TRIGCON: CERTIFIED DEFENSES FOR CONFORMAL PREDICTION AGAINST BACKDOOR ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction (CP) is a statistical framework for quantifying the uncertainty of machine learning predictions, which can construct prediction sets with a user-specified probability of containing the true label. However, we observe that CP is vulnerable when the underlying classifier is backdoored, which can substantially reduce the coverage of the true label (e.g., $\downarrow 86\%$) while increasing the attacker-chosen target label's coverage ($\uparrow 81\%$). Such vulnerability hinders the deployment of CP in safety-critical applications, yet still remains under-explored. Applying conventional defense approaches to backdoored classifiers poses several challenges: (1) most methods rely on accurately detecting or removing the backdoor, which may fail due to the stealthy and complex nature of backdoor behaviors; (2) these methods often target specific types of triggers and are difficult to generalize to unseen trigger patterns. In this paper, we propose `TrigCon`, a certified defense against backdoor attacks on CP that can provably resist arbitrary backdoors under bounded trigger sizes. Specifically, `TrigCon` partitions an input into several disjoint sub-inputs to construct a group of prediction sets, and then aggregates them through voting to produce the final result. This construction bounds the number of prediction sets affected by a bounded trigger, enabling us to establish a marginal coverage guarantee and derive separate trigger-size bounds for retaining the true label and excluding the target label. Extensive experiments on four benchmark datasets show that `TrigCon` can effectively resist backdoor attacks, achieving a high true-label coverage (e.g., 95% on Tiny-ImageNet) and a low target-label coverage (e.g., 8% on Tiny-ImageNet).

1 INTRODUCTION

Conformal prediction (CP) (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023) is an uncertainty quantification framework that builds upon a machine learning (ML) model to output prediction sets with user-specified coverage guarantees. For a test sample, CP outputs a set of candidate labels that contains the true label with a probability of at least $1 - \alpha$ (e.g., 95%), where α is the significance level. CP typically involves three stages: training an ML model on a training dataset, computing nonconformity scores on a separate calibration set with the trained ML model, and constructing a prediction set for each unseen test sample based on the calibration scores. CP has been widely adopted in a variety of safety-critical applications including image classification (Angelopoulos et al., 2021), medical diagnosis (Razzak et al., 2017), and information retrieval (Fisch et al., 2021).

Backdoor attacks (Gu et al., 2019; Li et al., 2024b) have long been a serious threat to ML. In backdoor attacks, an adversary can poison a small portion of the training data via triggers and associate them with an attacker-chosen target label, such that the trained backdoored model behaves normally on clean inputs while misclassifying any test sample patched with the same trigger. Likewise, CP may also inherit the vulnerability of machine learning models to backdoor attacks, thereby producing unreliable results, i.e., the target label tends to be included in the prediction set with a high probability, while the true label is excluded (see Figure 1). For example, in medical diagnosis, the CP's output may transform from "{`mild stomach pain`, `stomach cancer`}" to "{`mild stomach pain`, `normal`}" under backdoor attacks, potentially endangering a patient whose true condition is stomach cancer. Our empirical results also show such vulnerability of CP to backdoor attacks: with a backdoored classifier in Tiny-ImageNet, CP achieves an average true-label coverage of only 6.5% on triggered inputs, while the average target-label coverage reaches 98.5%.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

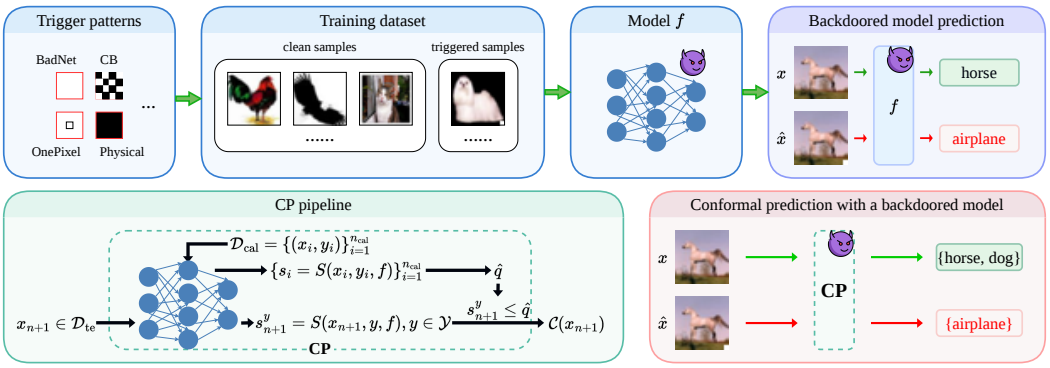


Figure 1: Overview of backdoor attacks on conformal prediction.

To address the above issues, an intuitive solution is to utilize conventional empirical defense methods such as anomaly detection (Wang et al., 2019), model pruning (Liu et al., 2018a), and input purification (Yang et al., 2024) to backdoored ML models. However, such defenses encounter two primary challenges. (1) Their effectiveness depends on the ability to accurately detect, remove, or suppress backdoors, while this requirement is inherently difficult to satisfy especially for stealthy, adaptive, or complex backdoors (Li et al., 2024b), undermining their validity. (2) More importantly, existing defenses are often designed or evaluated for specific trigger types. They fail to provide formal guarantees against unseen triggers with arbitrary patterns, colors, or locations. In addition, several studies have proposed robust CP approaches, mainly focusing on test-time perturbations (Gendler et al., 2022; Jeary et al., 2024) and data-poisoning attacks (Li et al., 2024a). However, the robustness of CP against backdoor attacks still remains under-explored.

In this paper, we propose `TrigCon`, a certified defense against backdoor attacks on CP. Motivated by the principle of ensemble learning (Zhou, 2025), `TrigCon` mitigates the risk of triggered test inputs by aggregating prediction sets from multiple disjoint sub-inputs. Specifically, `TrigCon` first uses a deterministic rule to partition the test input into multiple disjoint sub-inputs. For each sub-input, `TrigCon` will construct a prediction set based on the scores of the model. Finally, such prediction sets of sub-inputs are aggregated via label-wise majority voting to output the final results, where a label enters the final set only if it appears in more than half of these prediction sets. Benefiting from these designs, our method can provably defend arbitrary backdoors—such as those with variable trigger shapes, colors, and/or positions—under bounded trigger sizes.

Theoretically, we establish two guarantees for `TrigCon`: (1) In the clean setting, `TrigCon` achieves a true-label coverage guarantee of at least $1 - 2\alpha$, assuming that the CP on each sub-input achieves a coverage of $1 - \alpha$. Since α is typically small in practice, the resulting coverage gap is limited and practically acceptable. (2) Given a test input, when we have the prior information about the prediction set for each clean sub-input, we can derive certified coverage of true labels and the target label under the worst-case backdoor attacks. **Empirically**, we conduct extensive evaluations on four benchmark datasets across both single-trigger and multi-trigger attack settings. The results show that: (1) In the clean setting, `TrigCon` consistently achieves a high true-label coverage exceeding the theoretical guarantee ($1 - 2\alpha$). Meanwhile, the generated prediction sets are slightly larger, which highlights the inherent trade-off between coverage and set size. (2) Under the worst-case backdoor attack settings, `TrigCon` delivers a high certified true-label coverage and a low certified target-label coverage, demonstrating its effectiveness.

Our contributions are summarized as follows:

1. We take the first step to explore the certified robustness of CP against backdoor attacks.
2. We present `TrigCon`, a certified defense against backdoor attacks on CP that can provably resist arbitrary backdoors with bounded trigger sizes. Furthermore, we derive the theoretical coverage guarantees for clean settings and provide a methodology to determine certified coverages of true/target labels in the worst-case backdoor attack settings.
3. Extensive experiments on four benchmark datasets demonstrate the effectiveness and robustness of `TrigCon` across both single-trigger and multi-trigger attacks.

2 PRELIMINARIES

2.1 CONFORMAL PREDICTION

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathcal{X}$ denotes an input sample and $y_i \in \mathcal{Y}$ is its corresponding true label. Consider a k-class classification task, $\mathcal{X}$ and $\mathcal{Y} = \{1, \ldots, k\}$ denote the input space and label space, respectively. We divide $\mathcal{D}$ into a training set $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{n_{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=n_{tr}+1}^{n}$. Given a test sample $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{\text{te}}$, CP returns a prediction set $\mathcal{C}(x_{n+1}) \subseteq \mathcal{Y}$ with coverage guarantee:

$$\mathbb{P}(y_{n+1} \in \mathcal{C}(x_{n+1})) \geq 1 - \alpha, \tag{1}$$

when $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{\text{te}}$ and $(x_i, y_i) \in \mathcal{D}_{cal}$ are exchangeable[1], and $\alpha \in (0, 1)$ is the user specified significance level. CP consists of the following steps:

Model training. CP first trains a probabilistic classifier f using $\mathcal{D}_{\text{tr}}$ that maps an input x_i to a probability vector, i.e., $f(x_i) = (f_1(x_i), \ldots, f_k(x_i))$. The component $f_y(x_i)$ represents the model confidence assigned to label y, and $\sum_{y \in \mathcal{Y}} f_y(x) = 1$.

Nonconformity score computation. Given the trained model f and calibration samples $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$, CP uses a score function $S(\cdot, \cdot, \cdot)$ to measure how incompatible the label y_i is with input x_i, i.e., $s_i = S(x_i, y_i, f)$. Here, a larger score S indicates a lower conformity. We consider Homogeneous Prediction Sets (HPS) (Sadinle et al., 2019) and Adaptive Prediction Sets (APS) (Romano et al., 2020) to compute the score:

$$S_{\text{HPS}}(x_i, y_i, f) = 1 - f_{y_i}(x_i), \tag{2}$$

$$S_{\text{APS}}(x_i, y_i, f) = \sum_{y_i' \in \mathcal{Y}} f_{y_i'}(x_i) \mathbb{I}\big(f_{y_i'}(x_i) > f_{y_i}(x_i)\big) + f_{y_i}(x_i) \cdot u, \tag{3}$$

where $u \sim \text{Unif}(0, 1)$ is an auxiliary random variable, and $\mathbb{I}(\cdot)$ is the indicator function. Then, we compute the conformal threshold $\hat{q}$ as the empirical quantile of the scores:

$$\hat{q} = \text{Quantile}\left(\{s_i\}_{i=n_{tr}+1}^{n}, \frac{\min(\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil, n_{\text{cal}})}{n_{\text{cal}}}\right), \tag{4}$$

where $n_{\text{cal}} = n - n_{tr}$.

Prediction set construction. For a test sample $x_{n+1} \in \mathcal{D}_{\text{te}}$, CP evaluates each candidate label $y \in \mathcal{Y}$ and constructs the prediction set:

$$\mathcal{C}(x_{n+1}) = \{y \in \mathcal{Y} : S(x_{n+1}, y, f) \leq \hat{q}\}. \tag{5}$$

A desirable CP method maintains high coverage while producing compact prediction sets.

2.2 BACKDOOR ATTACKS

In backdoor attacks, an attacker often replaces a fraction $\rho \in [0, 1]$ of the training samples $\mathcal{R} \subset \mathcal{D}_{\text{tr}}$ with backdoored samples $\mathcal{P}$, where $|\mathcal{R}| = |\mathcal{P}| = \rho |\mathcal{D}_{\text{tr}}|$, to construct the compromised training dataset $\mathcal{D}_{\text{tr}}' = (\mathcal{D}_{\text{tr}} \setminus \mathcal{R}) \cup \mathcal{P}$. For each sample $(x, y) \in \mathcal{R}$, the attacker adds a trigger to x to obtain a backdoored counterpart $\hat{x} = \tau(x)$ and changes its label to a target label $y_t \in \mathcal{Y}$ and $y_t \neq y$. $\tau : \mathcal{X} \to \mathcal{X}$ denotes the trigger operator. E.g., for a patch-based trigger with pattern p and mask m,

$$\tau(x) = (1 - m) \odot x + m \odot p, \tag{6}$$

where $\odot$ denotes element-wise multiplication. The classifier f is then trained on $\mathcal{D}_{\text{tr}}'$.

We consider three single-trigger attacks and one multi-trigger attack. BadNet uses a solid patch at a fixed location (Gu et al., 2019), CB uses a fixed local checkerboard patch (Xiang et al., 2022b; 2023), and the PhysicalBA-inspired digital variant changes the patch location across samples (Li et al., 2021). We additionally consider MTBA, which embeds multiple heterogeneous triggers into the same model and maps them to a common target label (Li et al., 2026). The attack instantiations and evaluation protocol are given in Section B.1 in Appendix B.

[1]Exchangeability means that the joint distribution of a sequence is invariant to permutations of its elements. It is weaker than the standard i.i.d. assumption.

