PROVABLY ROBUST CONFORMAL PREDICTION AGAINST ARBITRARY DATA POISONING ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal Prediction (CP) provides a principled approach to uncertainty quantification in machine learning by constructing prediction sets that provably contain the true label with high coverage. However, recent studies have shown that CP is highly vulnerable to data poisoning attacks: manipulating even a small fraction of training or calibration samples can significantly inflate prediction set sizes or violate coverage guarantees, undermining CP's reliability. We propose `JointCP`, a robust CP framework to defend against arbitrary data poisoning attacks with formal coverage guarantees, i.e., the coverage is still valid when a fraction of training and calibration samples used in CP are arbitrarily perturbed (e.g., deleted, added, modified, or combined in any manner). Results on multiple image classification datasets demonstrate the effectiveness and robustness of `JointCP` from both empirical and theoretical perspectives.

1 INTRODUCTION

Conformal Prediction (CP) Vovk et al. (2005); Papadopoulos et al. (2002); Lei & Wasserman (2014) is a statistical framework for uncertainty quantification in machine learning (ML) that constructs prediction sets for individual test samples with rigorous coverage guarantees. Given training, calibration, and test samples drawn from an underlying distribution, CP proceeds as follows: (1) train an ML model on the training set; (2) compute conformal scores on the calibration set using the trained model and a conformal scoring function; and (3) use the calibrated scores to build a prediction set for each test sample that contains the true label with a prescribed coverage. In addition to coverage guarantees, an effective CP method aims to produce prediction sets of reasonable size to ensure practical utility.

CP has recently gained significant popularity due to its distribution-free and model-agnostic nature Angelopoulos et al. (2021); Angelopoulos & Bates (2022); Stutz et al. (2022); Einbinder et al. (2022). It has been widely applied to (high-stakes) ML tasks such as image classification Angelopoulos et al. (2021), information retrieval Fisch et al. (2021), medical diagnosis Razzak et al. (2017), question answering Li et al. (2024a), factuality verification Mohri & Hashimoto (2024), and enhancing human-machine agreement Gui et al. (2024); Jung et al. (2025). For example, in image classification, CP produces a set of predicted classes that provably contains the true label with a probability 90%.

However, just as standard ML models are vulnerable to data poisoning attacks Biggio et al. (2012); Jagielski et al. (2018), CP methods suffer from similar vulnerabilities Park et al. (2023); Li et al. (2024b). For instance, Li et al. (2024b) demonstrate that perturbing only 0.5% of the training data on Tiny-ImageNet can triple the prediction set size, substantially undermining the practical utility of CP. The fundamental reason is that manipulating a few training samples can largely affect the learned ML model, the conformal scores, and thus the prediction sets.

To address this challenge, we design a CP framework that is provably robust to data poisoning attacks, where an adversary is allowed to *arbitrarily* poison up to a bounded number of training **and** calibration samples without restriction on how those samples are perturbed (e.g., removed, added, modified, or combined in any manner). We propose `JointCP` to achieve this goal, drawing inspiration from ensemble learning Zhou (2025) (See Figure 1 an overview). Rather than relying on a single ML model or conformal function, `JointCP` learns a collection of independent ML models and conformal functions on disjoint splits of training and calibration data, resulting in multiple individual CPs. `JointCP` then aggregates the outputs of these CPs using a voting-based mechanism (parameterized by $\alpha > 1$) to produce the final prediction set. The key insight is that manipulating a

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.

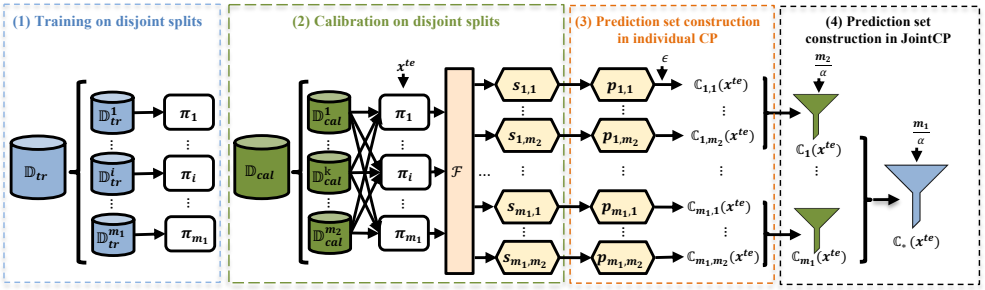


Figure 1: Overview of JointCP. Given training data $\mathbb{D}_{\text{tr}}$, calibration data $\mathbb{D}_{\text{cal}}$, a test sample (x^{te}, y^{te}), an aggregation parameter $\alpha > 1$, and a significance level ϵ, JointCP contains four steps: **(1)** Train m_1 models $\{\pi_i\}_{i=1}^{m_1}$ on m_1 disjoint training splits $\{\mathbb{D}_{tr}^i\}_{i=1}^{m_1}$ partitioned by $\mathbb{D}_{\text{tr}}$. **(2)** Partition $\mathbb{D}_{\text{cal}}$ into m_2 disjoint calibration splits $\{\mathbb{D}_{cal}^k\}_{k=1}^{m_2}$, which are used to compute conformal scores $\{s_{i,k}\}_{k=1}^{m_2}$ via a conformal function $\mathcal{F}$ for each trained model π_i. **(3)** Compute a p-value $p_{i,k}$ for x^{te} based on the conformal score $s_{i,k}$ and construct the prediction set $\mathbb{C}_{i,k}(x^{te})$ to form each CP $CP_{i,k}$. **(4)** Aggregate over m_2 prediction sets $\{\mathbb{C}_{i,k}(x^{te})\}_{k=1}^{m_2}$ with a threshold $\frac{m_2}{\alpha}$ to obtain an aggregated set $\mathbb{C}_i(x^{te})$ per model π_i, and further aggregate the m_1 prediction sets $\{\mathbb{C}_i(x^{te})\}_{i=1}^{m_1}$ over all models with a threshold $\frac{m_1}{\alpha}$ to produce the final prediction set $\mathbb{C}_*(x^{te})$. **Distribution-free coverage guarantee:** *(i) Under clean data*: $\Pr[y^{te} \in \mathbb{C}_*(x^{te})] \geq 1 - \frac{\alpha^2\epsilon}{(\alpha-1)^2}$. *(ii) Under arbitrary poisoned c_{tr} training samples and c_{cal} calibration samples*: $\Pr[y^{te} \in \mathbb{C}'_*(x^{te})] \geq 1 - \frac{\alpha}{\alpha-1} \cdot \frac{\epsilon + c_{cal}/|\mathbb{D}_{cal}|}{1-1/\alpha-c_{tr}/m_1}$.

single training or calibration sample affects at most one CP or conformal score relative to the clean setting. More generally, when multiple samples are arbitrarily poisoned, the number of affected CPs and conformal scores can be bounded. As a result, when a majority of individual CPs remain intact, the aggregated CP can still produce prediction sets with valid coverage and reasonable size.

(Theoretical guarantees). *(i)* Without attacks, if vanilla CP guarantees a coverage level of $1 - \epsilon$, JointCP guarantees a coverage of $1 - \frac{\alpha^2}{(\alpha-1)^2}\epsilon$, where larger α yields stronger coverage guarantee. *(ii)* Under arbitrary poisoning of up to c_{tr} training samples and c_{cal} calibration samples, JointCP provides a coverage guarantee that degrades gracefully with c_{tr} and c_{cal} (see the exact expression in Equation 12). *(iii)* When the clean prediction set produced by JointCP for a test sample is known *a priori*, it can be further exploited to derive a tighter coverage guarantee under data poisoning attacks.[1]

(Empirical results). Our experiments on three image classification datasets show that: *(i)* In the absence of attacks, JointCP consistently achieves empirical coverage above the guaranteed level, while producing slightly larger prediction sets than vanilla CP. *(ii)* Under worst-case poisoning attacks, JointCP maintains reasonable certified coverage without excessively enlarging certified prediction sets, as long as the number of poisoned samples stays within the certified bound. These results collectively demonstrate the robustness of the proposed CP framework against arbitrary training and calibration data poisoning. We summarize our main contributions as follows:

- We study provably robust conformal prediction (CP) against arbitrary data poisoning attacks.
- We develop JointCP to ensure certified coverage under both clean and poisoned data settings.
- Results on various image classification datasets validate the effectiveness and robustness of JointCP, from both empirical and theoretical perspectives.

2 Background and Problem Setup

2.1 Conformal Prediction (CP)

Let a dataset $\mathbb{D} \subset \mathbb{X} \times \mathbb{Y}$ be split into a training set $\mathbb{D}_{tr} = \{(x_i^{tr}, y_i^{tr})\}_{i=1}^{n_{tr}}$, and a disjoint calibration set $\mathbb{D}_{cal} = \{(x_i^{cal}, y_i^{cal})\}_{i=1}^{n_{cal}}$, with samples drawn from an unknown distribution P. Here, x_i and y_i represent the features and label of a sample; $\mathbb{X}$ and $\mathbb{Y}$ the feature space and label space, respectively. A CP typically contains the following three steps:

[1]We note a closely related work (Scholten & Günnemann, 2025). While both approaches defend against data poisoning attacks in CP, they differ in their objectives, methods, theoretical guarantees, and evaluation metrics. More detailed discussion and comparison are provided in Section 4.4.

