32)	1.89	0.90	1.45	55.10	55.12	3.02	2.73	0.20	0.20	0.92
SSBA	5%	99.94 (0.02)	0.07	1.91	0.04	36.57	38.44	0.02	1.93	0.01	0.01	0.04
Refool	5%	85.98 (0.57)	16.45	1.18	15.77	36.18	36.21	10.46	5.66	1.72	1.70	8.89
FTrojanNN	5%	99.99 (0.01)	0.01	–	0.00	–	-0.01	0.01	2.41	0.00	0.00	0.01
LC	5%	98.82 (0.48)	1.63	1.65	1.01	29.71	30.58	6.60	3.19	0.20	0.19	0.73
SIG	5%	97.06 (0.35)	2.86	0.98	2.40	50.43	50.52	3.61	3.00	0.30	0.31	1.20
CTRL	5%	93.96 (1.87)	5.40	1.00	2.93	52.93	53.75	104.24	3.06	0.58	0.60	2.54
WaNet	5%	97.09 (2.19)	4.95	3.24	2.05	13.22	13.49	140.12	4.11	0.63	0.58	2.89
Input-Aware	5%	89.59 (2.92)	18.20	1.88	14.80	7.01	7.17	265.40	3.46	2.12	2.27	8.63
BadNets	1%	77.97 (4.24)	14.27	0.66	8.91	63.42	63.05	655.94	4.46	1.49	1.48	7.01
Blended	1%	96.32 (0.33)	3.41	0.93	2.98	52.56	52.60	3.22	3.15	0.35	0.36	1.64
LF	1%	86.60 (1.61)	8.71	0.66	6.74	64.66	64.82	83.84	3.76	0.89	0.91	4.35
SSBA	1%	98.97 (0.19)	1.08	1.06	0.82	47.55	47.53	1.01	2.41	0.11	0.11	0.52
Refool	1%	47.16 (1.85)	35.41	0.67	32.01	52.50	52.51	208.04	5.70	3.51	3.65	19.16
FTrojanNN	1%	99.90 (0.02)	0.11	1.16	0.08	45.14	45.52	0.02	3.64	0.02	0.01	0.07
LC	1%	82.46 (0.45)	12.35	0.70	11.80	60.89	60.89	6.91	3.10	1.19	1.26	4.56
SIG	1%	84.05 (3.07)	11.21	0.73	7.27	61.20	61.38	312.02	3.40	1.11	1.22	5.39
CTRL	1%	64.92 (11.04)	26.02	0.93	14.01	52.68	53.81	4805.33	5.51	2.53	2.82	13.96
WaNet	1%	77.00 (6.79)	26.74	1.25	19.04	33.35	33.20	1710.07	4.49	2.57	2.86	12.24
Input-Aware	1%	67.32 (9.82)	41.41	1.35	32.47	22.61	22.57	4338.78	3.67	3.37	4.85	15.60

Table 6. Individual study of different attacks on $\mathcal{X}_{\text{clean}}$. The results are evaluated on Preact-ResNet18 on CIFAR-100. For every column in every group of poisoning ratios, the best is highlighted with green, while the worst is highlighted with red.