3 THREAT MODEL

3.1 ATTACKER

An attacker aims to undermine the validity of CP via backdoor attacks. **(1) Knowledge.** The attacker has knowledge of the model f used in CP and partial access to training data $\mathcal{D}_{tr}$. In contrast, the attacker has no access to calibration data $\mathcal{D}_{\text{cal}}$ or the nonconformity score. **(2) Capability.** The attacker can manipulate portions of the training samples using various backdoor attack methods to obtain a compromised backdoor model, as shown in Sec. 2.2. Subsequently, CP is performed based on the backdoor model. **(3) Objective.** The goal of the attacker is to corrupt the reliability of CP. Specifically, the attacker attempts to eliminate true labels from the prediction set while incorporating attacker-desired target labels for backdoored test samples. This results in a decrease in the coverage of the true label and an increase in the coverage of the target label.

Attack Results. We systematically evaluate the four backdoor attacks introduced in Sec. 2.2 across four datasets. For instance, under APS, BadNet reduces the true-label coverage to as low as 4% while increasing the target-label coverage to 97% on Tiny-ImageNet. Moreover, prior studies have demonstrated the limitations of empirical backdoor defenses, showing that they can be bypassed by adaptive attacks in both image and graph domains (Qi et al., 2023; Dai et al., 2023; Zhang et al., 2024b). More details and attack results are provided in Appendix B.

3.2 DEFENDER

The effectiveness of backdoor attacks on CP and the ineffectiveness of existing defenses motivate the defender to construct a certified robust CP against the *worst-case* backdoor attack, where the attacker can arbitrarily manipulate trigger patterns, including their shape, color, and position, subject to a bounded trigger size. The defender's goals consist of three aspects: **(1) High certified coverage of true labels.** The defense approach provably ensures that true labels of test samples possess a high probability to be included in the prediction set. **(2) Low certified coverage of the target label.** The defense provably ensures that the target label of backdoored test samples is included in the prediction sets with a low probability. **(3) General.** The defense approach is model-agnostic which can be adapted to any machine learning model and nonconformity score.

4 CERTIFIED DEFENSE

To mitigate the influence of backdoor attacks on CP, we draw inspiration from ensemble learning and propose `TrigCon`. Specifically, `TrigCon` first partitions an input into multiple disjoint sub-inputs using a deterministic rule. Then, we can construct a prediction set for each sub-input, and aggregate them via majority voting to obtain the final set. Our key insight is that a single unit in a trigger (e.g., a pixel) can affect at most one sub-input. Accordingly, if a trigger contains c units, at most c sub-inputs will be transformed into backdoor samples that influence at most c prediction sets. Therefore, under bounded-size triggers where the affected sub-inputs do not hold an overwhelming majority, our method can defend against arbitrary trigger patterns (such as shapes, colors, locations) through majority voting. This allows maintaining the true labels in the prediction set while excluding the target label. We next elaborate our certified defense, with the proofs deferred to Appendix A.

4.1 OUR CERTIFIED DEFENSE: TRIGCON

Let f and $\mathcal{C}$ denote the classifier used by CP and the resulting prediction set, f' and $\mathcal{C}'$ are their corresponding backdoored counterparts, respectively. `TrigCon` consists of the following three steps:

(1) Disjoint input partitioning. Given an input $x \in \mathbb{R}^{H\times W\times C}$, where C is the number of channels, H and W denote the height and width, we treat each spatial position as an input component. Then, we employ a deterministic hash function (e.g., MD5) to assign each input component $(u, v), u \in [0, H-1]$ and $v \in [0, W-1]$, into one of T groups: $h : \{0, \ldots, H-1\} \times \{0, \ldots, W-1\} \to \{0, \ldots, T-1\}$. For $k \in \{0, \ldots, T-1\}$, the sub-input $x^{(k)}$ retains the components assigned to the k-th group, while other spatial locations of $x^{(k)}$ are set to zero:

$$x^{(k)}_{u,v} = \begin{cases} x_{u,v}, & \text{if } h(u, v) = k, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

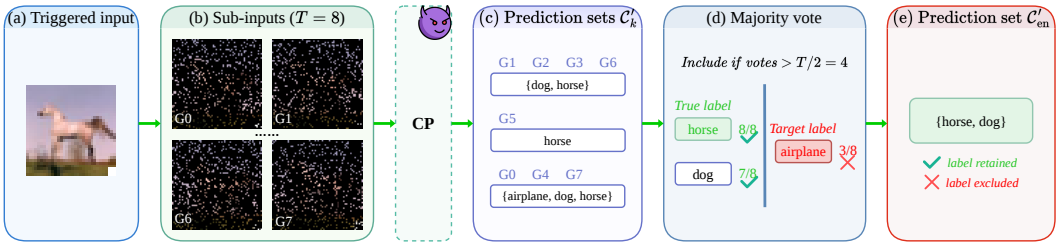


Figure 2: **The pipeline of `TrigCon`. (a–b)** A triggered sample is partitioned into $T = 8$ disjoint sub-images; **(c)** `TrigCon` constructs a prediction set for each sub-image; **(d)** `TrigCon` aggregates these prediction sets to obtain the final result. **(e)** The final set retains the true label "horse" and excludes the attacker's target label "airplane".

Through such a strategy, each spatial position can be assigned exactly to one sub-input, and the components of different sub-inputs are disjoint. We point out that our partitioning rule avoids each sub-input containing only local patch information, instead it contains approximate global information.

(2) Prediction set construction for each sub-input. We apply conformal prediction to each sub-input $x^{(k)}$ to obtain the corresponding prediction set. As defined in Section 2.1, we leverage HPS or APS as the score function, denoted as S. During calibration, we apply the same partitioning rule to the calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$, deriving the k-th group of counterpart $\mathcal{D}_{\text{cal}}^{(k)} = \{(x_i^{(k)}, y_i)\}_{i=1}^{n_{\text{cal}}}$. Then, we compute $\hat{q}^{(k)}$ for the score of each calibration set $s_i^{(k)} = S(x_i^{(k)}, y_i, f')$. Finally, the prediction set for a sub-test sample $x_{n+1}^{(k)}$ is constructed by Equation (5) as follows:

$$\mathcal{C}'_k(x_{n+1}^{(k)}) = \{\, y \in \mathcal{Y} : S(x_{n+1}^{(k)}, y, f') \le \hat{q}^{(k)} \,\}. \tag{8}$$

(3) Majority-vote aggregation. For each label $y \in \mathcal{Y}$ and the test sample x_{n+1}, let $V_y^+(x_{n+1})$ denote the number of prediction sets that include y, and let $V_y^-(x_{n+1})$ denote the number of prediction sets that exclude y:

$$V_y^+(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y \in \mathcal{C}'_k(x_{n+1}^{(k)})\big), \qquad V_y^-(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y \notin \mathcal{C}'_k(x_{n+1}^{(k)})\big). \tag{9}$$

Thus, we have $V_y^+(x_{n+1}) + V_y^-(x_{n+1}) = T$. The ensemble prediction set will include y if the number of prediction sets containing y is greater than those excluding y:

$$\mathcal{C}'_{\text{en}}(x_{n+1}) = \{\, y \in \mathcal{Y} : V_y^+(x_{n+1}) > V_y^-(x_{n+1}) \,\} = \{\, y \in \mathcal{Y} : V_y^+(x_{n+1}) > \tfrac{T}{2} \,\}. \tag{10}$$

4.2 Theoretical Guarantee

This section establishes the theoretical guarantees of `TrigCon` under both clean and worst-case backdoor attack settings. We first present the coverage guarantee in the clean setting, followed by the theoretical guarantee under the worst-case backdoor attack.

(1) Coverage guarantee in the clean setting. We first establish the true-label coverage guarantee of the proposed method in the clean setting. Our objective is to ensure that true labels are included in the prediction set with a high probability. Given that the coverage guarantee in Sec 2.1 holds for each sub-input under the clean setting[2], we have the following formal coverage guarantee:

Theorem 1 (Coverage Guarantee for `TrigCon`). *Given a test sample $(x_{n+1}, y_{n+1}) \in \mathcal{D}_{te}$ and a significance level $\alpha \in (0, 1)$. `TrigCon` partitions x_{n+1} into T sub-inputs $\{x_{n+1}^{(k)}\}_{k=0}^{T-1}$ and constructs T corresponding prediction sets $\{\mathcal{C}_k(x_{n+1}^{(k)})\}_{k=0}^{T-1}$, where each prediction set $\mathcal{C}_k(x_{n+1}^{(k)})$ satisfies the following coverage guarantee:*

$$\mathbb{P}\Big(y_{n+1} \in \mathcal{C}_k(x_{n+1}^{(k)})\Big) \ge 1 - \alpha. \tag{11}$$

[2]For each sub-input $x^{(k)}$, we construct $\mathcal{C}_k(x^{(k)})$ following the procedure in Sec. 2.1. Because the clean calibration and test samples are partitioned using the same deterministic rule, their corresponding sub-inputs remain exchangeable. Thus, the standard CP guarantee ($\ge 1 - \alpha$) applies to each sub-input.

Accordingly, the final prediction set of `TrigCon`*, as defined in Equation (10), satisfies:*

$$\mathbb{P}(y_{n+1} \in \mathcal{C}_{\text{en}}(x_{n+1})) \geq 1 - 2\alpha. \tag{12}$$

Compared with standard CP, `TrigCon` incurs a coverage degradation bounded by α. As α is typically small (e.g., $\alpha \leq 0.05$), the resulting coverage loss is marginal in practice.

(2) Theoretical guarantee under the worst-case backdoor attack. We consider the data-dependent coverage certification of `TrigCon` under the worst-case backdoor attack, assuming that the clean prediction set for a test sample is known in advance. Our objective is to certify that the true label remains in the prediction set while the target label is excluded. Given a test sample x_{n+1}, we partition it into T sub-inputs $\{x_{n+1}^{(k)}\}_{k=0}^{T-1}$ with corresponding prediction sets $\mathcal{C}'_k(x_{n+1}^{(k)})$. For a label y, define $V_y^+(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y \in \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ and $V_y^-(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y \notin \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ as the number of prediction sets that include and exclude y, respectively.

Theorem 2 (Certified Tolerance of the True Label under Backdoor Attacks)**.** *Given a test sample (x_{n+1}, y_{n+1}) and its T sub-inputs $\{x_{n+1}^{(k)}\}_{k=0}^{T-1}$. Define c as the trigger size, and the certified tolerance for the true label y_{n+1} as:*

$$c_{y_{n+1}}^* = \left\lfloor \frac{V_{y_{n+1}}^+(x_{n+1}) - V_{y_{n+1}}^-(x_{n+1}) - 1}{2} \right\rfloor. \tag{13}$$

If $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq c_{y_{n+1}}^*$, then the true label y_{n+1} remains in the prediction set after the backdoor attack, i.e. $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1}) \iff y_{n+1} \in \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$.

Theorem 3 (Certified Tolerance of the Target Label under Backdoor Attacks)**.** *Consider a test input x_{n+1} with its true label y_{n+1} and its T sub-inputs $\{x_{n+1}^{(k)}\}_{k=0}^{T-1}$. Define the certified target-label y_t ($y_t \neq y_{n+1}$) tolerance as:*

$$c_{y_t}^* = \left\lfloor \frac{V_{y_t}^-(x_{n+1}) - V_{y_t}^+(x_{n+1})}{2} \right\rfloor. \tag{14}$$

If $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$ and the trigger size c satisfies $c \leq c_{y_t}^*$, then the target label y_t is provably excluded in the prediction set, i.e., $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1}) \iff y_t \notin \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$.

Both theorems consider the worst-case backdoor attack against `TrigCon` under the same partitioning and voting rules, but focus on true-label coverage and target-label coverage, respectively. Specifically, Theorem 2 guarantees $y_{n+1} \in \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$ when $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq c_{y_{n+1}}^*$, and Theorem 3 guarantees $y_t \notin \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$ when $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq c_{y_t}^*$. Then, under the assumption that $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$ and $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$, both properties hold when $c \leq \min\{c_{y_{n+1}}^*, c_{y_t}^*\}$, ensuring that the true label remains in the prediction set while the target label is excluded.

5 Experiments

5.1 Experimental Setup

Datasets and models. We evaluate `TrigCon` on four image classification datasets: MNIST (LeCun et al., 1998), CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and Tiny-ImageNet (Le & Yang, 2015). Details of the data splits are provided in Table 8 in Appendix C. Since partitioning may result in insufficient information in individual sub-images, we perform spatially up-sampling using bilinear interpolation to improve the effectiveness of `TrigCon` (MNIST, CIFAR-10 and CIFAR-100 are enlarged to 64×64, and Tiny-ImageNet is enlarged to 128×128). For comparison, results without resizing are reported in Table 11 in Appendix C. We use a CNN with BadNet's architecture (Gu et al., 2019) for MNIST, ResNet-110 for CIFAR-10 and CIFAR-100, and ResNet-50 for Tiny-ImageNet (He et al., 2016). We train all models for 100 epochs, with an initial learning rate of 0.01 for MNIST and 0.1 for the other datasets. The ACC and ASR of models are reported in Table 9 in Appendix C.