(1) Model training. We learn an ML model π that maps a sample x to a probability vector $\pi(x) \in [0,1]^{|\mathbb{Y}|}$ by minimizing a loss ℓ on the training set $\mathbb{D}_{tr}$:

$$\ell(\pi; \mathbb{D}_{tr}) = \sum_{(x_i^{tr}, y_i^{tr}) \in \mathbb{D}_{tr}} \text{CE}(\pi(x_i^{tr}), y_i^{tr}). \tag{1}$$

(2) Calibration. Given the model π and a sample x, a conformal function $\mathcal{F}$ is used to compute a conformal score between x and each candidate label $y \in \mathbb{Y}$ based on $\pi_y(x)$. For example, when instantiating $\mathcal{F}$ using the Adaptive Prediction Set (APS) Romano et al. (2020), we have

$$\mathcal{F}(\pi, x, y) = \sum_{y' \in \mathbb{Y}} \pi_{y'}(x) \mathbf{1}(\pi_{y'}(x) > \pi_y(x)) + \pi_y(x) u, \tag{2}$$

where $\mathbf{1}(\cdot)$ is an indicator and u is a random sample from the uniform distribution $\text{Uniform}(0,1)$. Then, we use the calibration set $\mathbb{D}_{cal}$ to compute the conformal scores s for calibration samples. Specifically, given $\mathcal{F}$ and π, s are calculated as:

$$s(x_i^{cal}, y_i^{cal}) = \mathcal{F}(\pi, x_i^{cal}, y_i^{cal}), \forall (x_i^{cal}, y_i^{cal}) \in \mathbb{D}_{cal}. \tag{3}$$

(3) Prediction set construction. For a test sample (x^{te}, y^{te}) and candidate label $y \in \mathbb{Y}$, the conformal score is computed as $s(x^{te}, y) = \mathcal{F}(\pi, x^{te}, y)$. We compute the p-value associated with label y as:

$$p(x^{te}, y) = \frac{|\{i : s(x^{te}, y) \leq s(x_i^{cal}, y_i^{cal})\}| + 1}{|\mathbb{D}_{cal}| + 1}. \tag{4}$$

This p-value measures how compatible the test example (x^{te}, y) is with the calibration set $\mathbb{D}_{cal}$. Given a prescribed significance level $\epsilon \in (0,1)$, CP produces the prediction set for x^{te} as:

$$\mathbb{C}(x^{te}) = \{y \in \mathbb{Y} : p(x^{te}, y) > \epsilon\}. \tag{5}$$

Coverage guarantee (Angelopoulos & Bates, 2022): Assume exchangeable [2] calibration and test samples. Equation (5) ensures that $\mathbb{C}(x^{te})$ contains the true label y^{te} with a probability at least $1 - \epsilon$.

$$\Pr[y^{te} \in \mathbb{C}(x^{te})] \geq 1 - \epsilon. \tag{6}$$

This coverage guarantee is valid regardless of the underlying distribution P (a.k.a. **distribution-free coverage**), except for the exchangeability. Ideally, a CP method is desired to yield a *smaller* prediction set size $|\mathbb{C}|$ and a *higher* coverage guarantee $1 - \epsilon$. The coverage serves as a practical measure of how well the CP performs on unseen data.

2.2 Threat Model

Adversary: It aims to compromise CP via data poisoning.

- **Adversary knowledge.** To offer robustness under the *worst-case attack*, we assume an adversary with complete knowledge of the CP learning, e.g., the base model, calibration procedure, and data.
- **Adversary capability.** An adversary is allowed to manipulate both model training (Step 1) and calibration (Step 2) of CP through data poisoning. Let c_{tr} and c_{cal} be the maximal number of poisoned training and calibration samples, respectively. We consider *the worst-case attack*, where the adversary can *arbitrarily* perturb this number of samples (e.g., by adding, removing, or modifying them) to affect CP learning. We consider poisoning that occurs *either when training the model from scratch or during downstream model fine-tuning and conformal calibration*. In the downstream setting, we assume that a pretrained backbone is used and remains fixed and trusted.
- **Adversary objective.** The adversary aims to undermine the reliability of CP. In particular, the adversary seeks to reduce coverage by *excluding the true label* from the prediction set and/or to increase the prediction set size by *retaining as many incorrect labels* as possible.

Defender: It is to construct a CP framework that is provably robust to arbitrary data poisoned within the prescribed budgets c_{tr} and c_{cal}. The defender aims to ensure:

- **Guaranteed (high) coverage**: the prediction set provably contains the true label of a test sample with a predefined coverage, when no more than c_{tr} training samples and c_{cal} calibration samples are arbitrarily manipulated.
- **Guaranteed (low) prediction set size**: provably limit the unnecessary labels in the prediction set, thus maintaining a controlled and informative prediction set size.

[2] A finite sequence of random variables $Z_1, \ldots, Z_n$ is *exchangeable* if, for any permutation τ of $1, \ldots, n$, $(Z_1, \ldots, Z_n) \overset{d}{=} (Z_{\tau(1)}, \ldots, Z_{\tau(n)})$.

3 JOINTCP: DEFENDING AGAINST ARBITRARY TRAINING AND CALIBRATION POISONING ON CP

3.1 OVERVIEW

Existing CP methods are intrinsically vulnerable to poisoning attacks, as manipulating even a single training or calibration sample can largely affect the learned model and/or the conformal score. To mitigate this vulnerability, we propose to learn *a collection of* independent ML models and/or conformal functions using *disjoint* splits of the training and calibration data, thereby producing a set of individual CPs. We then carefully aggregate these individual CPs to form an aggregated CP, which we denote as `JointCP`. Throughout the certification analysis, we assume a fixed partition assignment and deterministic training of each submodel given its assigned partition. Consequently, arbitrarily manipulating a single sample can affect only one individual CP or conformal score. Further, when multiple samples are arbitrarily perturbed, the number of affected CPs and conformal scores can be bounded. Consequently, as long as a majority of individual CPs remain unaffected, a properly designed aggregation strategy can still yield prediction sets with reasonable size while preserving coverage guarantees. Specifically, to handle arbitrary poisoning of both training and calibration data, we split the training set and calibration set into (e.g., m_1 and m_2) disjoint splits. For each training subset, we train an ML model using one subset and calibrate it using the m_2 calibration splits, resulting in multiple individual CPs. We then aggregate the m_1 trained models in a principled manner such that a (lower-bounded) coverage guarantee can be derived. Figure 1 provides an overview of `JointCP`; detailed complexity and runtime analyses for `JointCP` are provided in Appendix C.

3.2 JOINTCP

We now present our `JointCP` in detail. Given a significance level ϵ and two positive integers m_1 and m_2. We partition the training set $\mathbb{D}_{tr}$ into m_1 disjoint splits $\{\mathbb{D}_{tr}^i\}_{i=1}^{m_1}$ and the calibration set $\mathbb{D}_{cal}$ into m_2 disjoint splits $\{\mathbb{D}_{cal}^k\}_{k=1}^{m_2}$. `JointCP` consists of the below steps:

(1) Train m_1 models on disjoint training splits. For each training subset $\mathbb{D}_{tr}^i$, we train an individual base predictive model π_i on it to minimize the training loss via Equation (1).

(2) Calibrate on m_2 disjoint calibration splits. For each trained model π_i, we perform the calibration by aggregating information across the m_2 calibration splits $\{\mathbb{D}_{cal}^k\}_{k=1}^{m_2}$. Specifically, on each calibration subset $\mathbb{D}_{cal}^k$, we compute the conformal score $s_{i,k}$ for each calibration sample $(x_{j,k}^{cal}, y_{j,k}^{cal}) \in \mathbb{D}_{cal}^k$ using a conformal function $\mathcal{F}$ as $s_{i,k}(x_{j,k}^{cal}, y_{j,k}^{cal}) = \mathcal{F}(\pi_i, x_{j,k}^{cal}, y_{j,k}^{cal}), \forall (x_{j,k}^{cal}, y_{j,k}^{cal}) \in \mathbb{D}_{cal}^k$. Each pair (i,k) thus defines an individual conformal predictor, which we denote as $CP_{i,k}$.

(3) Prediction set construction in individual CP. Given a test sample x^{te} and a candidate label $y \in \mathbb{Y}$, we compute the p-value on each CP $CP_{i,k}$ as

$$p_{i,k}(x^{te}, y) = \frac{|\{j : s_{i,k}(x^{te}, y) \leq s_{i,k}(x_{j,k}^{cal}, y_{j,k}^{cal})\}| + 1}{|\mathbb{D}_{cal}^k| + 1} \tag{7}$$

The corresponding prediction set $\mathbb{C}_{i,k}(x^{te})$ is built as:

$$\mathbb{C}_{i,k}(x^{te}) = \{y \in \mathbb{Y} : p_{i,k}(x^{te}, y) > \epsilon\}. \tag{8}$$

(4) Prediction set construction in `JointCP`. A key challenge is to aggregate the prediction sets $\{\mathbb{C}_{i,k}\}$ produced by individual CPs $CP_{i,k}$ in a principled manner so that the final prediction set, $\mathbb{C}_*$, satisfies the desired coverage guarantee. To this end, we propose a voting-based conformal prediction aggregation strategy in `JointCP`, as detailed below.

- For each π_i, we first aggregate $\{\mathbb{C}_{i,k}\}_{k=1}^{m_2}$ over m_2 CPs: $v_y^i = \sum_{k=1}^{m_2} \mathbb{I}\big(y \in \mathbb{C}_{i,k}(x^{te})\big)$, where v_y^i indicates the number of CPs that include the label y in their prediction set $\{\mathbb{C}_{i,k}\}$. Denote $\alpha > 1$ as an aggregation parameter. The aggregated prediction set for π_i is defined as

$$\mathbb{C}_i(x^{te}) = \left\{y \in \mathbb{Y} : v_y^i \geq \frac{m_2}{\alpha}\right\}, \tag{9}$$

- We further perform another aggregation on m_1 aggregated prediction sets $\{\mathbb{C}_i\}_{i=1}^{m_1}$ for all models $\{\pi_i\}_{i=1}^{m_1}$, and define the vote count v_y: as $v_y = \sum_{i=1}^{m_1} \mathbb{I}\big(y \in \mathbb{C}_i(x^{te})\big)$.
- The final prediction set in `JointCP` is given by

$$\mathbb{C}_*(x^{te}) = \left\{y \in \mathbb{Y} : v_y \geq \frac{m_1}{\alpha}\right\}. \tag{10}$$

3.3 COVERAGE GUARANTEE

3.3.1 DISTRIBUTION-FREE COVERAGE

This section derives the coverage guarantee for `JointCP` in both clean and poisoned settings. *Proofs are in Appendix B.*

Coverage guarantee under the clean setting. The goal is to ensure the label y^{te} of a test sample x^{te} included in the prediction set $\mathbb{C}_*(x^{te})$ outputted by `JointCP` with high probability.

Theorem 3.1 (Coverage guarantee of `JointCP` under the clean data.)**.** *Given a significance level $\epsilon \in (0,1)$, an aggregation parameter α, m_1 training splits, m_2 calibration splits, and a test sample (x^{te}, y^{te}). The prediction set $\mathbb{C}_*(x^{te})$ of `JointCP` for x^{te} on $m_1 \cdot m_2$ individual CPs $\{CP_{i,k}\}$ satisfies the following coverage guarantee:*

$$\Pr[y^{te} \in \mathbb{C}_*(x^{te})] \geq 1 - \frac{\alpha^2 \epsilon}{(\alpha-1)^2}. \tag{11}$$

Remark: α controls the trade-off between coverage and prediction set size. From Equation (11), a larger α yields a higher coverage guarantee. Meanwhile, Equations (10) and (9) show that increasing α lowers the voting threshold m_2/α and m_1/α, leading to larger prediction sets.