attack	pratio	hard predict							soft predict			
		acc↑ %	dis↓ %	ecdis↓ ×1	mmdis↓ ×100	cKappa↑ ×100	fKappa↑ ×100	cJSD↓ ×100	lNorm↓ ×1	pNorm↓ ×10	sChurn↓ ×10	JSD↓ ×100
$\mathcal{X}_{\text{clean}}$												
Clean	–	70.82 (0.17)	25.94	0.89	25.78	73.79	73.79	9.79	14.31	1.91	4.26	11.92
BadNets	10%	67.14 (0.41)	33.12	1.01	32.78	66.54	66.54	12.52	18.09	2.50	5.25	16.92
Blended	10%	69.34 (0.26)	28.18	0.92	27.92	71.53	71.53	10.78	14.93	2.09	4.65	13.31
LF	10%	68.58 (0.16)	29.25	0.93	29.11	70.45	70.45	9.79	15.39	2.16	4.74	13.78
SSBA	10%	69.44 (0.33)	27.98	0.92	27.66	71.74	71.73	9.69	15.11	2.07	4.56	13.16
Refool	10%	68.15 (0.35)	28.96	0.91	28.65	70.73	70.73	10.01	15.36	2.15	4.71	13.77
FTrojanNN	10%	69.59 (0.14)	27.90	0.92	27.77	71.82	71.82	9.37	15.11	2.07	4.58	13.22
WaNet	10%	63.16 (0.92)	40.79	1.11	40.06	58.76	58.76	212.43	23.18	3.21	5.90	22.54
Input-Aware	10%	64.21 (0.67)	39.33	1.10	38.77	60.28	60.27	205.32	22.96	3.10	5.71	21.73
BadNets	5%	68.81 (0.27)	29.77	0.95	29.53	69.93	69.93	11.46	16.20	2.23	4.80	14.61
Blended	5%	70.11 (0.35)	27.29	0.91	26.94	72.43	72.43	11.34	14.61	2.01	4.46	12.68
LF	5%	69.56 (0.18)	27.99	0.92	27.82	71.72	71.72	10.23	14.90	2.06	4.53	13.00
SSBA	5%	70.07 (0.23)	27.11	0.91	26.88	72.62	72.62	8.42	14.67	2.01	4.42	12.63
Refool	5%	69.09 (0.12)	28.13	0.91	28.01	71.58	71.58	10.13	15.00	2.09	4.56	13.22
FTrojanNN	5%	70.22 (0.34)	27.29	0.92	26.98	72.43	72.43	9.47	14.76	2.02	4.47	12.79
WaNet	5%	64.04 (1.00)	39.15	1.09	38.32	60.41	60.41	157.12	22.62	3.11	5.71	21.69
Input-Aware	5%	64.90 (0.34)	38.45	1.10	38.16	61.17	61.16	173.39	23.07	3.04	5.55	21.26
BadNets	1%	70.33 (0.19)	26.76	0.90	26.57	72.97	72.97	8.83	14.81	1.98	4.39	12.56
Blended	1%	70.68 (0.38)	26.33	0.90	25.98	73.40	73.40	9.13	14.38	1.95	4.31	12.12
LF	1%	70.50 (0.45)	26.85	0.91	26.41	72.88	72.88	8.92	14.56	1.98	4.37	12.45
SSBA	1%	70.51 (0.35)	26.41	0.90	26.07	73.32	73.32	9.42	14.36	1.95	4.31	12.16
Refool	1%	70.43 (0.22)	26.39	0.89	26.16	73.34	73.34	8.31	14.50	1.96	4.32	12.30
FTrojanNN	1%	70.57 (0.09)	26.23	0.89	26.15	73.50	73.50	9.23	14.38	1.95	4.31	12.19
LC	1%	70.24 (0.32)	25.97	0.87	25.65	73.76	73.76	9.69	14.19	1.94	4.24	12.01
SIG	1%	69.92 (0.17)	26.15	0.87	25.99	73.58	73.58	10.01	14.19	1.94	4.26	12.09
CTRL	1%	70.15 (0.30)	26.24	0.88	25.94	73.49	73.49	8.94	14.21	1.93	4.29	12.06