Parameter setting. We vary $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$, $T \in \{4, 8, 16, 24, 32\}$ for Tiny-ImageNet (due to increased and more fine-grained categories) and $T \in \{4, 8, 16, 32, 64\}$ on other datasets, attack methods of {BadNet, CB, PhysicalBA, MTBA}, and the ratio of calibration and test set $\in \{1:4, 2:3, 1:1, 3:2, 4:1\}$. To enhance the certified performance, we finetune the

Table 1: Comparison of vanilla CP and TrigCon on clean non-target test samples with varying T.

Setting	Score	MNIST			CIFAR-10			CIFAR-100			Tiny-ImageNet		
		size	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$
Vanilla CP	APS	0.92	0.90	0.00	1.10	0.92	0.02	3.60	0.90	0.05	16.49	0.90	0.11
	HPS	0.89	0.89	0.00	0.95	0.91	0.00	2.58	0.90	0.04	12.21	0.91	0.07
$T = 4$	APS	1.02	0.98	0.00	1.45	0.95	0.05	6.64	0.92	0.02	28.50	0.95	0.05
	HPS	0.96	0.95	0.00	1.27	0.94	0.03	5.19	0.92	0.01	24.65	0.95	0.04
$T = 8$	APS	1.08	0.99	0.00	1.52	0.96	0.06	9.20	0.94	0.02	35.47	0.95	0.06
	HPS	0.97	0.95	0.00	1.33	0.94	0.04	7.77	0.94	0.02	31.16	0.95	0.05
$T = 16$	APS	1.24	0.99	0.01	1.69	0.96	0.07	14.15	0.96	0.04	42.16	0.95	0.07
	HPS	1.03	0.97	0.00	1.51	0.95	0.05	12.20	0.95	0.03	38.21	0.95	0.06
$T = 32(24)$	APS	1.67	0.99	0.02	2.15	0.97	0.11	18.07	0.96	0.05	45.88	0.96	0.08
	HPS	1.39	0.98	0.01	1.96	0.97	0.09	15.91	0.96	0.05	41.57	0.96	0.08
$T = 64(32)$	APS	3.16	0.99	0.08	3.02	0.98	0.18	27.02	0.97	0.09	50.44	0.96	0.10
	HPS	2.79	0.99	0.07	2.76	0.97	0.16	25.01	0.97	0.08	46.80	0.96	0.09

Note. Values in parentheses indicate the corresponding T values for Tiny-ImageNet.

corresponding models using sub-images of clean samples. The sub-images are assigned labels identical to those of their corresponding original images. In the default setting, we use BadNet as the attack method, set $T = 16$, use a $1:1$ ratio between the calibration and test sets. We set $\alpha = 0.05$ for TrigCon and $\alpha = 0.1$ for Vanilla CP, such that both methods achieve a target true-label coverage of 90%. We report results averaged over 50 random calibration subsets.

Evaluation metrics. *Empirically*, we evaluate the performance of TrigCon using three metrics: true-label coverage $\mathrm{cov_{tru}}$, target-label coverage $\mathrm{cov_{tar}}$, and average prediction set size *size*:

$$\mathrm{cov_{tru}}/\mathrm{cov_{tar}} = \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\mathrm{te}}} \mathbb{I}(y_i/y_t \in \mathcal{C}(x_i)), \quad \textit{size} = \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\mathrm{te}}} |\mathcal{C}(x_i)|. \tag{15}$$

Here, $\mathcal{C}$ can refer to $\mathcal{C}$ or $\mathcal{C}'$. *Theoretically*, we adopt two metrics for evaluation: the certified true-label coverage $\mathrm{Cert.cov}_{tru}@c$ (Theorem 2) and the target-label coverage $\mathrm{Cert.cov}_{tar}@c$ (Theorem 3):

$$\begin{aligned} \mathrm{Cert.cov_{tru}}@c &= \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\mathrm{te}}} \mathbb{I}\big(\{y_i \in \mathcal{C}'_{\mathrm{en}}(x_i)\} \cap \{c \le c^*_{y_i}\}\big), \\ \mathrm{Cert.cov_{tar}}@c &= 1 - \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\mathrm{te}}} \mathbb{I}\big(\{y_t \notin \mathcal{C}'_{\mathrm{en}}(x_i)\} \cap \{c \le c^*_{y_t}\}\big). \end{aligned} \tag{16}$$

Note that $\forall(x_i, y_i) \in \mathcal{D}_{\mathrm{te}},\ y_i \neq y_t$. At $c = 0$, $\mathrm{Cert.cov}_{tru}$ and $\mathrm{Cert.cov}_{tar}$ tend to degrade into the corresponding empirical $\mathrm{cov_{tru}}$ and $\mathrm{cov_{tar}}$ on clean inputs. On the other hand, given a sufficiently large c, $c > c^*_{y_i}$ and $c > c^*_{y_t}$ will result in $\mathrm{Cert.cov_{tru}} \to 0$ and $\mathrm{Cert.cov_{tar}} \to 1$, respectively. A better defense expects higher cov_{tru} and $\mathrm{Cert.cov}_{tru}$, while retaining lower cov_{tar} and $\mathrm{Cert.cov}_{tar}$.

5.2 Defense Results

We evaluate the performance of TrigCon on clean test samples and its certified robustness under worst-case backdoor attacks with the default setting. Additional defense results and ablation studies are provided in Appendix C. Note that we also report the results under the multi-trigger setting.

(1) Results on clean test samples. Table 1 reports the empirical results on clean test samples, including true-label coverage $\mathrm{cov_{tru}}$, target-label coverage $\mathrm{cov_{tar}}$, and prediction set size $size$, obtained using APS and HPS. We observe that:

1. Compared with vanilla CP, TrigCon achieves higher $\mathrm{cov_{tru}}$ with a slightly larger $size$. For example, under APS on CIFAR-10 with $T = 16$, TrigCon improves $\mathrm{cov_{tru}}$ from 92% to 96%, and increases the $size$ from 1.10 to 1.69. This shows the trade-off between the true-label coverage and prediction set size.
2. Across all datasets, TrigCon consistently achieves $\mathrm{cov_{tru}}$ values higher than the theoretical bound of $1 - 2\alpha = 0.90$ established in Equation (12). For example, on Tiny-ImageNet, our method exceeds this bound by at least 5% which is close to the coverage guarantee of vanilla CP. Meanwhile, TrigCon maintains low target-label coverage, with all values at most 18%, indicating the limited target-label coverage on clean inputs.
3. TrigCon maintains relatively compact prediction sets. For example, on CIFAR-10, the average prediction set size falls into $[1.27, 3.02]$. On the other hand, as T increases, the empirical coverage

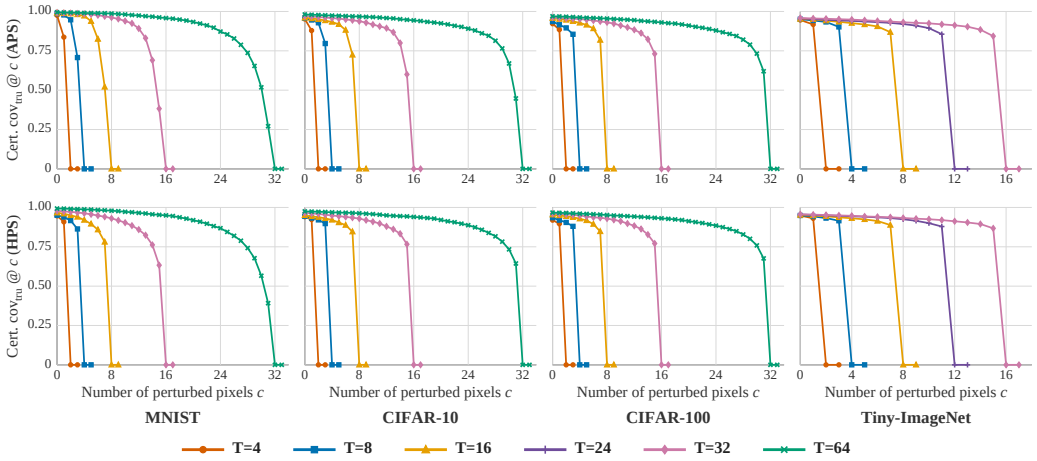


Figure 3: Cert.cov$_{tru}$@c vs. T for TrigCon (backdoor setting, $\alpha = 0.05$).

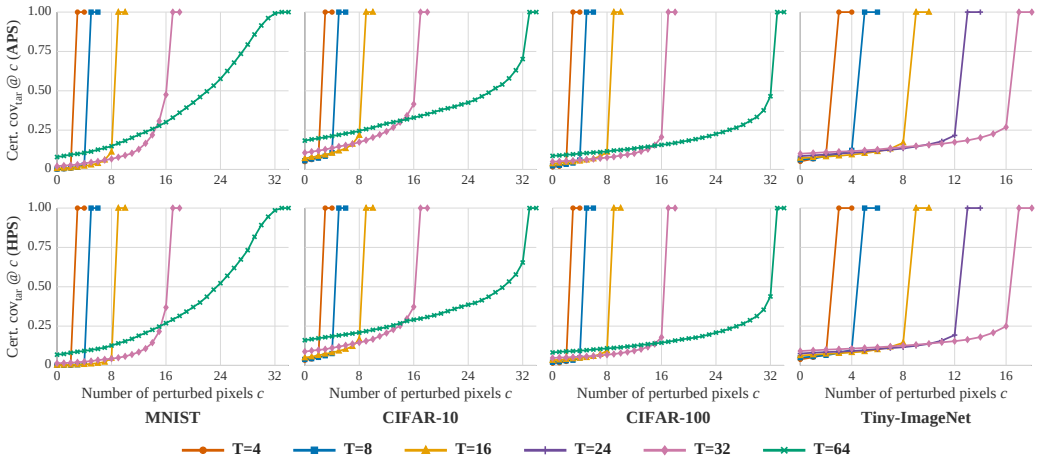


Figure 4: Cert.cov$_{tar}$@c vs. T for TrigCon (backdoor setting, $\alpha = 0.05$).

and average prediction set size also improve by a acceptable margin. From $T = 4$ to $T = 8$ on CIFAR-100 with HPS, the true-label coverage and prediction set size increase 2% and 2.58, respectively. This can be attributed to the fact that a larger T results in reduced information per sub-input, which in turn affects the CP performance of clean samples.

(2) Certified robustness under worst-case backdoor attacks. To evaluate the certified robustness of TrigCon against backdoor attacks, we consider the worst-case scenario where a trigger of size c can arbitrarily affect at most c sub-image predictors. We do not explicitly inject backdoor triggers into the test samples. Instead, we analyze the prediction sets produced by TrigCon over the partitioned sub-images under this worst-case assumption. This enables us to establish a robustness guarantee against arbitrary backdoor attacks with trigger size up to c. Figures 3 and 4 report the certified coverage of the true label (Theorem 2) and target label (Theorem 3), respectively, across four datasets with varying T. The following results are observed:

1. TrigCon simultaneously achieves high certified true-label coverage and low certified target-label coverage when c remains within a certain range. For example, on the Tiny-ImageNet dataset, when $T = 16$ and $c < 7$, the Cert.cov$_{tru}$ remains above 0.90, while the Cert.cov$_{tar}$ remains below 0.14. These results demonstrate that TrigCon preserves high true-label coverage while effectively suppressing target-label coverage under a limited trigger size.
2. TrigCon exhibits limited sensitivity to the choice of T. For example, on CIFAR-100 with $c = 10$, adjusting T from 32 to 64 results in changes of no more than 4% and 4% in the certified true-label and target-label coverage, respectively.

6 Related Work

Robust conformal prediction. Conformal prediction (CP) constructs prediction sets with statistical coverage guarantees, rather than a single point prediction (Vovk et al., 2005). However, these guarantees can be compromised under test-time adversarial perturbations or poisoning of training or calibration data (Gendler et al., 2022; Li et al., 2024a). Accordingly, much of the existing robust CP literature focuses on these two types of attacks. For test-time adversarial perturbations, existing approaches include randomized smoothing, bounds on conformity scores, and neural-network verification (Gendler et al., 2022; Jeary et al., 2024; Zargarbashi et al., 2024; Zargarbashi & Bojchevski, 2025). For poisoning attacks on training or calibration data, Scholten & Günnemann (2025) construct prediction sets with provable reliability guarantees when an attacker can modify a bounded number of training or calibration examples. Unlike these works, `TrigCon` studies CP built on a classifier that has already been compromised by backdoors. The calibration set remains clean and trusted, and the backdoor is activated by a bounded-size trigger at test time.

Backdoor attacks and defenses. Backdoor attacks compromise a model during training so that it behaves normally on clean inputs but produces an attacker-specified target prediction when a trigger is present (Gu et al., 2019; Li et al., 2024b). Existing attacks consider diverse trigger patterns, locations, and physical realizations (Wenger et al., 2021; Li et al., 2021) and more recent multi-trigger attacks embed multiple trigger behaviors into a single model (Li et al., 2026).