Coverage guarantee under the poisoned setting. We further analyze `JointCP` when an adversary can arbitrarily corrupt a bounded number of samples in the training and calibration sets. Our goal is to ensure that the true label of a test sample is still included in `JointCP`'s prediction set with the desired coverage guarantee, as long as the number of poisoned samples is below a threshold. The formal coverage guarantee is given below:

Theorem 3.2 (Coverage guarantee of `JointCP` under training and calibration data poisoning)**.** *Let c_{tr} and c_{cal} denote the number of poisoned training samples and calibration samples, respectively. Consider `JointCP` with m_1 training splits and m_2 calibration splits, and let each individual CP use significance level $\epsilon \in (0,1)$. Suppose c_{tr} satisfies $c_{tr} < m_1(1-1/\alpha)$. Then the prediction set $\mathbb{C}'_*(x^{te})$ given by `JointCP` satisfies the coverage guarantee*

$$\Pr[y^{te} \in \mathbb{C}'_*(x^{te})] \geq 1 - \frac{\alpha}{\alpha-1} \cdot \frac{\epsilon + \frac{c_{cal}}{|\mathbb{D}_{cal}|}}{1 - \frac{1}{\alpha} - \frac{c_{tr}}{m_1}} \tag{12}$$

Remark: Equation (12) quantifies how `JointCP`'s coverage degrades as a function of the poisoning level—a larger c_{tr} or/and c_{cal} can reduce the coverage with a given α. When $c_{tr} = c_{cal} = 0$, it reduces to clean data guarantee. Like Theorem 3.1, a smaller/larger α yields smaller/higher coverage. Note that the theorem assumes that c_{tr} and c_{cal} are known separately. If only the total attack budget $c = c_{tr} + c_{cal}$ is known while its split between training and calibration is unknown, one can obtain a conservative guarantee by replacing c_{tr} and c_{cal} with c in Equation (12). The effect of m_2 is absorbed into $|\mathcal{D}_{\text{cal}}|$ through Theorem A.3. When an upper bound on the poisoning budget is available, the defender can evaluate the certified coverage in Equation (12) for candidate values of α and select the smallest α for which the certified coverage meets the desired level under the assumed budget.

3.3.2 DATA-SPECIFIC COVERAGE

We further derive a fine-grained, data-specific coverage bound under poisoning.

Let $\mathbb{C}_*(x^{te})$ be the prediction set given by `JointCP` before attack, and $\mathbb{C}'_*(x^{te})$ be the corresponding prediction set after training and calibration data are poisoned. We aim to certify, for each label $y \in \mathbb{Y}$, whether it remains in the final prediction set $\mathbb{C}'_*(x^{te})$ after arbitrary poisoning, based on its status in the clean prediction set $\mathbb{C}_*(x^{te})$ and number of poisoned samples c_{tr} and c_{cal}.

Specifically, for each training subset index $i \in [m_1]$, let $G_i^{te}(y, c_{cal})$ and $c_i^*(y)$ denote the quantities in Equation (24) in Appendix A over the m_2 prediction sets $\{\mathbb{C}_{i,k}\}_{k=1}^{m_2}$. Further, let $c^*(y)$ be defined as in Equation (20) in Appendix A over the m_1 aggregated prediction sets $\{\mathbb{C}_i\}_{i=1}^{m_1}$.

Theorem 3.3 (Label-wise tolerance under training and calibration data poisoning)**.** *Consider `JointCP` with m_1 training splits and m_2 calibration splits and a test sample x^{te}. Consider any label $y \in \mathbb{Y}$. If c_{tr} and c_{cal} satisfy*

$$c_{tr} + |\{i : G_i^{te}(y, c_{cal}) > c_i^*(y)\}| < c^*(y), \tag{13}$$

then the status of label y in the final prediction set is unchanged after the poisoned attack. That is,

$$y \in \mathbb{C}_*(x^{te}) \iff y \in \mathbb{C}'_*(x^{te}), \quad \textit{and} \quad y \notin \mathbb{C}_*(x^{te}) \iff y \notin \mathbb{C}'_*(x^{te}).$$

Table 1: Empirical coverage (EmpC) and average prediction set size (AveS) for vanilla CP and JointCP ($\epsilon = 0.05, \alpha = 2$, clean setting).

Dataset	Vanilla CP		m = 10		m = 20		m = 50		m = 100	
	EmpC	AveS	EmpC	AveS	EmpC	AveS	EmpC	AveS	EmpC	AveS
SVHN	0.81	0.84	0.98	1.03	0.98	1.03	0.98	1.06	0.98	1.09
CIFAR10	0.81	0.86	0.96	1.37	0.96	1.41	0.98	1.61	0.97	1.77
CIFAR100	0.80	2.36	0.96	11.12	0.96	12.03	0.97	15.66	0.98	18.39

Remark: This result enables us to quantitatively assess the robustness of JointCP by introducing two metrics: *certified coverage* and *certified prediction set size*.

- **Certified coverage:** It is the probability that the true label y^{te} of a test sample x^{te} is included in $\mathbb{C}_*(x^{te})$, where c_{tr} and c_{cal} satisfy Equation (13). This provides a worst-case guarantee: under at most c_{tr} and c_{cal} poisoned samples, the true label will still be in the prediction set.
- **Certified prediction set size:** This is the size of $\mathbb{C}_*(x^{te})$ after adding any label $y \notin \mathbb{C}_*(x^{te})$ that satisfies Equation 13 and removing any label $y \in \mathbb{C}_*(x^{te})$ that violates Equation 13. The quantity measures the size of the prediction set under the poisoning of c_{tr} and c_{cal} samples.

4 Experiments

4.1 Experimental Setup

Datasets and models: We evaluate JointCP on three benchmark image classification datasets: CIFAR10, CIFAR100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011). Following (Gendler et al., 2021) (Yan et al., 2024), we utilize the pretrained ResNet-110 (He et al., 2016) model for CIFAR10 and CIFAR100. For SVHN, we follow Scholten & Günnemann (2025) to train a ResNet101 (He et al., 2016) base model from scratch. Unless otherwise specified, we use disjoint subsets of 5,000 samples for sub-model training, 2,500 for calibration, and 2,500 for testing. JointCP finetunes the pretrained ResNet-110 model or trains the ResNet101 model with 200 epochs on $m_1 = 100$ disjoint subsets of training samples (not used in pretraining) with equal size, resulting in 100 sub-models for calibration and prediction set construction for test samples. *Further experiments on data augmentation and scalability (e.g., Tiny ImageNet (Le et al., 2015)) are deferred to Appendix D.*

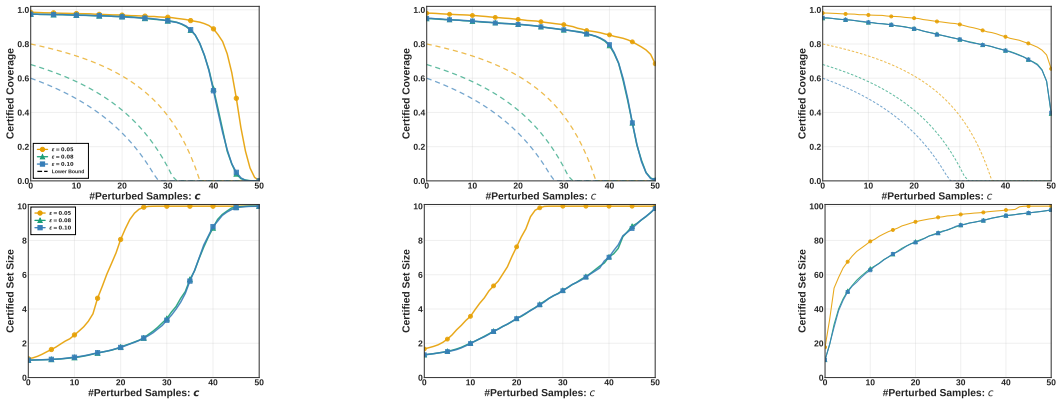
Parameter setting: We employ the Adaptive Prediction Sets (APS) method (Romano et al., 2020) as the conformal function. By default, we use $m_1 = m_2 = 100$ training and calibration splits. Unless otherwise specified, we assume the adversary poisons an equal maximal number of training samples c_{tr} and calibration samples c_{cal}. Consequently, the "#Perturbed Samples c" in our figures indicates $c_{\text{tr}} = c_{\text{cal}} = c$. Further, the default aggregation parameter $\alpha = 2$ and CP significance level $\epsilon = 0.05$. We will also study the impact of $\alpha \in \{1.5, 2.0, 3.0, 4.0\}$, $\epsilon \in \{0.05, 0.08, 0.10\}$, as well as different c_{tr} and c_{cal}. *Additional ablations on score functions and parameter sensitivity are in Appendix D.*

Evaluation metric: We evaluate CP performance on clean data using *empirical coverage* and *average prediction set size* over test samples, following Angelopoulos et al. (2021); Angelopoulos & Bates (2022). We report *certified coverage* and *certified prediction set size* to quantify robustness under arbitrary data poisoning. Higher certified coverage and lower certified prediction set size indicate better robustness. For a given dataset, the certified coverage provides a lower bound on the coverage that holds against any admissible attack, and the experiments empirically evaluate this guarantee.

4.2 Results on Clean Data

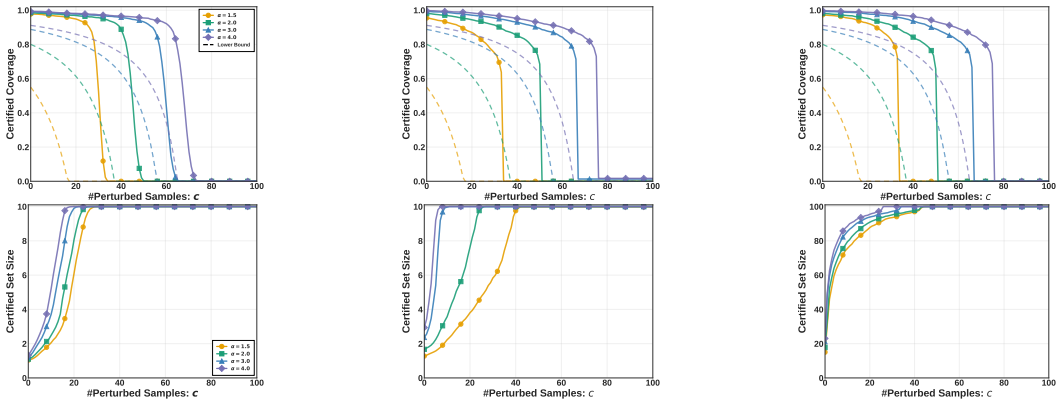
In this experiment, we compare JointCP with vanilla CP on the clean setting in order to show that JointCP maintains both the coverage and prediction set size. Given a significance level ϵ, vanilla CP ensures a coverage $1 - \epsilon$, while JointCP has a coverage $1 - 4\epsilon$ (Equation (11) with default $\alpha = 2$). For a fair comparison, when setting an ϵ in JointCP, we set the significance level of vanilla CP to be 4ϵ. Table 1 shows the comparison results under the same setting with $\epsilon = 0.05$.