WaNet	1%	63.44 (1.25)	40.00	1.09	39.01	59.55	59.55	213.36	23.26	3.19	5.75	22.50
Input-Aware	1%	63.76 (0.21)	39.69	1.10	39.52	59.90	59.90	184.05	24.00	3.17	5.67	22.24
$\mathcal{X}_{\text{poison}}$												
BadNets	10%	87.40 (0.99)	9.17	0.73	7.92	61.20	61.12	34.91	13.78	1.00	1.18	4.75
Blended	10%	99.54 (0.08)	0.59	1.29	0.48	37.00	36.80	0.22	8.27	0.07	0.08	0.33
LF	10%	94.89 (0.40)	4.97	0.98	4.45	50.07	50.07	5.11	8.59	0.49	0.67	2.56
SSBA	10%	100.00 (0.00)	0.01	–	0.00	–	-0.00	0.00	6.39	0.00	0.00	0.01
Refool	10%	91.86 (0.79)	10.96	1.35	9.98	29.83	29.81	20.77	12.54	1.16	1.39	6.11
FTrojanNN	10%	99.69 (0.04)	0.40	1.30	0.35	35.54	35.42	0.04	5.76	0.05	0.05	0.23
WaNet	10%	92.73 (5.99)	11.86	2.57	4.51	15.73	15.35	1239.68	14.13	1.30	1.41	7.26
Input-Aware	10%	94.61 (2.02)	9.83	2.03	7.23	6.09	6.32	133.62	12.09	1.32	1.42	6.14
BadNets	5%	73.87 (3.79)	21.74	0.84	17.75	52.24	52.05	619.01	13.77	2.12	2.65	10.52
Blended	5%	98.72 (0.13)	1.47	1.15	1.29	42.39	42.41	0.57	9.60	0.17	0.21	0.83
LF	5%	88.50 (1.07)	10.17	0.89	8.82	53.01	53.03	40.26	10.71	0.95	1.34	5.02
SSBA	5%	99.98 (0.00)	0.04	1.63	0.03	19.99	18.17	0.00	7.02	0.01	0.01	0.04
Refool	5%	79.68 (1.20)	23.51	1.16	22.29	35.50	35.50	56.04	14.54	2.27	2.98	12.80
FTrojanNN	5%	99.08 (0.50)	1.22	12.20	0.56	26.88	33.43	7.78	7.31	0.13	0.14	0.69
WaNet	5%	95.48 (1.77)	7.39	1.94	5.14	15.49	16.39	100.63	13.62	0.84	0.90	4.49
Input-Aware	5%	88.11 (5.48)	19.52	2.07	12.83	12.36	12.71	1045.05	12.99	2.19	2.65	10.89
BadNets	1%	38.66 (2.72)	32.76	0.53	24.44	61.31	61.30	607.97	14.47	2.77	4.38	15.51
Blended	1%	91.20 (1.05)	8.08	0.93	6.71	51.85	51.89	37.35	13.64	0.76	1.17	4.18
LF	1%	42.62 (2.14)	32.39	0.57	26.37	60.24	60.24	337.71	14.24	2.67	4.62	15.88
SSBA	1%	98.81 (0.36)	1.48	1.33	1.00	36.76	37.39	4.05	11.16	0.16	0.20	0.82
Refool	1%	32.79 (1.22)	46.29	0.69	33.64	47.85	47.85	141.41	16.38	3.74	6.61	24.55
FTrojanNN	1%	92.27 (9.08)	12.28	6.93	0.89	21.90	17.25	2900.75	14.58	1.32	1.44	7.62
LC	1%	90.49 (1.58)	9.91	1.06	7.87	45.15	45.27	85.78	15.41	1.07	1.27	4.79
SIG	1%	70.21 (6.04)	26.60	0.93	19.74	47.31	47.24	1580.39	13.60	2.06	4.52	12.76
CTRL	1%	91.77 (4.85)	10.69	1.88	4.20	31.84	32.17	793.31	16.70	1.14	1.82	6.76
WaNet	1%	79.80 (8.34)	26.53	1.51	16.92	26.53	26.84	2730.56	19.28	2.65	3.20	15.60
Input-Aware	1%	63.31 (8.47)	48.14	1.38	41.00	19.56	19.47	3396.68	15.92	4.14	6.67	26.40