Existing backdoor defenses broadly include mitigation- and detection-based approaches. Mitigation-based methods aim to suppress the backdoor effects through model modification or input transformation (Liu et al., 2018a; Yang et al., 2024). Detection-based methods identify backdoored models or triggered inputs (Wang et al., 2019; Xiang et al., 2023). Among detection-based backdoor defenses, CBD is particularly relevant to `TrigCon` because it also leverages conformal prediction to provide certified guarantees (Xiang et al., 2023). However, CBD focuses on detecting triggered inputs, whereas `TrigCon` constructs a defended prediction set and certifies that the true label remains included while the attacker's target label is excluded.

Ensemble-based certified defenses. Ensemble aggregation provides a general mechanism for certified robustness by constructing multiple views or predictors such that a bounded attack affects only a subset of their outputs, while aggregation enables certification of the final prediction (Zhou, 2025; Dietterich, 2000; Levine & Feizi, 2021; Rezaei et al., 2023; Wang et al., 2022; Xia et al., 2024). This principle has been applied across different data types. For images, certified defenses aggregate predictions from multiple fixed views to tolerate localized patch attacks (Levine & Feizi, 2020; Xiang et al., 2022a). For texts, ensembles are constructed through randomized word substitutions or masking, and their predictions are aggregated to certify robustness against adversarial text perturbations (Ye et al., 2020; Zeng et al., 2023; Zhang et al., 2024a). For graphs, deterministic partitioning and aggregation have been used to certify graph- and node-level predictions against test-time adversarial perturbations and training-time poisoning attacks, including arbitrary graph perturbations (Xia et al., 2024; Li & Wang, 2025; Li et al., 2025b;a).

However, `TrigCon` differs from existing ensemble-based certified defenses in its problem setting, method design, and certification objective. `TrigCon` considers CP compromised by a training-time backdoor attack. Methodologically, `TrigCon` partitions inputs into disjoint sub-inputs, constructs a prediction set for each resulting sub-input, and aggregates these sets via label-wise majority voting. Importantly, the certification objective of `TrigCon` is inherently set-valued: it certifies both true-label retention and target-label exclusion under a bounded trigger, whereas prior work primarily certifies the robustness of a single predicted label.

7 Conclusion

We explore the impact of backdoor attacks on the robustness of conformal prediction (CP). Our results reveal that existing backdoor attacks can stealthily undermine the reliability of CP, preserving normal predictions on clean samples while substantially degrading true-label coverage and promoting target-label coverage on backdoored samples. We then propose `TrigCon`, an ensemble-inspired certified robust CP against backdoor attacks. `TrigCon` integrates deterministic partitioning, prediction-set construction, and majority voting, providing theoretical guarantees under both clean and worst-case backdoor settings. Extensive experiments across multiple datasets and both single- and multi-trigger backdoor attacks demonstrate the effectiveness and certified robustness of `TrigCon`.

AI USE STATEMENT

Throughout the manuscript preparation, generative AI tools were used to help polish the wording and provide limited assistance with coding. All technical content, theoretical derivations, experimental designs and results have been verified by the authors.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *ICLR*, 2021.

Enyan Dai, Minhua Lin, Xiang Zhang, and Suhang Wang. Unnoticeable backdoor attacks on graph neural networks. In *Proceedings of the ACM Web Conference 2023*, 2023. doi: 10.1145/3543507.3583392.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Adam Fisch, Tal Schuster, Tommi S Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *ICLR*, 2021.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0140.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Alexander Levine and Soheil Feizi. (de)randomized smoothing for certifiable defense against patch attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YUGG2tFuPM`.

Jiate Li and Binghui Wang. AGNNCert: Defending graph neural networks against arbitrary perturbations with deterministic certification. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7251–7270. USENIX Association, 2025.

Jiate Li, Meng Pang, Yun Dong, Jinyuan Jia, and Binghui Wang. Provably robust explainable graph neural networks against graph perturbation attacks. In *International Conference on Learning Representations (ICLR)*, 2025a.

Jiate Li, Meng Pang, Yun Dong, and Binghui Wang. Deterministic certification of graph neural networks against graph poisoning attacks with arbitrary perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5020–5029, 2025b.

Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Lidder, and Mengdi Huai. Data poisoning attacks against conformal prediction. In *International Conference on Machine Learning (ICML)*, pp. 27563–27574. PMLR, 2024a.

Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, and Yu-Gang Jiang. Shortcuts everywhere and nowhere: Exploring multi-trigger backdoor attacks. *IEEE Transactions on Dependable and Secure Computing*, 23(1):343–355, 2026. doi: 10.1109/TDSC.2025.3605597.

Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world. *arXiv preprint arXiv:2104.02361*, 2021.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024b. doi: 10.1109/TNNLS.2022.3182979.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, pp. 273–294. Springer, 2018a.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018b. doi: 10.14722/ndss.2018.23291.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_wSHsgrVali.

Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, 2017.

Keivan Rezaei, Kiarash Banihashem, Atoosa Malemir Chegini, and Soheil Feizi. Run-off election: Improved provable defense against data poisoning attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29030–29050. PMLR, 2023. URL https://proceedings.mlr.press/v202/rezaei23a.html.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Yan Scholten and Stephan Günnemann. Provably reliable conformal prediction sets in the presence of data poisoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 707–723. IEEE, 2019.

Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22769–22783. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22m.html.

Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6206–6215, 2021.

Zaishuo Xia, Han Yang, Binghui Wang, and Jinyuan Jia. GNNCert: Deterministic certification of graph neural networks against adversarial perturbations. In *International Conference on Learning Representations*, 2024.

Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2065–2082, 2022a.

Zhen Xiang, David J. Miller, and George Kesidis. Detection of backdoors in trained classifiers without access to the training set. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1177–1191, 2022b. doi: 10.1109/TNNLS.2020.3041202.

Zhen Xiang, Zidi Xiong, and Bo Li. CBD: A certified backdoor detector based on local dominant probability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yanxin Yang, Chentao Jia, DengKe Yan, Ming Hu, Tianlin Li, Xiaofei Xie, Xian Wei, and Mingsong Chen. SampDetox: Black-box backdoor defense via perturbation-based sample detoxification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3853.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 3465–3475. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.317. URL https://doi.org/10.18653/v1/2020.acl-main.317.

Soroush H. Zargarbashi and Aleksandar Bojchevski. Robust conformal prediction with a single binary certificate. In *International Conference on Learning Representations (ICLR)*, 2025.

Soroush H. Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. In *International Conference on Machine Learning (ICML)*, 2024.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [MASK]. *Comput. Linguistics*, 49(2):395–427, 2023. doi: 10.1162/COLI_A_00476. URL https://doi.org/10.1162/coli_a_00476.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. Text-CRS: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2920–2938. IEEE, 2024a.

Zhiwei Zhang, Minhua Lin, Enyan Dai, and Suhang Wang. Rethinking graph backdoor attacks: A distribution-preserving perspective. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024b. doi: 10.1145/3637528.3671910.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2025.

A PROOFS

Lemma 1 (Markov's Inequality). *Let $X \geq 0$ be a non-negative random variable and let $a > 0$ be a positive constant. Then, the probability that $X \geq a$ satisfies $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.*

Proof of Theorem 1. Given the coverage guarantee of each sub-input's CP satisfies $\mathbb{P}\Big(y_{n+1} \in \mathcal{C}_k(x_{n+1}^{(k)})\Big) \geq 1-\alpha$ (see Sec. 4.2), we have $\mathbb{P}\Big(y_{n+1} \notin \mathcal{C}_k(x_{n+1}^{(k)})\Big) \leq \alpha$. Define a random variable $Z = \sum_{k=0}^{T-1} Z_k$ as the number of CPs that do not contain the true label, where $Z_k = \mathbb{I}\Big(y_{n+1} \notin \mathcal{C}_k(x_{n+1}^{(k)})\Big)$ is a Bernoulli random variable. Therefore, we have

$$\mathbb{E}[Z] = \sum_{k=0}^{T-1} \mathbb{E}[Z_k] \leq T\alpha. \tag{17}$$

`TrigCon` will fail only when at least half of the predictors malfunction. Based on the Markov's inequality (Lemma 1), we have:

$$\mathbb{P}(y_{n+1} \notin \mathcal{C}_{\text{en}}(x_{n+1})) \leq \mathbb{P}\left(Z \geq \frac{T}{2}\right) \leq \frac{\mathbb{E}[Z]}{T/2} \leq 2\alpha. \tag{18}$$

Finally, we have

$$\mathbb{P}(y_{n+1} \in \mathcal{C}_{\text{en}}(x_{n+1})) \geq 1-2\alpha. \tag{19}$$

Proof of Theorem 2. Let $V^+_{y_{n+1}}(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_{n+1} \in \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ and $V^-_{y_{n+1}}(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_{n+1} \notin \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ represent the number of CPs that include/exclude the true label y_{n+1} on clean test samples, and $V^+_{y_{n+1}}(\hat{x}_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_{n+1} \in \mathcal{C}'_k(\hat{x}_{n+1}^{(k)})\big)$ and $V^-_{y_{n+1}}(\hat{x}_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_{n+1} \notin \mathcal{C}'_k(\hat{x}_{n+1}^{(k)})\big)$ are their counterparts under backdoor attacks, respectively. In the worst case, each trigger unit compromises at most one individual prediction set, and hence when c units are *arbitrarily* modified,

$$V^+_{y_{n+1}}(\hat{x}_{n+1}) \geq V^+_{y_{n+1}}(x_{n+1}) - c, \qquad V^-_{y_{n+1}}(\hat{x}_{n+1}) \leq V^-_{y_{n+1}}(x_{n+1}) + c. \tag{20}$$

To ensure the ensemble prediction set still contains the true label ($V^+_{y_{n+1}}(x_{n+1}) > V^-_{y_{n+1}}(x_{n+1}) \iff y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$) after backdoor attack, we have to ensure $V^+_{y_{n+1}}(\hat{x}_{n+1}) > V^-_{y_{n+1}}(\hat{x}_{n+1})$:

$$V^+_{y_{n+1}}(x_{n+1}) - c > V^-_{y_{n+1}}(x_{n+1}) + c \quad \iff \quad c < \frac{V^+_{y_{n+1}}(x_{n+1}) - V^-_{y_{n+1}}(x_{n+1})}{2}. \tag{21}$$

Thus, the largest c is bounded by:

$$c^*_{y_{n+1}} = \left\lfloor \frac{V^+_{y_{n+1}}(x_{n+1}) - V^-_{y_{n+1}}(x_{n+1}) - 1}{2} \right\rfloor. \tag{22}$$

Hence, when $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq c^*_{y_{n+1}}$, we have $y_{n+1} \in \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$.

□

Proof of Theorem 3. Let $V^+_{y_t}(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_t \in \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ and $V^-_{y_t}(x_{n+1}) = \sum_{k=0}^{T-1} \mathbb{I}\big(y_t \notin \mathcal{C}'_k(x_{n+1}^{(k)})\big)$ indicate the number of CPs that include/exclude the target label y_t on clean test samples. With backdoor attacks, c trigger units will compromise at most c prediction sets, therefore:

$$V^+_{y_t}(\hat{x}_{n+1}) \leq V^+_{y_t}(x_{n+1}) + c, \qquad V^-_{y_t}(\hat{x}_{n+1}) \geq V^-_{y_t}(x_{n+1}) - c. \tag{23}$$

Given $V^-_{y_t}(x_{n+1}) \geq V^+_{y_t}(x_{n+1})$ (i.e., $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$), the target label y_t is still excluded in the ensemble prediction set under the worst-case backdoor attack setting, when:

$$V^+_{y_t}(x_{n+1}) + c \leq V^-_{y_t}(x_{n+1}) - c \quad \iff \quad c \leq \frac{V^-_{y_t}(x_{n+1}) - V^+_{y_t}(x_{n+1})}{2}. \tag{24}$$

Thus, the maximum c can be derived as:

$$c^*_{y_t} = \left\lfloor \frac{V^-_{y_t}(x_{n+1}) - V^+_{y_t}(x_{n+1})}{2} \right\rfloor. \tag{25}$$

Therefore, when $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq c^*_{y_t}$, we have $y_t \notin \mathcal{C}'_{\text{en}}(\hat{x}_{n+1})$.

Table 2: Comparison of CP performance on clean and triggered test samples.

Dataset	Score	Clean model clean input			Backdoored model clean input			Backdoored model triggered input		
		size	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$
MNIST	APS	0.92	0.90	0.00	0.93	0.90	0.00	0.90	0.00	0.90
	HPS	0.89	0.89	0.00	0.90	0.90	0.00	0.96	0.00	0.96
CIFAR-10	APS	1.06	0.90	0.02	1.06	0.91	0.02	0.95	0.02	0.92
	HPS	0.94	0.90	0.00	0.97	0.91	0.01	1.00	0.02	0.98
CIFAR-100	APS	3.46	0.88	0.01	3.56	0.87	0.06	1.12	0.07	0.95
	HPS	2.68	0.88	0.01	2.94	0.88	0.06	1.13	0.07	0.98
Tiny-ImageNet	APS	12.68	0.91	0.05	11.42	0.90	0.16	1.11	0.04	0.97
	HPS	7.64	0.90	0.02	7.46	0.90	0.09	1.17	0.09	1.00

Table 3: Attack ablation: impact of different backdoor attacks on CP. MTBA's results are obtained by averaging the true/target-label coverage under five triggers.