We have the following observations: First, compared to vanilla CP, JointCP achieves higher empirical coverage at the cost of larger average prediction sets, reflecting the tradeoff between coverage and set size. Second, JointCP consistently attains empirical coverage substantially higher than its guaranteed coverage. Under the default setting with $\epsilon = 0.05$ and $\alpha = 2$, Equation (11) gives a worst-case lower bound of $1 - 4\epsilon = 0.8$, while JointCP achieves empirical coverage around 0.97–0.98 across all evaluated datasets. This gap indicates that the bound is a conservative distribution-free certificate rather than a tight empirical estimate, while aggregation helps preserve true labels even when some constituent models are inaccurate.



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 2: Impact of ϵ on `JointCP` ($\alpha = 2.0$): certified coverage (top) and set size (bottom).



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 3: Impact of α on `JointCP` ($\epsilon = 0.05$): certified coverage (top) and set size (bottom). Dashed lines denote the distribution-free lower bound in Theorem 3.2.

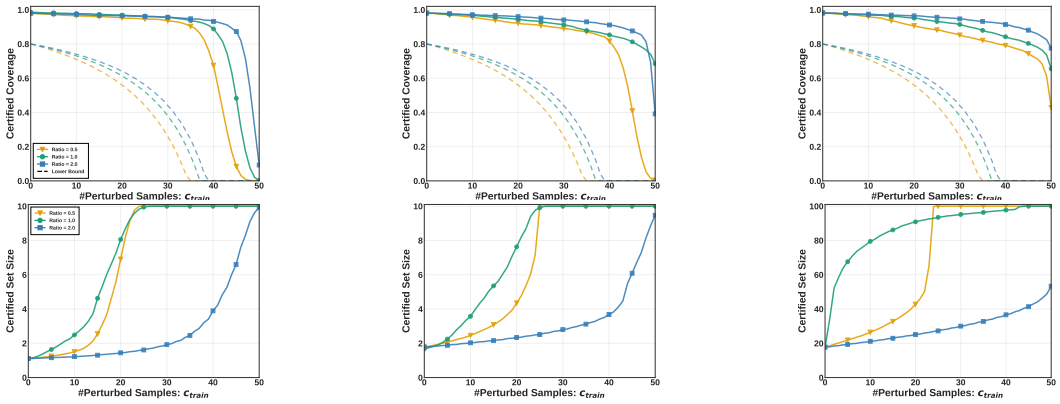
4.3 Results on Poisoned Data

In this experiment, we assess the provable robustness of `JointCP` via reporting the certified coverage and prediction set size under the worst-case attack scenario derived from our theoretical results in Theorem 3.2 and Theorem 3.3.

Impact of α on `JointCP`. Figure 3 displays the results. The key observations are as follows:

- **Larger α yields stronger certified coverage.** Increasing α consistently shifts the certified coverage curves to the right, meaning `JointCP` tolerates more perturbed samples. This matches the role of α in Eqs (9) and (10), where a larger α lowers the voting threshold and makes the aggregation more conservative, as well as reflecting the theoretical result in Theorem 3.2.
- **α controls the tradeoff between certified coverage and certified prediction set size.** As α increases, certified set size also becomes larger. This is because a stricter coverage guarantee requires more conservative label inclusion.
- **Coverage guarantees derived from Theorem 3.3 are consistently stronger than those from Theorem 3.2.** Theorem 3.3 provides data-dependent coverage guarantees by exploiting instance- and label-specific information, thereby yielding finer-grained and less conservative bounds. In contrast, Theorem 3.2 establishes distribution-free coverage guarantees that hold uniformly over all data distributions; while more conservative, these guarantees remain broadly applicable.

Impact of ϵ on `JointCP`. This experiment tests `JointCP` under different significance levels $\epsilon \in \{0.05, 0.08, 0.10\}$, with results shown in Figure 2. First, ϵ governs the tradeoff between certified coverage and certified prediction set size: smaller ϵ yields higher certified coverage at the cost of larger certified prediction sets. Second, for all values of ϵ, the data-dependent certified coverage consistently exceeds the lower-bounded (distribution-free) coverage.



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 4: Impact of poisoning ratio c_{tr}/c_{cal} on JointCP ($\alpha = 2.0$, $\epsilon = 0.05$).

Impact of poison ratio c_{tr}/c_{cal} on JointCP. This experiment studies how the poison ratio c_{tr}/c_{cal} affects JointCP's robustness. We vary the poison ratio $\text{ratio} = c_{\text{tr}}/c_{\text{cal}} \in \{0.5, 1.0, 2.0\}$; and results are reported in Figure 4. The key observations are as follows:

- **Certified coverage degrades/improves as the ratio decreases/increases.** Across all three datasets, smaller ratios (i.e., more calibration corruption) consistently yield lower certified coverage. This trend agrees with our arbitrary poisoning analysis: calibration poisoning increases the effective miscoverage level (e.g., through the term depending on $c_{\text{cal}}/|\mathbb{D}_{cal}|$), so decreasing ratio monotonically enlarges c_{cal} and weakens the coverage certificate.
- **Certified size is not necessarily monotone w.r.t. the poison ratio.** In contrast to coverage, the certified prediction set size curves do not preserve a strict ordering with respect to ratio. For instance, on CIFAR10 and SVHN, $\text{ratio} = 1.0$ can yield the *largest* certified size in the low-to-moderate budget range, while $\text{ratio} = 0.5$ stays smaller initially and then abruptly saturates to the maximum size. This behavior is consistent with our label-wise tolerance characterization. Unlike coverage, which aggregates degradation smoothly over test samples, calibration corruption increases the number of calibration splits that can be adversarially flipped, but a label can be forced into the final set only when the combined effect of $(c_{\text{tr}}, c_{\text{cal}})$ crosses the vote margins.

4.4 Comparison with RPS (Scholten & Günnemann, 2025)

Key differences between JointCP and RPS: JointCP differs from RPS in the following aspects. *(i) Objective:* RPS primarily characterizes robustness through *pointwise* prediction-set stability, whereas JointCP aims to maintain certified coverage while controlling certified prediction-set size under joint training and calibration poisoning. *(ii) Methodology:* RPS adopts a *one-level prediction-set aggregation*, following score aggregation across training-partition models. In contrast, JointCP introduces a *two-level prediction-set aggregation* framework tailored to CP: it first aggregates prediction sets across calibration splits for each model, and then aggregates the resulting prediction sets across models trained on different training splits. *(iii) Theory:* RPS provides pointwise, data-specific certificates for prediction-set stability under arbitrary data poisoning. In contrast, JointCP provides not only a data-specific certificate in Theorem 3.3, but also a *distribution-free* coverage guarantee in Theorem 3.2. *(iv) Metric:* Consistent with its pointwise guarantees, RPS evaluates stability (coverage/size) ratios, i.e., the fraction of test points for which labels are certified against removal/addition under poisoning. In contrast, JointCP evaluates certified coverage and prediction-set size via Theorem 3.2, as well as pointwise robustness via Theorem 3.3.

Adapting our theoretical results to RPS metrics. Since RPS does not directly provide our certified metrics, we instead adapt Theorem 3.3 to evaluate JointCP using the reliability metrics of RPS. Specifically, RPS evaluates pointwise certification using three metrics: coverage reliability (CovR), the fraction of test data for which all labels in the clean prediction set are certified to remain included after poisoning; size reliability (SizeR), the fraction for which all labels outside the clean prediction set are certified to remain excluded; and robust reliability (RobR), the fraction for which both properties hold. Theorem 3.3 determines whether each label is guaranteed to remain included or excluded under poisoning, and thus directly enables the computation of these RPS metrics for JointCP.

Table 2: Comparison between JointCP and RPS under the RPS evaluation protocol on SVHN.

Method	$c = 1$			$c = 3$			$c = 6$		
	CovR	SizeR	RobR	CovR	SizeR	RobR	CovR	SizeR	RobR
JointCP	75.2%	67.4%	51.7%	32.3%	34.0%	13.1%	13.6%	20.4%	0.03%
RPS	54.1%	46.7%	28.8%	25.5%	10.5%	9.9%	13.4%	1.5%	0.0%

Specifically, we use SVHN for comparison, as both methods train the model from scratch. We follow the SVHN setup of RPS, including its model architecture and calibration/test splits. We use $m_1 = 100$ training partitions and $m_2 = 40$ calibration partitions, with significance level $\epsilon = 0.1$. Although both methods allow different training- and calibration-poisoning budgets, we consider the symmetric setting $c_{tr} = c_{cal} = c$ to match the experiment reported in RPS. As shown in Table 2, JointCP achieves higher CovR, SizeR, and RobR across all evaluated poisoning budgets, demonstrating higher prediction-set reliability than RPS under the RPS evaluation metrics.

5 Related Work

Conformal prediction (CP). CP provides an uncertainty quantification framework to construct prediction sets with a user-specified confidence level (Papadopoulos et al., 2002; Vovk et al., 2005; Shafer & Vovk, 2008) without requiring any distributional assumptions on the data (Angelopoulos & Bates, 2022; Angelopoulos et al., 2021; Romano et al., 2020; Cauchois et al., 2021; Bates et al., 2021). CP is deployed on top of a pretrained model, treating the underlying model as "black box" to compute conformal scores on a separate calibration set. However, more recent works incorporate CP objectives (e.g., coverage or set-efficiency surrogates) into training or fine-tuning, enabling joint optimization of the underlying model and resulting conformal predictor (Gendler et al., 2021; Einbinder et al., 2022; Ghosh et al., 2023; Yan et al., 2024; Jeary et al., 2024; Yeh et al., 2025). This perspective has extended broader optimization directions for CP, e.g., directly optimizing prediction-set efficiency (length/size) under coverage or conditional-validity constraints (Kiyani et al., 2024).

Adversarial attacks on CP. While CP offers coverage guarantees, it remains vulnerable to adversarial perturbations in various phases. Adversarial attacks can be categorized based on the stage of the CP pipeline they target: training, calibration, or test. At training, poisoning attacks can bias the underlying model's scoring function, leading to significantly large prediction sets (Li et al., 2024b). At calibration, specific forms of label noise (Penso & Goldberger, 2024) or distribution shift between calibration and test distributions (Cauchois et al., 2024) can bias the calibrated conformal threshold, leading to miscoverage or larger prediction sets. At testing, adversarial perturbations are crafted to shift individual test input across the decision boundary, causing the CP set to exclude the ground truth label or excessively expand in size (Gendler et al., 2021; Ghosh et al., 2023; Liu et al., 2024).