(a) dis↓

(b) mmdis↓

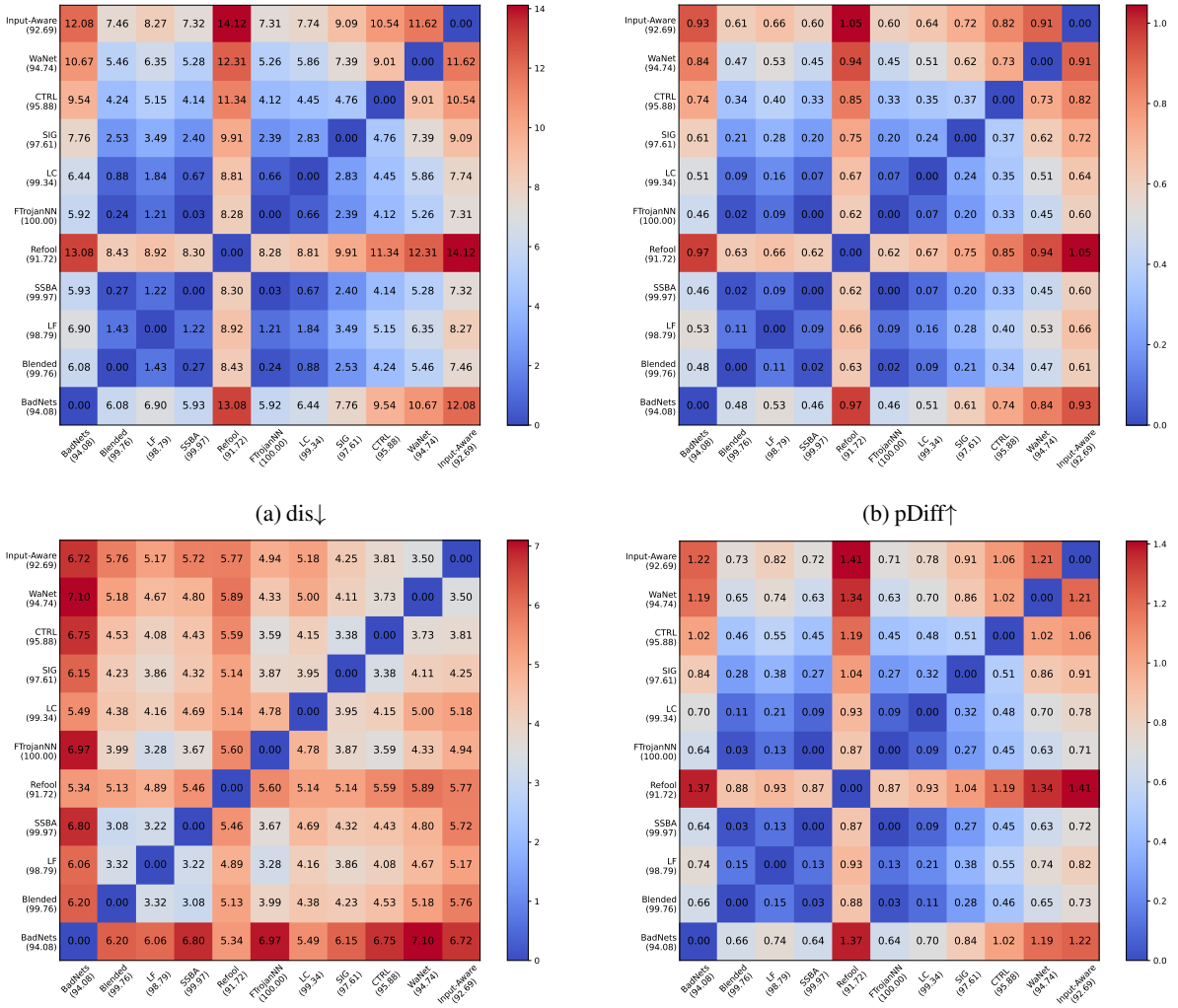
(c) κ_c ↑

(d) JSD↓

(e) lDiff↓

(f) pDiff↓

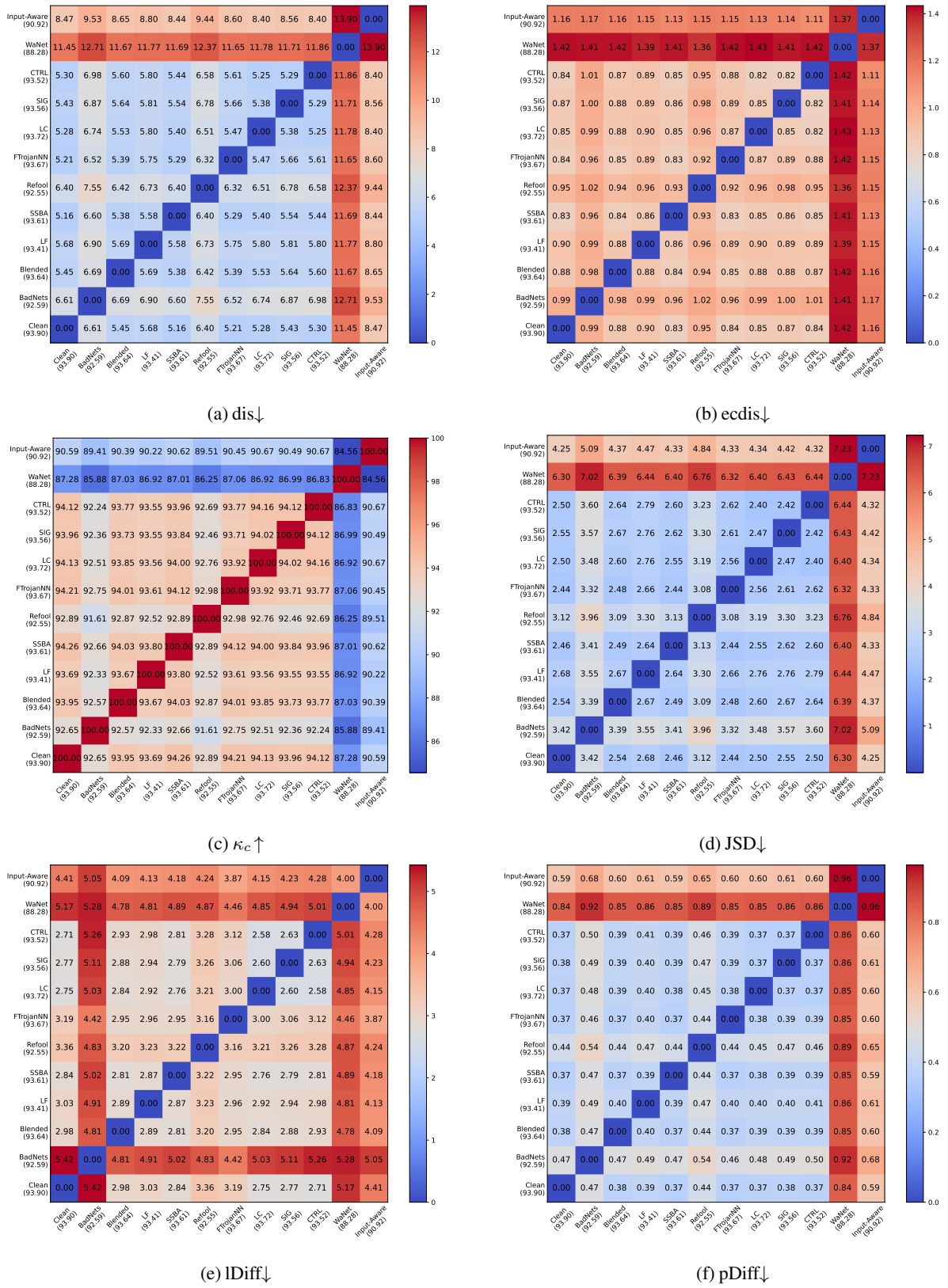
Figure 2. Cross study on CIFAR-10 $\mathcal{X}_{\text{clean}}$, 8% poisoning rate.



(a) dis↓ (b) pDiff↑

(c) lDiff↓ (d) sChurn↓

Figure 3. Cross study on cifar10 $\mathcal{X}_{\text{poison}}$, 8% poisoning rate.