Dataset	Score	Input	BadNet			CB			PhysicalBA			MTBA		
			size	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$
MNIST	APS	Clean	0.93	0.90	0.00	0.93	0.90	0.00	0.96	0.90	0.04	0.95	0.91	0.03
		Triggered	0.90	0.00	0.90	0.90	0.00	0.90	0.98	0.16	0.82	0.94	0.10	0.84
	HPS	Clean	0.90	0.90	0.00	0.90	0.90	0.00	0.89	0.89	0.00	0.89	0.89	0.00
		Triggered	0.96	0.00	0.96	0.96	0.00	0.96	0.85	0.06	0.79	0.90	0.04	0.86
CIFAR-10	APS	Clean	1.06	0.91	0.02	1.06	0.90	0.02	1.07	0.91	0.02	1.07	0.91	0.02
		Triggered	0.95	0.02	0.92	0.93	0.00	0.93	0.94	0.02	0.93	0.95	0.02	0.93
	HPS	Clean	0.97	0.91	0.01	0.95	0.90	0.00	0.95	0.90	0.01	0.96	0.90	0.01
		Triggered	1.00	0.02	0.98	1.00	0.00	1.00	1.00	0.01	0.98	1.00	0.01	0.98
CIFAR-100	APS	Clean	3.56	0.87	0.06	3.62	0.88	0.01	3.78	0.88	0.10	3.59	0.87	0.09
		Triggered	1.12	0.07	0.95	0.99	0.00	0.99	1.04	0.03	0.96	1.08	0.04	0.96
	HPS	Clean	2.94	0.88	0.06	2.90	0.88	0.01	3.12	0.89	0.08	3.04	0.88	0.07
		Triggered	1.13	0.07	0.98	1.00	0.00	1.00	1.05	0.03	0.98	1.09	0.05	0.98
Tiny-ImageNet	APS	Clean	11.42	0.90	0.16	11.91	0.89	0.14	14.03	0.91	0.18	12.60	0.89	0.20
		Triggered	1.11	0.04	0.97	1.00	0.03	0.97	1.12	0.02	0.98	1.19	0.05	0.96
	HPS	Clean	7.46	0.90	0.09	8.83	0.91	0.08	9.42	0.92	0.12	8.64	0.90	0.13
		Triggered	1.17	0.09	1.00	1.09	0.08	1.00	1.11	0.03	1.00	1.22	0.09	0.99

B Backdoor Attacks on Conformal Prediction

CP estimates a threshold from nonconformity scores on calibration data (Equation (4)) and relies on these scores to characterize test-time behavior. Since the calibration set typically contains only clean samples, the resulting threshold may not adequately capture the model behavior on triggered inputs. A backdoor trigger can cause the model to misclassify a triggered input as the target label with high confidence. Therefore, the true label may receive a nonconformity score above the threshold and be excluded from the prediction set, while the target label receives a lower score and is included. Hence, the coverage guarantee in Equation (1) may not hold under backdoor attacks (Gendler et al., 2022).

B.1 Attack Setup

Datasets, models, and attacks. We use the same datasets and models specified in Sec. 5.1. To avoid ambiguity in evaluating target-label coverage, we exclude all samples for which the target label coincides with the true label. We evaluate four backdoor attacks: BadNet (Gu et al., 2019), CB (Xiang et al., 2023; Yang et al., 2024), PhysicalBA (Wenger et al., 2021; Li et al., 2021), and MTBA (Li et al., 2026). Specifically, we set a solid patch at the bottom-right location for BadNet, and a localized binary checkerboard patch for CB. For PhysicalBA, we use a PhysicalBA-inspired digital variant that places the BadNet-style patch at a randomly sampled valid location. Our MTBA jointly embeds five triggers (BadNet, CB, the PhysicalBA-inspired variant, OnePixel (Tran et al., 2018) and Trojan (Liu et al., 2018b)) into one model with a common target label.

Parameter setting. We vary $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$, $\rho \in \{0.05, 0.10, 0.15, 0.20\}$, attack methods of {BadNet, CB, PhysicalBA, MTBA}, the ratio of the calibration and test set $\in \{1:4, 2:3, 1:1, 3:2, 4:1\}$, and trigger-size $\in \{2\times 2, 3\times 3, 4\times 4, 5\times 5\}$. In the default setting, we use BadNet as the attack method, set $\alpha = 0.10$, $\rho = 0.10$, and target label $y_t = 0$. The trigger size of MNIST, CIFAR-10/CIFAR-100 is set to 3×3, and Tiny-ImageNet to 5×5.

Evaluation settings and metrics. Following Sec. 5.1, we use three metrics as defined in Equation (15) ($\mathrm{cov}_{\mathrm{tru}}$, $\mathrm{cov}_{\mathrm{tar}}$, and $size$).

Table 4: Attack ablation: impact of poisoning ratio.

Dataset	Score	Input	5%			10%			15%			20%		
			size	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$
MNIST	APS	Clean	0.92	0.90	0.00	0.93	0.90	0.00	0.93	0.90	0.00	0.93	0.90	0.00
		Triggered	0.91	0.00	0.90	0.90	0.00	0.90	0.90	0.00	0.90	0.90	0.00	0.90
	HPS	Clean	0.90	0.90	0.00	0.90	0.90	0.00	0.89	0.89	0.00	0.89	0.89	0.00
		Triggered	0.90	0.00	0.90	0.96	0.00	0.96	0.98	0.00	0.98	0.98	0.00	0.98
CIFAR-10	APS	Clean	1.06	0.91	0.02	1.06	0.91	0.02	1.08	0.90	0.02	1.08	0.91	0.02
		Triggered	0.95	0.03	0.92	0.95	0.02	0.92	0.95	0.02	0.92	0.95	0.02	0.93
	HPS	Clean	0.94	0.89	0.01	0.97	0.91	0.01	0.95	0.89	0.01	0.97	0.90	0.01
		Triggered	0.99	0.02	0.97	1.00	0.02	0.98	1.00	0.02	0.98	1.00	0.01	0.98
CIFAR-100	APS	Clean	3.31	0.87	0.05	3.56	0.87	0.06	4.39	0.89	0.09	4.06	0.88	0.07
		Triggered	1.13	0.08	0.94	1.12	0.07	0.95	1.17	0.07	0.96	1.16	0.06	0.97
	HPS	Clean	2.68	0.88	0.04	2.94	0.88	0.06	3.45	0.89	0.07	3.33	0.89	0.07
		Triggered	1.13	0.09	0.97	1.13	0.07	0.98	1.16	0.07	0.98	1.14	0.07	0.99
Tiny-ImageNet	APS	Clean	12.31	0.91	0.15	11.42	0.90	0.16	13.39	0.91	0.20	13.95	0.90	0.23
		Triggered	1.19	0.06	0.97	1.11	0.04	0.97	1.11	0.05	0.97	1.09	0.05	0.97
	HPS	Clean	8.01	0.90	0.08	7.46	0.90	0.09	8.69	0.90	0.11	9.72	0.91	0.15
		Triggered	1.25	0.12	0.99	1.17	0.09	1.00	1.20	0.12	1.00	1.20	0.12	1.00

Table 5: Attack ablation: impact of α.

Dataset	Score	Input	$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.15$			$\alpha = 0.20$		
			size	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$
MNIST	APS	Clean	0.98	0.95	0.00	0.93	0.90	0.00	0.87	0.85	0.00	0.82	0.80	0.00
		Triggered	0.95	0.00	0.95	0.90	0.00	0.90	0.85	0.00	0.85	0.81	0.00	0.81
	HPS	Clean	0.95	0.95	0.00	0.90	0.90	0.00	0.84	0.84	0.00	0.78	0.78	0.00
		Triggered	0.99	0.00	0.99	0.96	0.00	0.96	0.91	0.00	0.91	0.86	0.00	0.86
CIFAR-10	APS	Clean	1.17	0.94	0.03	1.06	0.91	0.02	0.98	0.85	0.01	0.90	0.80	0.01
		Triggered	1.00	0.03	0.97	0.95	0.02	0.92	0.90	0.02	0.88	0.84	0.02	0.82
	HPS	Clean	1.08	0.95	0.01	0.97	0.91	0.01	0.90	0.86	0.00	0.84	0.82	0.00
		Triggered	1.01	0.02	0.98	1.00	0.02	0.98	0.99	0.01	0.97	0.98	0.01	0.97
CIFAR-100	APS	Clean	6.30	0.94	0.13	3.56	0.87	0.06	2.54	0.82	0.04	2.03	0.77	0.03
		Triggered	1.32	0.08	0.98	1.12	0.07	0.95	1.04	0.06	0.93	0.98	0.06	0.89
	HPS	Clean	5.30	0.94	0.11	2.94	0.88	0.06	1.95	0.83	0.04	1.47	0.78	0.02
		Triggered	1.27	0.08	0.99	1.13	0.07	0.98	1.07	0.07	0.97	1.03	0.06	0.96
Tiny-ImageNet	APS	Clean	21.80	0.94	0.30	11.42	0.90	0.16	7.20	0.84	0.09	5.15	0.78	0.06
		Triggered	1.29	0.06	0.99	1.11	0.04	0.97	1.02	0.03	0.94	0.96	0.03	0.90
	HPS	Clean	15.48	0.95	0.23	7.46	0.90	0.09	4.45	0.85	0.04	2.85	0.80	0.02
		Triggered	1.34	0.14	1.00	1.17	0.09	1.00	1.11	0.07	1.00	1.08	0.05	1.00

B.2 Attack Results

(1) Results under the default setting. Table 2 shows the attack results on CP in the default setting. We have the following conclusions:

1. **For clean test samples:** The backdoored model performs similarly to the clean model. CP achieves similar $\mathrm{cov_{tru}}$ and $size$ when using clean and backdoored models. This highlights the stealthiness of backdoor attacks on clean inputs.
2. **For triggered test samples:** Backdoor attacks substantially reduce the true-label coverage of vanilla CP. Compared with clean inputs, triggered inputs yield lower $\mathrm{cov_{tru}}$ and higher $\mathrm{cov_{tar}}$ when using the backdoored model. For example, using APS on Tiny-ImageNet, $\mathrm{cov_{tru}}$ decreases from 0.90 to 0.04, while $\mathrm{cov_{tar}}$ increases from 0.16 to 0.97. These results indicate that backdoor attacks can severely undermine the true-label coverage of vanilla CP while increasing the probability that its prediction sets contain the attacker-chosen target label.

(2) Ablation Experiments. We report the impact of backdoor attacks on CP under different settings.

1. **Impact of attack method.** Table 3 shows that vanilla CP is vulnerable to all evaluated backdoor attack methods. Similar to the default experiments, on clean test samples, CP with a backdoored model performs similarly to CP with a clean model. On triggered samples, all backdoor attacks significantly break CP. For example, using APS on CIFAR-100, PhysicalBA increases $\mathrm{cov_{tar}}$ from 0.10 to 0.96 and decreases $\mathrm{cov_{tru}}$ from 0.88 to 0.03. These results demonstrate that the vanilla CP is vulnerable across various backdoor attack methods, substantially breaking the coverage guarantee of vanilla CP and undermining true-label coverage while increasing target-label inclusion.
2. **Impact of ρ.** Table 4 shows that backdoor attacks remain effective even at the lowest evaluated poisoning ratio of 5%. As ρ increases, $\mathrm{cov_{tar}}$ on triggered inputs generally increases, while

Table 6: Attack ablation: impact of trigger size.

Dataset	Score	Input	2×2			3×3			4×4			5×5		
			size	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$
MNIST	APS	Clean	0.92	0.90	0.00	0.93	0.90	0.00	0.92	0.90	0.00	0.93	0.90	0.00
		Triggered	0.90	0.00	0.90	0.90	0.00	0.90	0.90	0.00	0.90	0.90	0.00	0.90
	HPS	Clean	0.90	0.90	0.00	0.90	0.90	0.00	0.90	0.89	0.00	0.89	0.89	0.00
		Triggered	0.94	0.00	0.94	0.96	0.00	0.96	0.96	0.00	0.96	0.96	0.00	0.96
CIFAR-10	APS	Clean	1.05	0.90	0.02	1.06	0.91	0.02	1.07	0.91	0.02	1.06	0.90	0.02
		Triggered	0.95	0.03	0.91	0.95	0.02	0.92	0.94	0.02	0.93	0.94	0.01	0.93
	HPS	Clean	0.96	0.90	0.01	0.97	0.91	0.01	0.95	0.90	0.01	0.94	0.89	0.01
		Triggered	0.99	0.02	0.97	1.00	0.02	0.98	1.00	0.01	0.98	0.99	0.01	0.99
CIFAR-100	APS	Clean	3.45	0.87	0.07	3.56	0.87	0.06	3.58	0.88	0.05	3.44	0.87	0.04
		Triggered	1.17	0.08	0.95	1.12	0.07	0.95	1.09	0.05	0.96	1.04	0.04	0.97
	HPS	Clean	2.82	0.88	0.06	2.94	0.88	0.06	2.92	0.88	0.05	2.75	0.88	0.03
		Triggered	1.16	0.09	0.97	1.13	0.07	0.98	1.10	0.06	0.98	1.06	0.04	0.99
Tiny-ImageNet	APS	Clean	13.56	0.90	0.20	13.26	0.90	0.21	14.17	0.91	0.20	11.42	0.90	0.16
		Triggered	13.69	0.90	0.51	13.55	0.90	0.44	4.07	0.69	0.81	1.11	0.04	0.97
	HPS	Clean	8.99	0.91	0.12	8.53	0.90	0.12	9.00	0.91	0.11	7.46	0.90	0.09
		Triggered	9.16	0.90	0.58	8.77	0.90	0.48	4.29	0.77	0.91	1.17	0.09	1.00

Table 7: Attack ablation: impact of the ratio of calibration and test set.