Robust CP. Previous work has adapted CP to various forms of data imperfection, including covariate shift (Tibshirani et al., 2019), broader distribution shifts (Gibbs & Candes, 2021), and label noise (Sesia et al., 2025). More recently, growing concerns over adversarial attacks have motivated the development of CP with provable robustness guarantees. Existing works mainly focus on test-time perturbations (Gendler et al., 2021; Ghosh et al., 2023; Yan et al., 2024; Zargarbashi et al., 2024; Jeary et al., 2024). For training-time defenses, ensemble aggregation (Zhou, 2025; Dietterich, 2000) is a general strategy, with partition-and-aggregation (Levine & Feizi, 2021; Rezaei et al., 2023; Wang et al., 2022) providing provable robustness by limiting the influence of any single corrupted sample. RPS (Scholten & Günnemann, 2025) applies this principle to CP by aggregating prediction scores across submodels and provides pointwise certificates under arbitrary data poisoning. In contrast, JointCP employs a two-level aggregation framework tailored to CP and establishes both pointwise certificates and distribution-free coverage guarantees under arbitrary data poisoning.

6 Conclusion

We propose JointCP, a provably robust conformal prediction framework that provides certified guarantees against *arbitrary* poisoning on the training and calibration data in CP. At a high level, JointCP leverages ensemble-based aggregation over disjoint training and calibration subsets to bound the influence of poisoned samples. This design enables the derivation of both distribution-free and data-dependent certified coverage and, importantly, certified prediction set size under arbitrary data poisoning attacks. Our results on multiple image classification datasets validate that JointCP consistently maintains high coverage and compact prediction sets under strong poisoning attacks.

AI USE STATEMENT

Generative AI tools were used for language editing, wording refinement, and organizational suggestions during manuscript preparation. All technical content, theoretical derivations, experimental design and reported results were verified by the authors.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2022.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 2021.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, 2012.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119 (548):3033–3044, 2024.

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

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. In *Neural Information Processing Systems*, 2022.

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

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, 2023.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Neural Information Processing Systems*, 2021.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Neural Information Processing Systems*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2016.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy (SP)*, 2018.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. In *Neural Information Processing Systems*, 2024.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025.

Shayan Kiyani, George J Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

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.

Shuo Li et al. Traq: Trustworthy retrieval augmented question answering via conformal prediction. In *ACL*, 2024a.

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*, 2024b.

Ziquan Liu, Yufei Cui, Yan Yan, Yi Xu, Xiangyang Ji, Xue Liu, and Antoni B Chan. The pitfalls and promise of conformal inference under adversarial attacks. In *International Conference on Machine Learning*, 2024.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, 2024.

Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011. URL https://api.semanticscholar.org/CorpusID:16852518.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, 2002.

Sangdon Park, Osbert Bastani, and Taesoo Kim. {ACon^ 2}: Adaptive conformal consensus for provable blockchain oracles. In *USENIX Security Symposium*, 2023.

Coby Penso and Jacob Goldberger. A conformal prediction score that is robust to label noise. *arXiv preprint arXiv:2405.02648*, 2024.

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 Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

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

Matteo Sesia, YX Rachel Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):796–815, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Neural Information Processing Systems*, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *International Conference on Learning Representations*, 2024.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. *arXiv preprint arXiv:2510.08748*, 2025.

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

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

A PREREQUISITES

Before deriving the theoretical results of `JointCP` against joint training and calibration data poisoning, we first analyze two special poisoning scenarios: *(i) only training data poisoning*; and *(ii) only calibration data poisoning*. Accordingly, we denote `JointCP-T` and `JointCP-C` as the variant of `JointCP` to defend against scenarios (i) and (ii), respectively.

We first introduce below lemmas.

Lemma 1 (Markov's Inequality). For any non-negative random variable X and any positive number $a > 0$:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{14}$$

Lemma 2. Given a test sample (x^{te}, y^{te}) and let $\mathbb{C}_j(x^{te})$ and $\mathbb{C}_*(x^{te})$ denote the prediction set produced by an individual CP CP_j and by aggregating a set of m CPs on the clean data. Then the coverage guarantee for individual CP and the aggregate CP are respectively:

$$\Pr[y^{te} \in \mathbb{C}_j(x^{te})] \geq 1 - \epsilon. \tag{15}$$

$$\Pr[y^{te} \in \mathbb{C}_*(x^{te})] \geq 1 - \frac{\alpha\epsilon}{\alpha - 1}. \tag{16}$$

Proof. First, each individual CP satisfies $\Pr[y^{te} \in \mathbb{C}_j(x^{te})] \geq 1 - \epsilon$ due to the CP property.

Let B be the random variable representing the number of individual CPs for which the true label is not in the prediction set, i.e., $B = \sum_{j=1}^{m} \mathbb{I}(y^{te} \notin \mathbb{C}_j(x_{n+1}))$. The probability that at least $1 - 1/\alpha$ fraction of the $\{CP_j\}_{j=1}^{m}$ fail to include the true label is $\mathbb{P}\left(B \geq (1 - 1/\alpha)m\right)$. By Markov's inequality, we have

$$\mathbb{P}\left(B \geq (1 - 1/\alpha)m\right) \leq \frac{\mathbb{E}[B]}{(1 - 1/\alpha)m}. \tag{17}$$

By Equation (15), each individual CP guarantees that $\mathbb{P}(y^{te} \notin \mathbb{C}_j(x_{n+1})) \leq \epsilon$. Since each $\mathbb{I}(y^{te} \notin \mathbb{C}_j(x_{n+1}))$ is a Bernoulli random variable with success probability at most ϵ, we have

$$\mathbb{E}[B] = \sum_{j=1}^{m} \mathbb{E}[\mathbb{I}(y^{te} \notin \mathbb{C}_j(x_{n+1}))] \leq m\epsilon. \tag{18}$$

Thus, the probability that the ensemble conformal predictor fails to include the true label is bounded by:

$$\begin{aligned}\mathbb{P}\left(y^{te} \notin \mathbb{C}_*(x_{n+1})\right) =& \mathbb{P}\left(B \geq (1 - \frac{1}{\alpha})m\right) \\ \leq& \frac{m\epsilon}{(1 - \frac{1}{\alpha})m} = \frac{\alpha\epsilon}{\alpha - 1}.\end{aligned}$$

Hence, the coverage guarantee for the ensemble CP is:

$$\mathbb{P}\left(y^{te} \in \mathbb{C}_*(x_{n+1})\right) \geq 1 - \frac{\alpha\epsilon}{\alpha - 1}.$$

□

The main theories of `JointCP-T` are as follows:

Theorem A.1 (Coverage Guarantee for `JointCP-T` against Training Data Poisoning)**.** *Consider a test sample (x^{te}, y^{te}) and an aggregated CP of m individual CPs with significance level $\epsilon \in (0, 1)$. Suppose the number of poisoned training samples c_{tr} is at most $(1/\beta - 1/\alpha)m$, for some $1 < \beta < \alpha$. Then, the prediction set $\mathbb{C}'_*(x^{te})$ given by this aggregated CP satisfies the coverage guarantee:*

$$\Pr[y^{te} \in \mathbb{C}'_*(x^{te})] \geq 1 - \frac{\beta\epsilon}{\beta - 1}, \tag{19}$$

Proof. We consider `JointCP-T` constructed from the same set of individual predictors $\{CP_j\}_{j=1}^m$, using an aggregation parameter $\beta > 1$. Let $\mathbb{C}_*(x^{te})$ denote the prediction set produced by `JointCP` on the clean data. By Equation (16), this clean prediction set satisfies the coverage guarantee

$$\Pr\left[y^{te} \in \mathbb{C}_*(x^{te})\right] \geq 1 - \frac{\beta\epsilon}{\beta - 1}.$$

We then show that $\mathbb{C}_*(x^{te}) \subseteq \mathbb{C}'_*(x^{te})$, i.e., the post-attack aggregated prediction set still contains all labels from the clean prediction set. Thus, we have $\mathbb{C}'_*(x^{te})$ satisfies the same coverage guarantee as $\mathbb{C}_*(x^{te})$.

For any label $y \in \mathbb{C}_*(x^{te})$, it must hold that

$$v_y = \sum_{j=1}^{m} \mathbb{I}(y \in \mathbb{C}_j(x^{te})) \geq \frac{m}{\beta},$$

where $\mathbb{C}_j(x^{te})$ is the prediction set given by each individual predictor CP_j. Suppose c_{tr} satisfied

$$c_{tr} \leq \left(\frac{1}{\beta} - \frac{1}{\alpha}\right) m.$$

Then, at most c_{tr} individual predictors can be corrupted. Let $\mathbb{C}'_j(x^{te})$ be the corresponding post-attack prediction set from predictor CP_j. For this label $y \in \mathbb{C}_*(x^{te})$, the post-attack vote satisfies

$$v'_y = \sum_{j=1}^{m} \mathbb{I}(y \in \mathbb{C}'_j(x^{te})) \geq \frac{m}{\beta} - c_{tr} \geq \frac{m}{\alpha}.$$

Therefore, this label y is also contained in the post-attack aggregated prediction set $\mathbb{C}'_*(x^{te})$. □

Theorem A.1 gives a worst-case bound on the number of training samples that can be attacked while preserving the clean label set. The next theorem derives a data-dependent, label-wise bound, providing a finer-grained certification on coverage.

Theorem A.2 (Label-wise Training Poisoning Tolerance w.r.t. Number of Poisoned Training Samples)**.** *Given m individual CPs and a test sample x^{te}, consider any label $y \in \mathcal{Y}$. Define*

$$V^{te}_{\in,y} = \sum_{j=1}^{m} \mathbb{I}(y \in \mathbb{C}_j(x^{te})), V^{te}_{\notin,y} = \sum_{j=1}^{m} \mathbb{I}(y \notin \mathbb{C}_j(x^{te})),$$

as the number of individual CPs that include and exclude label y in their prediction sets for a given x^{te} under the clean data, respectively. If c_{tr} satisfies:

$$c_{tr} < c^*(y) := \frac{|(\alpha - 1)V^{te}_{\in,y} - V^{te}_{\notin,y}|}{\alpha}, \tag{20}$$

then the status of label y in the aggregated prediction set is provably unchanged after the poisoned attack. That is,

$$y \in \mathbb{C}_*(x^{te}) \iff y \in \mathbb{C}'_*(x^{te}),$$
$$y \notin \mathbb{C}_*(x^{te}) \iff y \notin \mathbb{C}'_*(x^{te}).$$

Proof. Let $\widetilde{V}^{te}_{\in,y}$ and $\widetilde{V}^{te}_{\notin,y}$ be the corresponding vote counts after training data has been perturbed by c_{tr} poisoned samples. Since each poisoned sample can affect the output of at most one model, we have

$$V^{te}_{\in,y} - c_{tr} \leq \widetilde{V}^{te}_{\in,y} \leq V^{te}_{\in,y} + c_{tr},$$
$$V^{te}_{\notin,y} - c_{tr} \leq \widetilde{V}^{te}_{\notin,y} \leq V^{te}_{\notin,y} + c_{tr}.$$