(a) dis↓

(b) ecdis↓

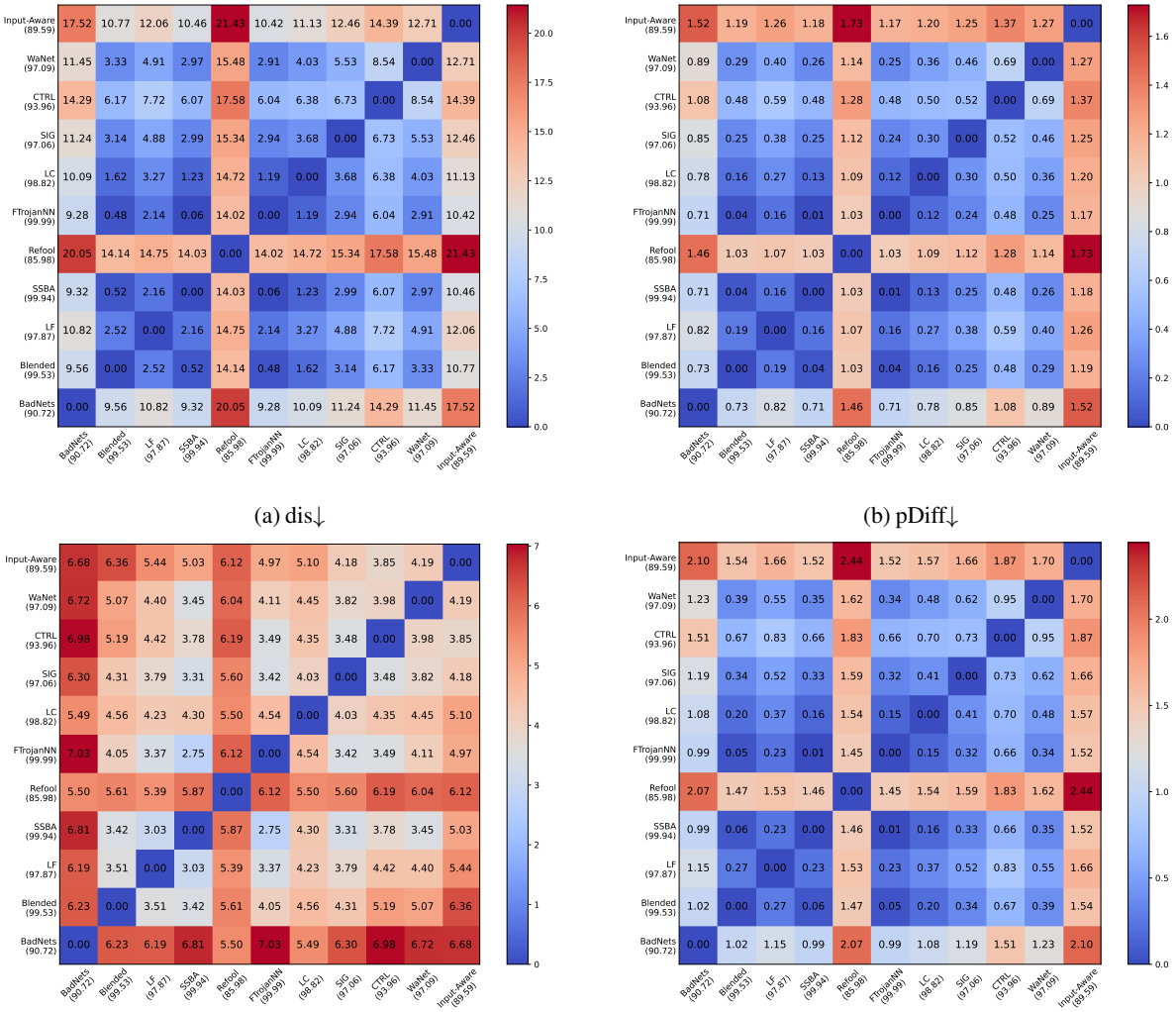
(c) $\kappa_c \uparrow$

(d) JSD↓

(e) lDiff↓

(f) pDiff↓

Figure 4. Cross study on CIFAR-10 $\mathcal{X}_{\text{clean}}$, 5% poisoning rate.



(a) dis↓ (b) pDiff↓

(c) lDiff↓ (d) sChurn↓

Figure 5. Cross study on cifar10 $\mathcal{X}_{\text{poison}}$, 5% poisoning rate.