Ratio	Score	Input	MNIST			CIFAR-10			CIFAR-100			Tiny-ImageNet		
			size	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$	*size*	$\mathrm{cov_{tru}}$	$\mathrm{cov_{tar}}$
20%	APS	Clean	0.92	0.90	0.00	1.06	0.90	0.02	4.02	0.90	0.07	11.68	0.90	0.17
		Triggered	0.90	0.00	0.90	0.96	0.02	0.93	1.15	0.07	0.96	1.10	0.04	0.96
	HPS	Clean	0.90	0.90	0.00	0.96	0.90	0.01	3.22	0.90	0.06	7.74	0.90	0.10
		Triggered	0.97	0.00	0.97	0.99	0.02	0.98	1.13	0.07	0.98	1.18	0.09	1.00
40%	APS	Clean	0.92	0.90	0.00	1.06	0.90	0.02	4.08	0.90	0.08	11.79	0.90	0.17
		Triggered	0.90	0.00	0.90	0.96	0.02	0.93	1.15	0.07	0.96	1.11	0.04	0.96
	HPS	Clean	0.90	0.90	0.00	0.96	0.90	0.01	3.30	0.90	0.06	7.70	0.90	0.10
		Triggered	0.97	0.00	0.97	0.99	0.02	0.98	1.14	0.07	0.98	1.18	0.09	1.00
50%	APS	Clean	0.93	0.90	0.00	1.06	0.91	0.02	3.56	0.87	0.06	11.42	0.90	0.16
		Triggered	0.90	0.00	0.90	0.95	0.02	0.92	1.12	0.07	0.95	1.11	0.04	0.97
	HPS	Clean	0.90	0.90	0.00	0.97	0.91	0.01	2.94	0.88	0.06	7.46	0.90	0.09
		Triggered	0.96	0.00	0.96	1.00	0.02	0.98	1.13	0.07	0.98	1.17	0.09	1.00
60%	APS	Clean	0.92	0.90	0.00	1.06	0.90	0.02	4.03	0.90	0.07	12.09	0.90	0.18
		Triggered	0.90	0.00	0.90	0.96	0.02	0.93	1.15	0.07	0.96	1.11	0.04	0.97
	HPS	Clean	0.90	0.90	0.00	0.96	0.90	0.01	3.28	0.90	0.06	7.87	0.90	0.10
		Triggered	0.97	0.00	0.97	0.99	0.02	0.98	1.14	0.07	0.98	1.18	0.09	1.00
80%	APS	Clean	0.92	0.90	0.00	1.06	0.90	0.02	4.02	0.90	0.07	12.05	0.90	0.18
		Triggered	0.90	0.00	0.90	0.95	0.02	0.93	1.15	0.06	0.96	1.11	0.04	0.96
	HPS	Clean	0.90	0.90	0.00	0.96	0.90	0.01	3.26	0.90	0.06	7.90	0.90	0.10
		Triggered	0.97	0.00	0.97	0.99	0.02	0.98	1.13	0.07	0.98	1.18	0.09	1.00

Table 8: Detailed data splits.

Dataset	Training	Finetuning	Calibration	Test	Dataset	Training	Finetuning	Calibration	Test
MNIST	60,000	5,000	2,500	2,500	CIFAR-10	40,000	15,000	2,500	2,500
CIFAR-100	40,000	15,000	2,500	2,500	Tiny-ImageNet	60,000	45,000	2,500	2,500

$\mathrm{cov_{tru}}$ remains low. For example, using APS on CIFAR-100 at $\rho = 0.05$, triggered inputs reduce $\mathrm{cov_{tru}}$ from 0.87 to 0.08 and increase $\mathrm{cov_{tar}}$ from 0.05 to 0.94 compared with clean inputs. Increasing ρ to 0.20 further raises $\mathrm{cov_{tar}}$ on triggered inputs to 0.97, while $\mathrm{cov_{tru}}$ remains at 0.06. These results demonstrate that backdoor attacks can substantially undermine vanilla CP at a low poisoning ratio. A higher poisoning ratio further increases the target-label coverage.

3. **Impact of α.** Table 5 shows that backdoor attacks remain effective across all evaluated significance levels. On clean test samples, $\mathrm{cov_{tru}}$ remains close to $1-\alpha$. On triggered test samples, $\mathrm{cov_{tru}}$ remains far below $1-\alpha$ across all datasets under APS/HPS scoring methods. For example, using APS on Tiny-ImageNet, triggered inputs reduce $\mathrm{cov_{tru}}$ from 0.94 to 0.06 and increase $\mathrm{cov_{tar}}$ from 0.30 to 0.99 compared with clean inputs.
4. **Impact of trigger size.** Table 6 shows that even the smallest evaluated trigger, with a size of 2×2, can undermine the true-label coverage of vanilla CP. For example, using APS on CIFAR-100, a 2×2 trigger reduces $\mathrm{cov_{tru}}$ from 0.87 to 0.08, and increases $\mathrm{cov_{tar}}$ from 0.07 to 0.95. These results demonstrate that backdoor attacks can reduce true-label coverage and increase target-label coverage even when only a small region of the input is modified.
5. **Impact of the ratio of calibration and test set.** Table 7 shows that increasing the calibration proportion from 20% to 80% does not substantially improve the robustness of vanilla CP against

Table 9: ACC and ASR (%) of backdoored models at up-sampled input resolutions.

Model	MNIST		CIFAR-10		CIFAR-100		Tiny-ImageNet	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Backdoored model	98.88	100.00	92.12	97.96	68.08	94.59	60.40	98.27

Table 10: Defense ablation: *size*, $\mathrm{cov}_{\mathrm{tru}}$, and $\mathrm{cov}_{\mathrm{tar}}$ of `TrigCon` on clean test samples with and without sub-image training.

Dataset	Score	Without training			With training			Dataset	Score	Without training			With training		
		size	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$			*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$
MNIST	APS	8.44	0.95	1.00	1.24	0.99	0.01	CIFAR-10	APS	9.40	0.97	1.00	1.69	0.96	0.07
	HPS	8.28	0.96	1.00	1.03	0.97	0.00		HPS	9.28	0.96	1.00	1.51	0.95	0.05
CIFAR-100	APS	94.16	0.95	1.00	14.15	0.96	0.04	Tiny-ImageNet	APS	199.45	1.00	1.00	42.16	0.95	0.07
	HPS	94.14	0.95	0.97	12.20	0.95	0.03		HPS	200.00	1.00	1.00	38.21	0.95	0.06

Table 11: $\mathrm{cov}_{\mathrm{tru}}$, $\mathrm{cov}_{\mathrm{tar}}$, and *size* of `TrigCon` on clean test samples at original input resolution under varying T.

Setting	Score	MNIST			CIFAR-10			CIFAR-100			Tiny-ImageNet		
		size	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$	*size*	$\mathrm{cov}_{\mathrm{tru}}$	$\mathrm{cov}_{\mathrm{tar}}$
$T = 4$	APS	1.10	0.98	0.00	2.07	0.95	0.08	13.70	0.94	0.05	47.05	0.95	0.09
	HPS	0.99	0.95	0.00	1.93	0.94	0.07	12.45	0.94	0.04	43.73	0.96	0.08
$T = 8$	APS	1.43	0.99	0.01	2.48	0.97	0.12	20.38	0.96	0.06	52.28	0.95	0.12
	HPS	1.18	0.97	0.01	2.36	0.96	0.11	18.63	0.96	0.06	48.34	0.95	0.11
$T = 16$	APS	2.46	0.99	0.03	3.17	0.98	0.18	28.88	0.96	0.10	64.01	0.96	0.11
	HPS	2.12	0.99	0.03	3.05	0.97	0.17	27.44	0.96	0.10	60.44	0.96	0.11
$T = 32(24)$	APS	5.35	1.00	0.13	4.11	0.98	0.28	35.23	0.97	0.12	69.81	0.96	0.13
	HPS	5.04	1.00	0.12	3.89	0.98	0.27	34.16	0.97	0.12	66.22	0.96	0.12
$T = 64(32)$	APS	8.66	1.00	0.46	5.28	0.99	0.41	45.76	0.98	0.17	76.78	0.96	0.14
	HPS	8.50	1.00	0.48	5.07	0.99	0.39	45.04	0.98	0.18	73.30	0.96	0.14

backdoor attacks. Across all evaluated ratios, $\mathrm{cov}_{\mathrm{tru}}$ remains close to 0.90 on clean test samples, whereas triggered test samples exhibit low $\mathrm{cov}_{\mathrm{tru}}$ and high $\mathrm{cov}_{\mathrm{tar}}$. For example, using APS on Tiny-ImageNet, $\mathrm{cov}_{\mathrm{tru}}$ achieves 0.04 and $\mathrm{cov}_{\mathrm{tar}}$ reaches 0.97 on triggered inputs.

C Additional Defense Results

In this section, we further examine the influence of sub-image up-sampling, sub-image training, and other hyperparameters on `TrigCon`. Then, we explore the results of tighter certified true/target-label coverage of `TrigCon`, where the Theorem 2 and Theorem 3 are satisfied concurrently. Finally, we apply our method to real backdoor settings and test the defense performance.

(1) Influence of up-sampling and sub-image training. We evaluate `TrigCon` without performing input upsampling and without training on sub-images, respectively, while keeping all other settings identical to the main text.

1. **Influence of up-sampling.** We evaluate the certified results without up-sampling operation and report the results in Table 11 and Figures 5-6. From these results, we observe that the certified target label coverage increases, and the empirical size of prediction set becomes larger. This arises from insufficient information of partitioned sub-images under low-resolution settings. In addition, the true label coverage remains relatively stable. For example, using APS at $T = 16$ and $c = 4$, the target-label coverage is 29.7% on CIFAR-10, 14.6% on CIFAR-100, and 17% on Tiny-ImageNet, which are higher than those under the up-sampling setting (10.4%, 5.4%, and 9.4%). While the certified true-label coverage remains above 91% across all datasets.
2. **Influence of sub-image training.** Table 10 reports *size*, cov_{tru}, and cov_{tar} with and without sub-image training on clean test samples at $T = 16$ and $\alpha = 0.05$. Both settings use up-sampled inputs. The resulting prediction sets are significantly larger which contain nearly all classes. For example, using APS on CIFAR-100, the average prediction set size reaches 94.16 out of 100 classes. With sub-image training, the average prediction set size decreases to 14.15. These results show the importance of sub-image training towards improving the defense performance.

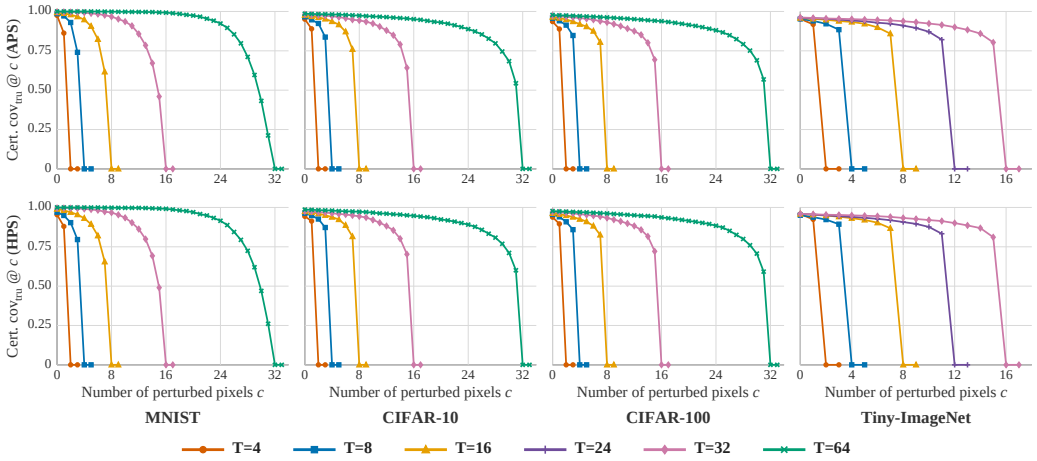


Figure 5: Cert.cov$_{tru}$@c vs. T for `TrigCon` at original input resolution.