The status of label y in the aggregated prediction set remains unchanged if the majority vote condition is preserved under perturbation, i.e.,

$$\widetilde{V}^{te}_{\in,y} > \frac{m}{\alpha} = \frac{\widetilde{V}^{te}_{\in,y} + \widetilde{V}^{te}_{\notin,y}}{\alpha} \iff (\alpha-1)\widetilde{V}^{te}_{\in,y} > \widetilde{V}^{te}_{\notin,y}, \qquad \text{if } (\alpha-1)V^{te}_{\in,y} > V^{te}_{\notin,y},$$
$$(\alpha-1)\widetilde{V}^{te}_{\in,y} \le \widetilde{V}^{te}_{\notin,y}, \qquad \text{otherwise.}$$

To ensure this holds in the worst-case perturbation scenario where the attacker maximally increases one vote count and decreases the other, we require

$$\begin{cases} (\alpha-1)(V^{te}_{\in,y} - c_{tr}) > V^{te}_{\notin,y} + c_{tr} & \text{if } (\alpha-1)V^{te}_{\in,y} > V^{te}_{\notin,y}, \\ (\alpha-1)(V^{te}_{\in,y} + c_{tr}) \le V^{te}_{\notin,y} - c_{tr} & \text{otherwise.} \end{cases}$$

Both cases are satisfied if and only if $|(\alpha-1)V^{te}_{\in,y} - V^{te}_{\notin,y}| > \alpha c_{tr}$, which is equivalent to the certified bound

$$c_{tr} < \frac{|(\alpha-1)V^{te}_{\in,y} - V^{te}_{\notin,y}|}{\alpha}.$$

□

The main theories for `JointCP-C` are as follows:

Theorem A.3 (Coverage Guarantee for `JointCP-C` against Calibration Data Poisoning)**.** *Consider a new test sample (x^{te}, y^{te}) and an aggregated CP of m above learnt individual CPs with significance level $\epsilon \in (0,1)$. Then, the prediction set $\mathbb{C}'_*(x^{te})$ given by this aggregated CP against calibration data poisoning satisfies:*

$$\Pr[y^{te} \in \mathbb{C}'_*(x^{te})] > 1 - \frac{\alpha}{\alpha-1}(\epsilon + \frac{c_{cal}}{|\mathbb{D}_{cal}|}). \tag{21}$$

Our following analysis only requires **exchangeability between the test sample and the clean (pre-attack) calibration samples**, which is the standard assumption in Conformal Prediction. We do **not** assume that exchangeability holds after the calibration data has been poisoned. Specifically, the rank k is defined relative to the **clean** calibration scores. The poisoning step is handled as a separate perturbation: since the adversary modifies at most c_{cal} samples, the calibration count (and thus the resulting p-value) can shift by at most $c_{cal}/(n_{cal}+1)$. Our proof relies on bounding this shift rather than assuming any distributional symmetry in the poisoned set.

Proof. Suppose each individual CP_j is poisoned with c_j calibration samples, where the total number of poisoned samples is $c_{cal} = \sum_{j=1}^{m} c_j$. For simplicity, we illustrate the argument using CP_1; the same reasoning applies to the remaining individual CPs in `JointCP-C`.

Assume that the conformal scores of calibration samples for CP_1 are sorted in ascending order, i.e.,

$$s_1(x_{n_{tr}+1}, y_{n_{tr}+1}) < \cdots < s_1(x_{n_{tr}+\bar{n}_{cal}}, y_{n_{tr}+\bar{n}_{cal}}).$$

Suppose $s_1(x^{te}, y^{te})$ of the test sample would be inserted at rank k in this sequence, where $\bar{n}_{cal} = \frac{|\mathbb{D}_{cal}|}{m}$ is the size of the calibration set of individual CP.

By definition of p-value, let $p'_1(x,y)$ denote the p-value produced by CP_1 after the attack. Accounting for the effect of poisoning c_1 calibration samples, the perturbed p-value $p'_1(x^{te}, y^{te})$ in poisoned calibration setting satisfies:

$$\frac{\bar{n}_{cal} - k + 2 - c_1}{\bar{n}_{cal}+1} \le p'_1(x^{te}, y^{te}) \le \frac{\bar{n}_{cal} - k + 2 + c_1}{\bar{n}_{cal}+1}.$$

Therefore, the label y^{te} is included in the prediction set $\mathbb{C}'_1(x^{te})$ after the attack if and only if $p'_1(x^{te}, y^{te}) > \epsilon$, which is guaranteed when

$$\frac{\bar{n}_{cal} - k + 2 - c_1}{\bar{n}_{cal}+1} > \epsilon.$$

Due to exchangeability of the calibration samples $(x_{n_{tr}+1}, y_{n_{tr}+1}), \ldots, (x_{n_{tr}+\bar{n}_{cal}}, y_{n_{tr}+\bar{n}_{cal}})$ and the test sample (x^{te}, y^{te}), we have

$$\begin{aligned}\Pr[y^{te} \in \mathbb{C}'_1(x^{te})] &= \frac{\lceil (\bar{n}_{cal}+1)(1-\epsilon) - c_1 \rceil}{\bar{n}_{cal}+1} \\ &\geq 1 - \epsilon - \frac{c_1}{\bar{n}_{cal}+1}. \end{aligned} \tag{22}$$

Next, we analyze the aggregated CP.

Define $B' = \sum_{j=1}^{m} \mathbb{I}(y^{te} \notin \mathbb{C}'_j(x^{te}))$ as the number of individual CPs that exclude label y^{te}.

Recall the Markov's inequality: for any non-negative random variable X and any positive number $a > 0$:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{23}$$

By plugging $X = B'$ and $a = m - \frac{m}{\alpha}$ into Equation (23) gives

$$\Pr[B' \geq m - \frac{m}{\alpha}] \leq \frac{\mathbb{E}[B']}{m - \frac{m}{\alpha}}.$$

From Equation (22), the expected value satisfies

$$\mathbb{E}[B'] = \sum_{j=1}^{m} \mathbb{E}[\mathbb{I}(y^{te} \notin \mathbb{C}'_j(x^{te}))] \leq m\epsilon + \frac{c_{cal}}{\bar{n}_{cal}+1}$$

Thus, we have

$$\begin{aligned}&\Pr\left[y^{te} \notin \mathbb{C}'_*(x^{te})\right] = \Pr[B' \geq m - \frac{m}{\alpha}] \\ &\leq \frac{m\epsilon + \frac{\sum_{j=1}^{m} c_j}{\bar{n}_{cal}+1}}{m - \frac{m}{\alpha}} = \frac{\alpha}{\alpha - 1}(\epsilon + \frac{c_{cal}}{(\bar{n}_{cal}+1)m}).\end{aligned}$$

Therefore, the coverage guarantee for the aggregated CP after the attack is

$$\begin{aligned}&\Pr[y^{te} \in \mathbb{C}'_*(x^{te})] \geq 1 - \frac{\alpha}{\alpha - 1}(\epsilon + \frac{c_{cal}}{(\bar{n}_{cal}+1)m}) \\ &> 1 - \frac{\alpha}{\alpha - 1}(\epsilon + \frac{c_{cal}}{\bar{n}_{cal} m}) = 1 - \frac{\alpha}{\alpha - 1}(\epsilon + \frac{c_{cal}}{|\mathbb{D}_{cal}|}),\end{aligned}$$

completing the proof. □

Theorem A.4 (Label-wise Calibration Poisoning Tolerance)**.** *Given m individual CPs and a test sample x^{te}, consider any label $y \in \mathcal{Y}$. If the number of poisoned calibration samples c_{cal} satisfies:*

$$G^{te}(y, c_{cal}) \leq c^*(y), \quad c^*(y) := \frac{|(\alpha - 1)V^{te}_{\in,y} - V^{te}_{\notin,y}|}{\alpha}, \tag{24}$$

then the status of label y in the aggregated prediction set is provably unchanged after the poisoned attack. That is,

$$\begin{aligned}y \in \mathbb{C}_*(x^{te}) &\Longleftrightarrow y \in \mathbb{C}'_*(x^{te}), \\ y \notin \mathbb{C}_*(x^{te}) &\Longleftrightarrow y \notin \mathbb{C}'_*(x^{te}),\end{aligned}$$

here, $G^{te}(y, c_{cal}) = \max\left\{t \in [m_2] : \sum_{j=1}^{t} g^{te}_{(j)}(y) \leq c_{cal}\right\}$ and $g^{te}_j(y) = \lfloor |(1-\epsilon)\bar{n}_{cal} - \epsilon - k + 2| \rfloor + 1$. k denotes the rank at which $s_j(x^{te}, y)$ would be inserted in the sorted list and $\bar{n}_{cal} = |\mathbb{D}_{cal}|/m_2$.

Proof. By definition, $G^{te}(y, c_{cal})$ represents the maximum number of individual CPs whose predictions on label y can be altered when up to c_{cal} calibration samples are poisoned. These correspond to the CPs with the smallest values $g^{te}_j(y)$. In the worst case, an attacker perturbs the calibration samples to flip the predictions of these $G^{te}(y, c_{cal})$ most vulnerable CPs. According to Theorem A.2, when $G^{te}(y, c_{cal}) \leq c^*_y$, the majority vote condition is preserved even under the worst-case perturbation. Thus, y's status in the final prediction set remains unchanged. □

B PROOFS

B.1 PROOF OF THEOREM 3.1

Proof. Based on Equation (16), `JointCP-C` aggregates m_2 calibration splits for each model, yielding coverage at least $1-\epsilon^*$, where $\epsilon^* = \frac{\alpha}{\alpha-1}\epsilon$. The symbol ϵ^* can be viewed as the effective miscoverage of each individual CP. By utilizing the coverage guarantee for `JointCP-T` in Equation (16), we can obtain the final coverage at least $1-\frac{\alpha\epsilon^*}{\alpha-1}$.

Hence, the coverage guarantee for the aggregated CP is:

$$\Pr[y^{te} \in \mathbb{C}_*(x^{te})] \geq 1 - \frac{\alpha^2\epsilon}{(\alpha-1)^2}.$$

□

B.2 PROOF OF THEOREM 3.2

Proof. By Theorem A.3, for each CP with a clean training subset, `JointCP-C` that aggregates the m_2 calibration splits yields coverage at least $1-\epsilon'$, where $\epsilon' = \frac{\alpha}{\alpha-1}(\epsilon + \frac{c_{cal}}{|\mathbb{D}_{cal}|})$. Then we can treat ϵ' as the effective miscoverage of each individual CP corresponding to the clean training subset. By assumption, at most $c_{\text{tr}} \leq (\frac{1}{\beta} - \frac{1}{\alpha})m_1$ training splits are poisoned; hence at least $m_1 - c_{\text{tr}}$ of the m_1 individual CPs satisfy the marginal coverage lower bound $1-\epsilon'$. Applying the coverage guarantee for `JointCP-T` in Theorem A.1 and the above constraint on c_{tr} yields the final coverage at least $1-\frac{\beta\,\epsilon'}{\beta-1}$. Further, the tightest bound can be obtained by setting $\beta = \frac{1}{1/\alpha + c_{tr}/m_1}$, producing the coverage guarantee as shown in Equation (12).

□

B.3 PROOF OF THEOREM 3.3

Proof. Fix a label $y \in \mathcal{Y}$. For each training subset $i \in [m_1]$, consider its `JointCP-C` instance. By Theorem A.4, the inclusion decision for y in this `JointCP-C` can be flipped by calibration poisoning of budget c only if $G_i^{te}(y,c) > c_y^*(i)$. Hence, calibration poisoning can change the local decision for at most $F_{\text{cal}} = \left|\{\, i \in [m_1] : G_i^{te}(y,c) > c_y^*(i) \,\}\right|$ training splits.

Independently, training poisoning can arbitrarily alter the decisions of at most c_{tr} training splits. Therefore, across the m_1 votes aggregated by `JointCP-T`, the total number of local votes that can change is at most $c_{\text{tr}} + F_{\text{cal}}$.

Let c_y^* denote the training-layer vote margin before the attack. By Theorem A.2, if

$$c_{\text{tr}} + F_{\text{cal}} \;<\; c_y^*,$$

then the vote for y in `JointCP-T` cannot be changed. □

C COMPLEXITY ANALYSIS

We analyze the computational complexity of `JointCP` across all steps.

Model training. The full training set $\mathbb{D}_{tr}$ is partitioned into m_1 disjoint splits, each of size $|\mathbb{D}_{tr}|/m_1$. Assuming the training cost of a model is linear with respect to the number of training samples, training a model on a subset incurs a complexity of $\mathcal{O}(|\mathbb{D}_{tr}|/m_1)$. Consequently, the total cost for training all m_1 models sequentially is: $m_1 \times \mathcal{O}\left(\frac{|\mathbb{D}_{tr}|}{m_1}\right) = \mathcal{O}(|\mathbb{D}_{tr}|)$.

Notably, this total complexity of training is theoretically equivalent to training a single model on the full dataset $\mathbb{D}_{tr}$ in vanilla CP. Furthermore, since the m_1 models are independent, they can be trained in parallel.

Model aggregation/prediction. Evaluating the aggregation requires the predictions from all m_1 trained models. This incurs a computational complexity of $\mathcal{O}(m_1 T_\pi)$, where T_π denotes the cost of a single forward pass for a base model.

Table 3: Runtime comparison for standard prediction on SVHN.

Method	Model inference (s)	Pred. set construction (s)
CP	7.48	0.04
`JointCP`	632.91	0.24

Prediction set construction. This step involves both predictions on individual and aggregated CPs.

- *On individual CPs:* Computing the $m_1 \times m_2$ individual p-values $\{p_{i,k}\}$ per label y involves comparing the test score against each calibration subset $\mathbb{D}^k_{cal}$, resulting the operations sum to $\sum_{i=1}^{m_1}\sum_{k=1}^{m_2}|\mathbb{D}^k_{cal}| = m_1|\mathbb{D}_{cal}|$. Across all labels, the total complexity is $\mathcal{O}(m_1|\mathbb{D}_{cal}||\mathbb{Y}|)$.
- *On CP aggregation:* Summing the votes across m_2 calibration splits and then across m_1 training splits incurs $\mathcal{O}(m_1 m_2)$ time, which is far smaller than $\mathcal{O}(m_1|\mathbb{D}_{cal}||\mathbb{Y}|)$ as often $|\mathbb{D}_{cal}| \gg m_2$.

Thus the computational complexity in this step is dominated by $\mathcal{O}(m_1|\mathbb{D}_{cal}||\mathbb{Y}|)$.

Summary. The total time complexity of `JointCP` is thus $\mathcal{O}(m_1 T_\pi + m_1|\mathbb{D}_{cal}||\mathbb{Y}|) + O(|\mathbb{D}_{tr}|)$.

Runtime. Table 3 reports the runtime of standard prediction on SVHN with JointCP, including model inference and prediction-set construction, but excluding the data-specific certification in Theorem 3.3. Since the offline training stage has the same Big-O complexity as vanilla CP under the linear training-cost assumption discussed above, we focus on the runtime of model inference and prediction set construction. The results are consistent with our complexity analysis. The dominant computational cost of `JointCP` arises from inference with the $m_1 = 100$ base models: the total inference time (632.91 s) is approximately $100\times$ that of vanilla CP (7.48 s). In contrast, prediction set construction is lightweight for both methods, with comparable runtimes of 0.04 s for CP and 0.24 s for `JointCP`. These results indicate that the practical overhead of joint calibration and aggregation is negligible relative to model inference.

Compute resources. All experiments were run on a Python 3 Google Compute Engine backend in Google Colab with an NVIDIA A100 GPU. The machine has 40GB GPU memory and 83.5GB system RAM. For one certified robustness run under a fixed hyperparameter setting, sweeping the poisoning budget from $c = 1$ to $c = 100$ takes approximately 4.5 hours on SVHN and CIFAR10, and approximately 22 hours on CIFAR100. CIFAR100 requires substantially longer runtime due to its larger label space, which increases the cost of evaluating label-wise certified prediction sets. Separately, for the full certification experiment on CIFAR10, model inference and score computation over the data used for calibration and testing take approximately 90 minutes, while the subsequent Theorem 3.3 certification sweep over all 2,500 test samples, candidate labels, and 100 poisoning budgets takes approximately 3 hours.

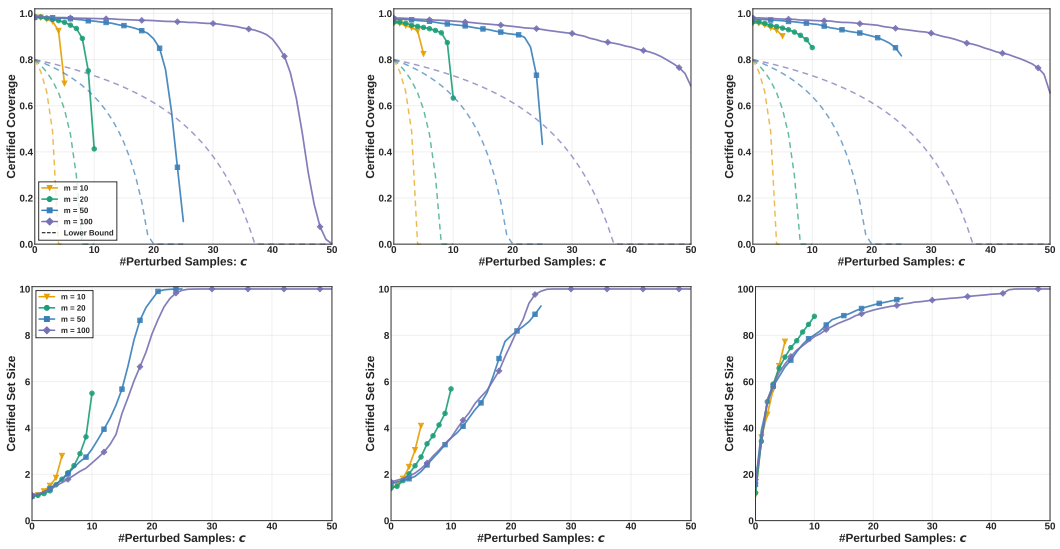
D Experiment

D.1 Impact of m

The number of splits m_1, m_2 controls the trade-off between robustness and per-split sample size. We conduct additional experiments with varying $m = m_1 = m_2$ (all other settings follow the default setup). Average Accuracy of the m_1 finetuned models are in Table 6 and impact of m on `JointCP` are shown in Figure 5. The results demonstrate the well-known robustness–accuracy trade-off: larger m provides stronger certified robustness, as corrupted data affects fewer sub-models, while smaller m allows more samples per sub-model, leading to higher accuracy.

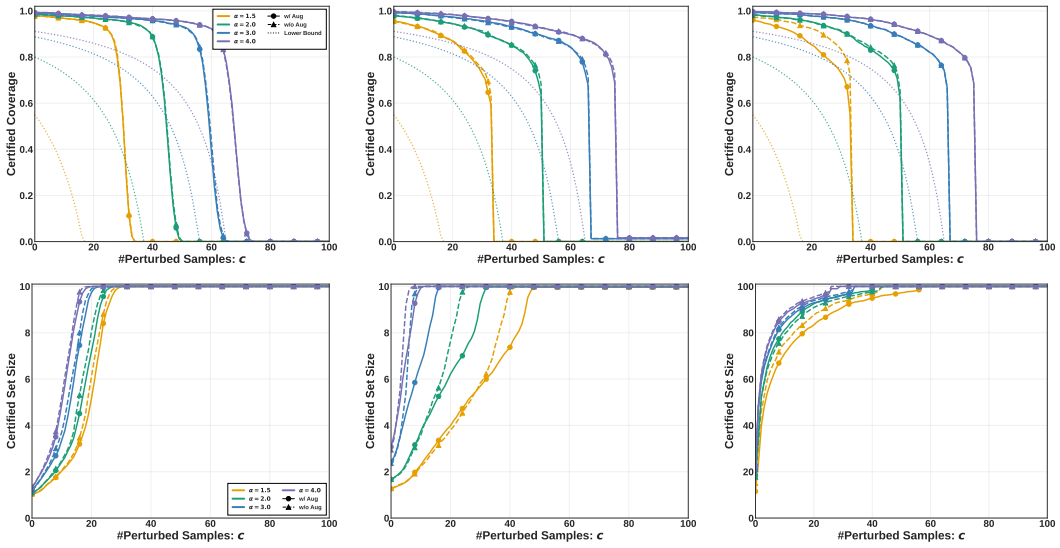
D.2 Extension to other CP variants

We additionally evaluate HPS on SVHN. The results in Table 4 show the same overall robustness trend as APS, confirming that the effectiveness of JointCP is not tied to a specific CP. For instance, with the default setting of $m_1 = m_2 = 100$, $\alpha = 2$, $\epsilon = 0.05$ in the clean setting ($c = 0$), JointCP with HPS achieves 98.0% certified coverage with an average set size of 1.05, which is very close to APS (98.0% coverage, size 1.09). Under a corruption budget of $c = 20$, HPS maintains a highly robust certified coverage of 95.3%, mirroring the 96.9% coverage of APS.



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 5: Impact of m on `JointCP` ($\epsilon = 0.05$, $\alpha = 2$, $c_{max} = m/\alpha$). We include the (lower bounded) distribution-free coverage in Theorem 3.2 for reference.