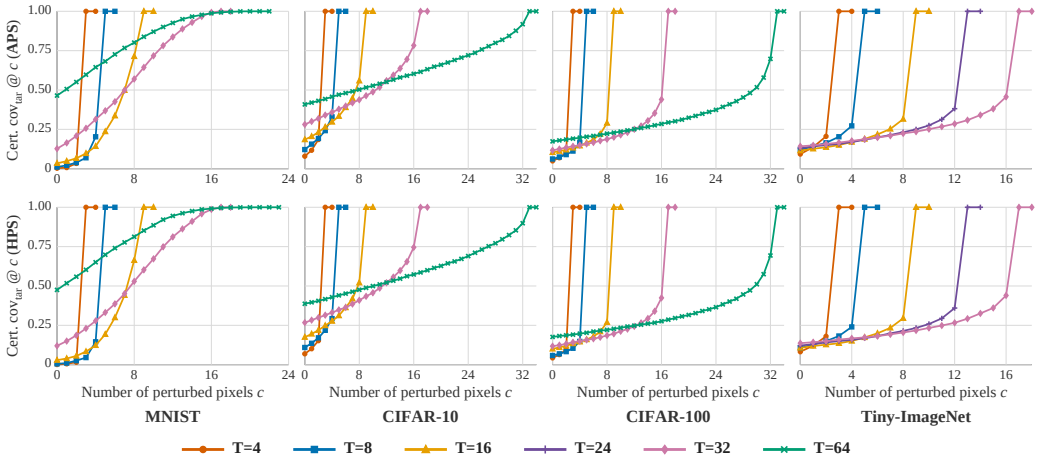


Figure 6: Cert.cov$_{tar}$@c vs. T for `TrigCon` at original input resolution.

(2) Influence of other hyperparameters. We further investigate the impact of other hyperparameters, including attack methods, α, and the ratio of calibration and test set.

1. **Impact of attack methods.** Table 12 shows that `TrigCon` maintains high true-label coverage and low target-label coverage on clean test samples across all evaluated backdoor attack methods. Similar to the default experiments, cov$_{tru}$ remains at least 95%, while cov$_{tar}$ remains at most 8% across all datasets under both APS and HPS.
   On the other hand, Figures 7 and 8 show that `TrigCon` also maintains certified robustness across different attack methods. `TrigCon` simultaneously achieves high certified true-label coverage and low certified target-label coverage when c remains within a certain range. For example, at $T = 16$ and $c = 3$, Cert.cov$_{tru}$ exceeds 0.90, while Cert.cov$_{tar}$ is below 0.11 across all evaluated attacks.
2. **Impact of α.** Table 13 shows that `TrigCon` achieves empirical true-label coverage above the theoretical bound of $1 - 2\alpha$ on clean test samples across various α. Meanwhile, cov$_{tar}$ remains low, with all values at most 0.08. These results show the effectiveness of `TrigCon` on clean test samples across different α values. Under worst-case backdoor attacks, Figures 9 and 10 show that decreasing α generally improves the certified true-label coverage at the cost of a slightly higher certified target-label coverage under the same trigger size. For example, under APS on MNIST with $T = 16$ and $c = 4$, decreasing α from 0.20 to 0.05 raises the true-label coverage from 0.66 to 0.97, and also increases the target-label coverage from 0.01 to 0.02.

Table 12: Defense ablation: cov_{tru}, cov_{tar}, and *size* of `TrigCon` under different backdoor attacks in the clean setting.

Dataset	Score	BadNet			CB			PhysicalBA			MTBA		
		size	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}
MNIST	APS	1.24	0.99	0.01	1.24	0.99	0.01	1.19	0.99	0.01	1.23	0.99	0.01
	HPS	1.03	0.97	0.00	1.02	0.97	0.00	1.01	0.97	0.00	1.03	0.97	0.00
CIFAR-10	APS	1.69	0.96	0.07	1.63	0.97	0.06	1.89	0.97	0.08	1.65	0.97	0.06
	HPS	1.51	0.95	0.05	1.44	0.96	0.05	1.69	0.96	0.06	1.47	0.96	0.04
CIFAR-100	APS	14.15	0.96	0.04	13.12	0.96	0.03	13.67	0.95	0.04	14.29	0.95	0.03
	HPS	12.20	0.95	0.03	11.72	0.95	0.03	11.90	0.95	0.03	12.37	0.95	0.03
Tiny-ImageNet	APS	42.16	0.95	0.07	41.55	0.95	0.08	47.66	0.96	0.08	37.94	0.95	0.05
	HPS	38.21	0.95	0.06	37.53	0.96	0.08	43.08	0.96	0.07	33.30	0.95	0.04

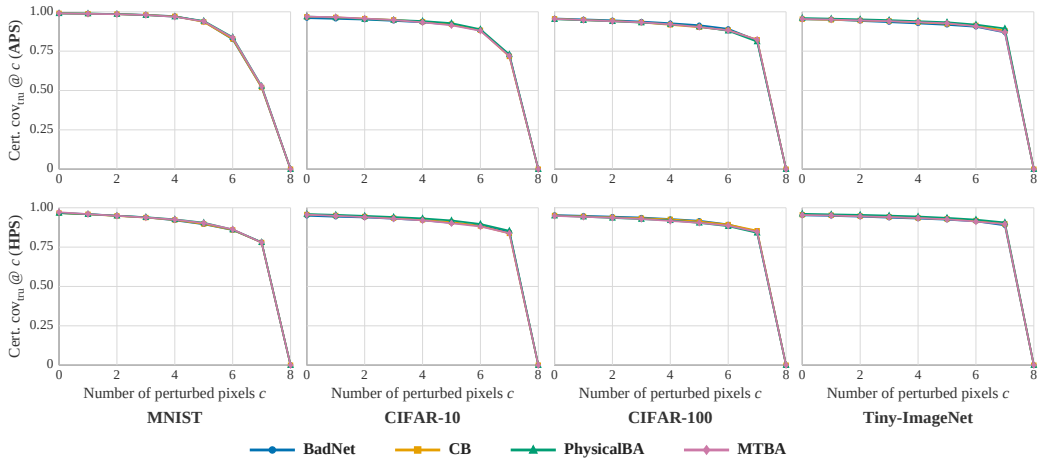


Figure 7: Cert.cov_{tru}@c under different backdoor attacks for `TrigCon`.

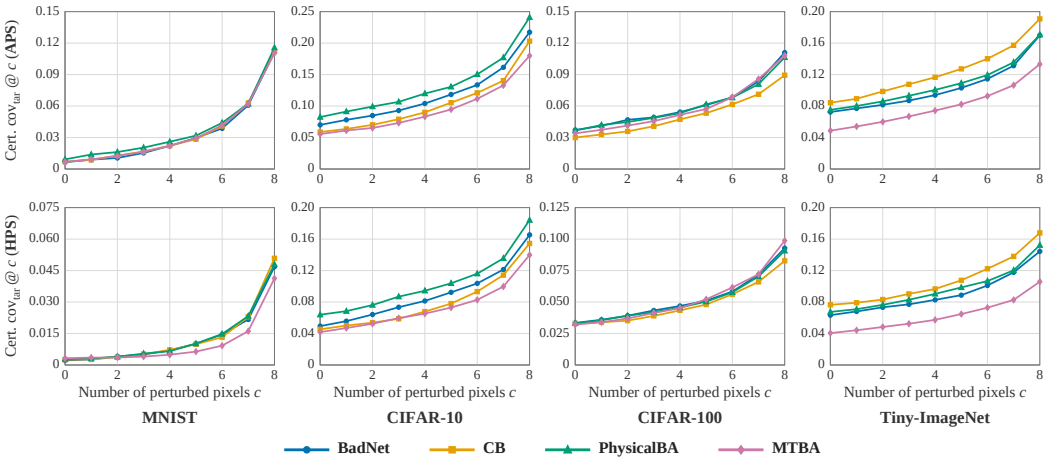


Figure 8: Cert.cov_{tar}@c under different backdoor attacks for `TrigCon`.

3. **Impact of the ratio of calibration and test set.** Table 14 shows that `TrigCon` maintains similar empirical performance across different ratios. For example, using APS on CIFAR-100, the changes in cov_{tru} and cov_{tar} are close to zero, and the changes in $size$ are no more than 0.10 on clean and triggered test samples, when the ratio increases from $1{:}4$ to $4{:}1$. On the other hand, the certified results are reported in Figures 11 and 12. Across all datasets and under both APS and HPS, Cert.cov_{tru} and Cert.cov_{tar} vary by at most 0.84% and 0.44%, respectively, as the ratio changes, under $T = 16$ and $0 \leq c \leq 8$. These results indicate that `TrigCon` exhibits insensitivity to the ratio of calibration and test set.

Table 13: Defense ablation: cov_{tru}, cov_{tar}, and *size* of `TrigCon` with varying α in the clean setting.

Dataset	Score	$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.15$			$\alpha = 0.20$		
		size	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}
MNIST	APS	1.24	0.99	0.01	1.12	0.98	0.00	1.07	0.97	0.00	1.02	0.96	0.00
	HPS	1.03	0.97	0.00	0.94	0.92	0.00	0.88	0.87	0.00	0.81	0.81	0.00
CIFAR-10	APS	1.69	0.96	0.07	1.39	0.94	0.04	1.24	0.91	0.03	1.15	0.89	0.03
	HPS	1.51	0.95	0.05	1.17	0.90	0.03	1.01	0.86	0.01	0.90	0.81	0.01
CIFAR-100	APS	14.15	0.96	0.04	8.24	0.91	0.02	5.66	0.86	0.01	4.16	0.82	0.01
	HPS	12.20	0.95	0.03	6.65	0.91	0.02	4.25	0.85	0.01	2.91	0.79	0.01
Tiny-ImageNet	APS	42.16	0.95	0.07	26.32	0.91	0.04	18.41	0.87	0.03	13.43	0.83	0.02
	HPS	38.21	0.95	0.06	22.23	0.91	0.03	14.38	0.87	0.02	9.59	0.82	0.02

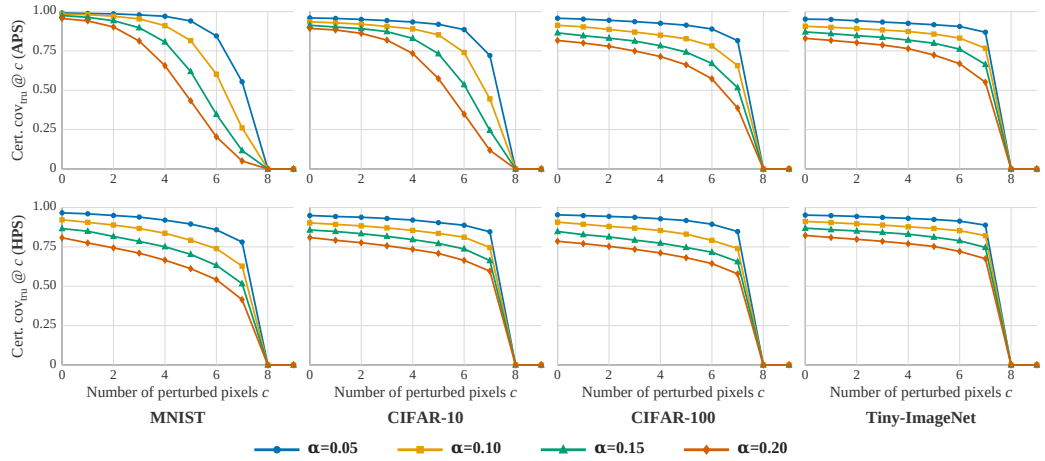


Figure 9: $\text{Cert.cov}_{tru}@c$ vs. α for `TrigCon`.

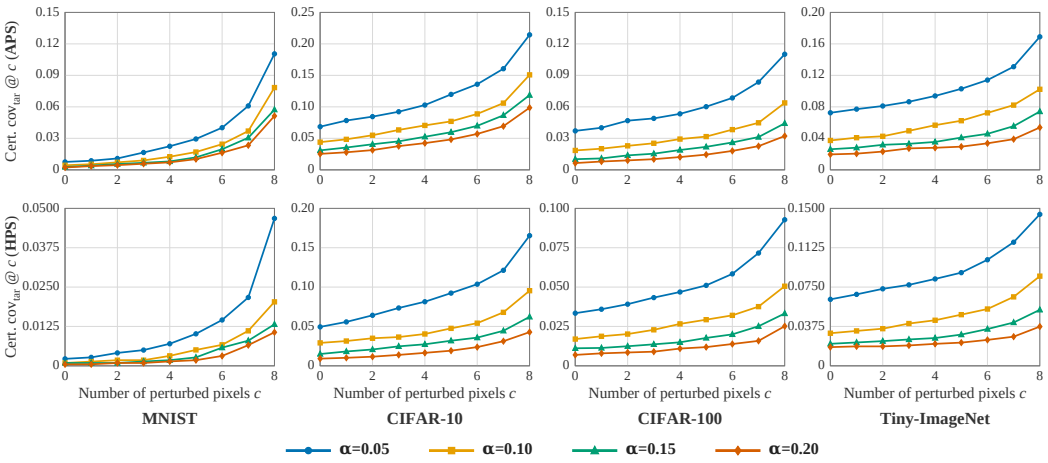


Figure 10: $\text{Cert.cov}_{tar}@c$ vs. α for `TrigCon`.