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 6: Impact of training data with augmentation vs without augmentation on `JointCP` ($\epsilon = 0.05$, aug=5). We include the (lower bounded) distribution-free coverage in Theorem 3.2 for reference.

D.3 Size-aware Conformal Training via Data Augmentation

Partitioning the training data into m_1 disjoint subsets improves certified robustness, since each poisoned training sample can affect only a limited number of sub-models. However, it also reduces the number of training samples available to each sub-model, which may hurt the quality of individual predictors. To mitigate this issue, we evaluate whether increasing the training sample size through data augmentation can improve the efficiency of `JointCP`.

`JointCP` can additionally incorporate a size-aware conformal training procedure to suppress large prediction sets by considering the possible perturbations on the training / calibration data. Concretely, for each sub-model π_i, we finetune the classifier head using a mixture of clean and augmented

Table 4: Results of CP variants ($m_1 = m_2 = 100$, $\alpha = 2$, $\epsilon = 0.05$).

CP variants	c = 0		c = 10		c = 20		c = 30		c = 40		c = 50	
	EmpC	AveS	CertC	CertS	CertC	CertS	CertC	CertS	CertC	CertS	CertC	CertS
APS	0.98	1.09	0.98	2.49	0.97	8.06	0.96	10	0.89	10	0	10
HPS	0.98	1.05	0.97	2.55	0.95	5.95	0.92	9.9	0.85	10	0.25	10

Table 5: Results of Tiny ImageNet ($\epsilon = 0.05$, $\alpha = 2$, $m_1 = m_2 = 100$).

	c = 0 (Emp)	c = 1	c = 2	c = 3	c = 4	c = 5
CertC	0.96	0.96	0.96	0.96	0.96	0.96
CertS	4.30	19.99	36.16	49.31	60.67	71.25

perturbed samples, where the *augmentation number* aug denotes the number of augmented variants per clean example. Thus, $\mathrm{aug} = 5$ corresponds to a $1:5$ mixture (one clean view plus five augmented views per sample). Our augmentation pipeline uses geometric transforms (e.g., color jitter and random crop), followed by optional Gaussian noise.

Inspired by Stutz et al. (2022), we optimize a size-aware objective: $\ell = \ell_{\mathrm{CE}} + \ell_{\mathrm{size}}$, where ℓ_{CE} is a cross-entropy loss (in Equation 1) and ℓ_{size} is a differentiable surrogate of the expected prediction set size. For each iteration, we split the mini-batch into two halves: a calibration half and a prediction half. On the calibration half, we compute the conformal threshold τ as the α-quantile of the ground-truth class probabilities, ensuring the target coverage level is met. On the prediction half, we compute a differentiable approximation of the prediction set size. A label k is considered included if its probability p_k exceeds τ. We approximate this inclusion using a sigmoid function: $\ell_{\mathrm{size}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sigma\left(\gamma \cdot (p_{i,k} - \tau)\right)$, where γ is a sharpness parameter and N is the number of samples. This training encourages non-true labels to fall below the CP threshold while preserving classification accuracy, reducing the resulting prediction set size.

Comparison of data augmentation vs. no augmentation. We investigate the effectiveness of our size-aware conformal finetuning. Figure 6 compares `JointCP-T` trained with 5 augmented training samples to training without augmentation. We observe that data augmentation has only a marginal effect on certified coverage, but can noticeably reduce the prediction set size. This effect is particularly evident on CIFAR10 when the numbers of poisoned training and calibration samples are relatively large (e.g., $c_{tr} = c_{cal} > 20$). These results suggest that the size-aware loss ℓ_{size}, when combined with diverse augmented samples, can improve effectiveness without compromising certified coverage.

Impact of the number of augmentations on `JointCP`. We further study the effect of the number of augmentations, $\mathrm{aug} \in \{0, 1, 5, 10\}$, on `JointCP`. Figure 7 reports results on CIFAR10, which exhibits the largest reduction in prediction set size among the three datasets. The certified coverage curves largely overlap across different augmentation levels, demonstrating that size-aware conformal training does not affect coverage. In contrast, the certified prediction set size decreases as the amount of augmented data increases, but stabilizes when the augmentation level becomes sufficiently large, e.g., $\mathrm{aug} = 5$ and 10 have close prediction set size across different perturbed samples.

D.4 Generalize JointCP to larger vision dataset/classes.

To evaluate whether `JointCP` scales to larger label spaces, we conduct additional experiments on Tiny ImageNet Le et al. (2015) with 200 classes. Specifically, we use the ImageNet-1K pretrained model from Torchvision and finetune with 100,000 samples from Tiny ImageNet to obtain the base model. For the `JointCP` pipeline, we use 50,000, 2,500, and 500 samples for finetuning, calibration, and testing, respectively, with hyperparameters set to $\epsilon = 0.05$, $\alpha = 2$, $m_1 = m_2 = 100$. The results in Table 5 under the default setting show a similar trend.

D.5 Calibration-time Augmentation

Similarly, partitioning the calibration data into m_2 disjoint subsets reduces the number of calibration samples available to each individual CP. Although the coverage guarantee remains valid under the

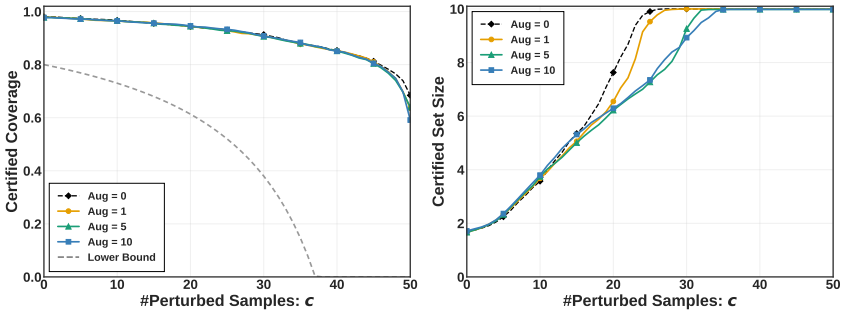


Figure 7: Impact of training-time augmentation for `JointCP` on CIFAR10 ($\epsilon = 0.05$, $\alpha = 2.0$).

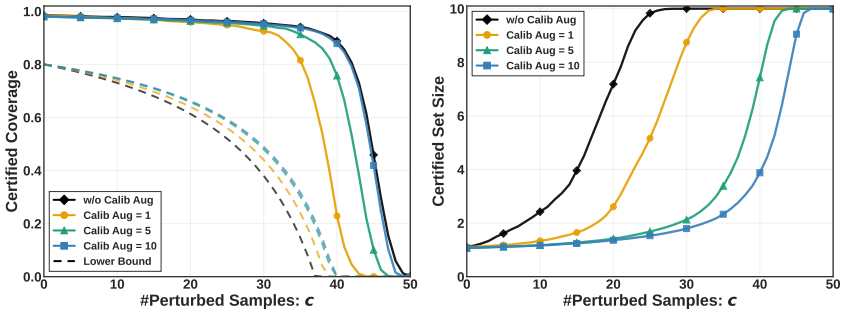


Figure 8: Impact of calibration-time augmentation on `JointCP` in SVHN ($\epsilon = 0.05$, $train_{aug} = 5$, $\alpha = 2$, $c_{max} = m/\alpha$). We include the (lower bounded) distribution-free coverage in Theorem 3.2 for reference.

Table 6: Average accuracy of the m finetuned models.

Dataset	$m = 1$	$m = 10$	$m = 20$	$m = 50$	$m = 100$
SVHN	0.96	0.95	0.95	0.95	0.94
CIFAR10	0.90	0.88	0.88	0.88	0.87
CIFAR100	0.71	0.67	0.66	0.65	0.64

stated assumptions, a small calibration split can lead to less stable conformal scores and larger certified prediction sets. To test whether increasing the effective calibration sample size improves efficiency, we additionally evaluate calibration-time data augmentation.

Here, our augmentation uses geometric transforms including color jitter and random crop, followed by optional Gaussian noise. In this experiment, the submodels are trained with augmentation strength 5, while the augmentation levels refer to calibration-time augmentation. The results in the default setting ($\epsilon = 0.05$, $\alpha = 2$, $m_1 = m_2 = 100$) are shown in Figure 8. We observe that augmentation maintains the certified coverage, while significantly reducing the certified prediction set size. For example, with 10 augmented views per calibration sample, the certified prediction set size decreases from 7.18 to 1.36 when 20 training samples and 20 calibration samples are arbitrarily perturbed.

D.6 Fixed Pretrained Backbone without Fine-tuning

We evaluate a fixed pretrained backbone without downstream fine-tuning, with $\alpha = 2$, $\epsilon = 0.05$, $m_1 = 1$, $m_2 = 100$, $c_{\text{tr}} = 0$, and varying $c_{\text{cal}} > 0$. The fixed pretrained model obtains clean empirical coverage 0.965 with an average prediction-set size of 1.053. Its certified results are in Table 7. As the calibration poisoning budget increases, certified coverage decreases and certified set size increases, revealing the trade-off between coverage and set size under calibration poisoning.

Table 7: Results of fixed pretrained backbone without fine-tuning.

#Perturbed Samples: c_{cal}	10	20	30	40	50
CertC	0.96	0.94	0.92	0.90	0.89
CertS	1.27	1.46	3.84	7.94	10.00

E LIMITATION

This work provides certified coverage guarantees for conformal prediction under arbitrary poisoning of training and calibration samples. We discuss several limitations and potential future work.

First, our theoretical guarantees focus on marginal coverage, which is the standard and most widely studied validity notion in conformal prediction literature. While marginal coverage provides distribution-free guarantees over the test distribution, it does not imply conditional coverage for every subpopulation or individual test point. Extending the proposed partition-and-aggregation framework to stronger notions such as group-conditional or approximate conditional coverage under poisoning is an important direction for future work.

Second, `JointCP` inherits the robustness-efficiency trade-off common to certified defenses. Increasing the number of training and calibration splits improves certified robustness because each poisoned sample can affect only a limited number of individual predictors or calibration scores. However, larger numbers of splits also reduce the number of samples available to each sub-model or calibration split, which may increase prediction set size or reduce the accuracy of individual sub-models when data are limited. Our experiments suggest that data augmentation can mitigate this issue.

Third, `JointCP` introduces additional inference cost compared with standard CP because the final prediction set aggregates the outputs of multiple models. Although training is modular and can be parallelized across splits, inference through multiple base models can be the dominant overhead. The prediction-set construction itself remains lightweight.

Finally, our empirical evaluation focuses primarily on image classification benchmarks. Although the theory is model-agnostic and does not rely on dataset-specific properties, further experiments on additional modalities, larger-scale tasks, and different conformal score functions would strengthen the empirical validation of the framework.