(3) Tighter certified coverage. Based on the conditions $y_{n+1} \in \mathcal{C}'_{\text{en}}(x_{n+1})$ and $y_t \notin \mathcal{C}'_{\text{en}}(x_{n+1})$ and $c \leq \min\{c^*_{y_{n+1}}, c^*_{y_t}\}$, we evaluate whether `TrigCon` can retain the true label and exclude the target label simultaneously for a sample x_{n+1} under worst-case backdoor attacks. Table 15 and Figure 13 report the clean-input and certified results, respectively. Under this stricter requirement, `TrigCon` maintains high certified true-label coverage and low certified target-label coverage under a limited trigger size. For example, at $T = 16$ and $c = 2$, $\text{Cert.cov}_{tru}@c$ exceeds 0.86, while $\text{Cert.cov}_{tar}@c$ remains below 0.09 across all four datasets.

Table 14: Defense ablation: cov_{tru}, cov_{tar}, and *size* of TrigCon on clean and triggered test samples under varying ratios of calibration and test set.

Ratio	Score	Input	MNIST			CIFAR-10			CIFAR-100			Tiny-ImageNet		
			size	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}
20%	APS	Clean	1.24	0.99	0.01	1.75	0.96	0.08	14.18	0.96	0.04	43.33	0.95	0.07
		Triggered	1.24	0.99	0.02	1.78	0.96	0.08	14.78	0.96	0.05	43.73	0.95	0.08
	HPS	Clean	1.03	0.97	0.00	1.56	0.96	0.06	12.29	0.95	0.04	38.88	0.95	0.06
		Triggered	1.03	0.96	0.01	1.58	0.95	0.06	12.68	0.96	0.05	39.20	0.95	0.07
40%	APS	Clean	1.24	0.99	0.01	1.75	0.96	0.08	14.33	0.96	0.04	43.28	0.95	0.07
		Triggered	1.24	0.99	0.02	1.78	0.96	0.08	14.93	0.96	0.05	43.67	0.95	0.08
	HPS	Clean	1.03	0.97	0.00	1.57	0.96	0.06	12.35	0.96	0.04	38.63	0.95	0.06
		Triggered	1.03	0.96	0.01	1.58	0.96	0.06	12.74	0.96	0.05	38.94	0.95	0.07
50%	APS	Clean	1.24	0.99	0.01	1.74	0.96	0.08	14.17	0.96	0.04	43.40	0.95	0.07
		Triggered	1.24	0.99	0.02	1.77	0.96	0.08	14.78	0.96	0.05	43.80	0.95	0.08
	HPS	Clean	1.03	0.97	0.00	1.56	0.96	0.06	12.21	0.95	0.04	38.87	0.95	0.06
		Triggered	1.03	0.97	0.01	1.58	0.95	0.06	12.60	0.95	0.04	39.19	0.95	0.07
60%	APS	Clean	1.24	0.99	0.01	1.75	0.96	0.08	14.21	0.96	0.04	43.48	0.96	0.08
		Triggered	1.24	0.99	0.02	1.78	0.96	0.08	14.81	0.96	0.05	43.87	0.95	0.08
	HPS	Clean	1.03	0.97	0.00	1.57	0.96	0.06	12.31	0.96	0.04	38.80	0.95	0.06
		Triggered	1.03	0.96	0.01	1.59	0.96	0.06	12.70	0.96	0.04	39.11	0.95	0.07
80%	APS	Clean	1.25	0.99	0.01	1.75	0.96	0.08	14.20	0.96	0.04	43.43	0.96	0.07
		Triggered	1.24	0.99	0.02	1.78	0.96	0.08	14.81	0.96	0.05	43.81	0.96	0.08
	HPS	Clean	1.03	0.97	0.00	1.57	0.96	0.06	12.33	0.96	0.04	38.86	0.95	0.06
		Triggered	1.03	0.96	0.01	1.58	0.95	0.06	12.71	0.96	0.04	39.14	0.95	0.07

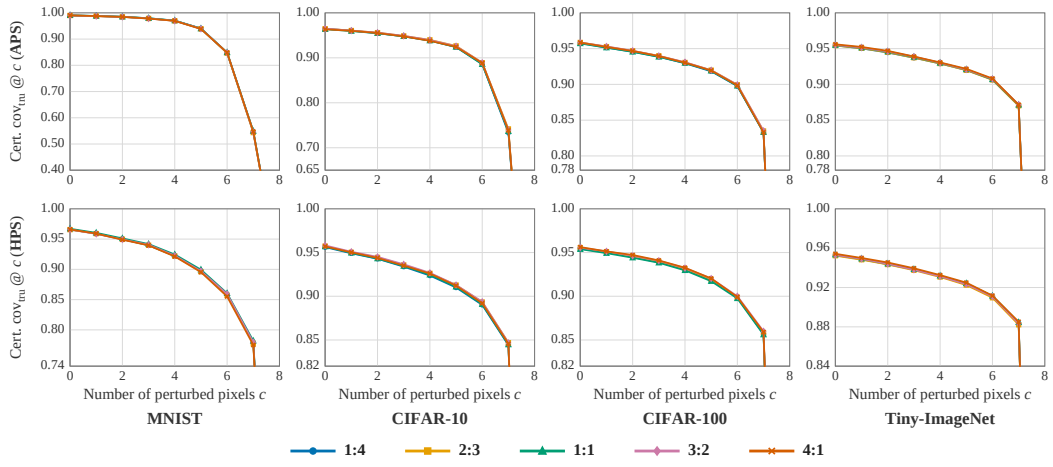


Figure 11: Cert.cov_{tru}@c vs. the ratio of calibration and test set for TrigCon.

Table 15: Tighter coverage of TrigCon on clean test samples with varying T using APS.

Dataset	Score	$T=4$			$T=8$			$T=16$			$T=32(24)$			$T=64(32)$		
		size	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}	*size*	cov_{tru}	cov_{tar}
MNIST	APS	1.02	0.97	0.00	1.08	0.98	0.00	1.24	0.99	0.01	1.67	0.97	0.02	3.16	0.92	0.08
CIFAR-10	APS	1.45	0.91	0.05	1.52	0.91	0.06	1.69	0.90	0.07	2.15	0.87	0.11	3.02	0.80	0.18
CIFAR-100	APS	6.64	0.91	0.02	9.20	0.91	0.02	14.15	0.92	0.04	18.07	0.91	0.05	27.02	0.89	0.09
Tiny-ImageNet	APS	28.50	0.90	0.05	35.47	0.89	0.06	42.16	0.88	0.07	45.88	0.87	0.08	50.44	0.86	0.10

(4) Performance on real backdoor settings. Table 16 shows the results of TrigCon on real backdoor test samples. We can observe that TrigCon simultaneously achieves high true-label coverage and low target-label coverage on real triggered inputs. For example, at $T = 16$, cov_{tru} ranges from 95% to 99%, while cov_{tar} remains at most 8% across all four datasets. These results further demonstrate the effectiveness of TrigCon.

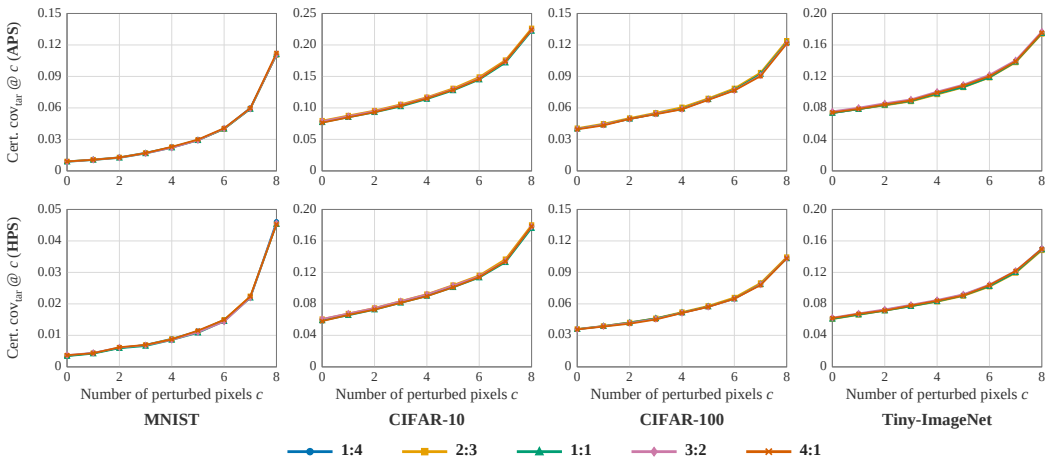


Figure 12: Cert.cov$_{tar}$@c vs. the ratio of calibration and test set for TrigCon.

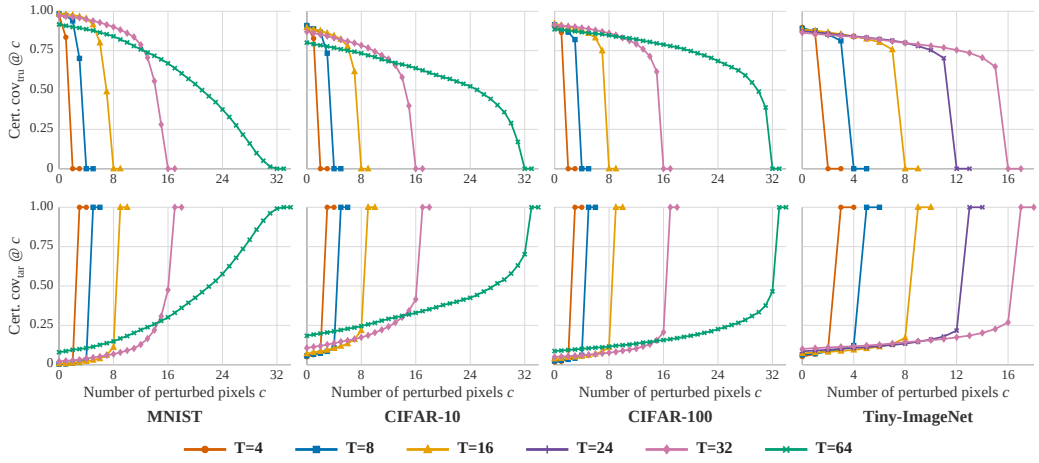


Figure 13: Tighter Cert.cov$_{tru}$@c (top) and Cert.cov$_{tar}$@c (bottom) vs. T for TrigCon with APS.

Table 16: cov$_{\text{tru}}$, cov$_{\text{tar}}$, and *size* of TrigCon on triggered test samples under varying T. Values in parentheses denote T for Tiny-ImageNet.

Dataset	Score	$T=4$			$T=8$			$T=16$			$T=32(24)$			$T=64(32)$		
		size	cov$_{\text{tru}}$	cov$_{\text{tar}}$	*size*	cov$_{\text{tru}}$	cov$_{\text{tar}}$	*size*	cov$_{\text{tru}}$	cov$_{\text{tar}}$	*size*	cov$_{\text{tru}}$	cov$_{\text{tar}}$	*size*	cov$_{\text{tru}}$	cov$_{\text{tar}}$
MNIST	APS	1.02	0.97	0.01	1.08	0.99	0.01	1.24	0.99	0.01	1.66	0.99	0.03	3.12	1.00	0.10
	HPS	0.95	0.94	0.00	0.97	0.95	0.00	1.03	0.96	0.01	1.38	0.98	0.02	2.74	0.99	0.09
CIFAR-10	APS	1.47	0.95	0.05	1.53	0.96	0.06	1.72	0.96	0.07	2.16	0.97	0.10	3.05	0.98	0.19
	HPS	1.27	0.94	0.04	1.34	0.94	0.04	1.52	0.95	0.05	1.99	0.96	0.09	2.78	0.97	0.16
CIFAR-100	APS	7.27	0.93	0.04	9.68	0.94	0.03	14.78	0.96	0.05	18.76	0.96	0.06	27.63	0.97	0.11
	HPS	5.50	0.92	0.02	8.07	0.93	0.03	12.61	0.95	0.04	16.42	0.96	0.06	25.54	0.97	0.10
Tiny-ImageNet	APS	28.90	0.94	0.07	35.83	0.95	0.07	42.51	0.95	0.08	46.30	0.95	0.10	51.00	0.96	0.11
	HPS	24.95	0.95	0.06	31.41	0.95	0.06	38.50	0.95	0.07	41.89	0.96	0.09	47.27	0.96	0.10