Game-Theoretic Defenses for Adversarially Robust Conformal Prediction

Anonymous authors
Paper under double-blind review

Abstract

Adversarial attacks pose major challenges to the reliability of deep learning models in safety-critical domains such as medical imaging and autonomous driving. In such high-stakes applications, providing reliable uncertainty quantification alongside adversarial robustness becomes crucial for safe deployment. Although conformal prediction can provide certain guarantees for model performance under such conditions, unknown attacks may violate the exchangeability assumption, resulting in the loss of coverage guarantees or excessively large predictive uncertainty. To address this, we propose a synergistic framework that integrates conformal prediction with game-theoretic defense strategies by modeling the adversarial interaction as a discrete, zero-sum game between attacker and defender. Our framework yields a Nash Equilibrium defense strategy, which we prove maintains valid coverage while minimizing the worst-case prediction set size against an optimal adversary operating within the defined attack space. Experimental results on CIFAR-10, CIFAR-100, and ImageNet further demonstrate that, under Nash equilibrium, defense models within our framework achieve valid coverage and minimal prediction set size. By bridging adversarial robustness and uncertainty quantification from a game-theoretic perspective, this work provides a verifiable defense paradigm for deploying safety-critical deep learning systems, particularly when adversarial distributions are unknown or dynamically evolving but contained within a known attack space.

1 Introduction

The reliability of deep learning technologies is facing systematic challenges posed by adversarial attacks. Such attacks, through carefully crafted subtle input perturbations, can cause medical imaging analysis models to produce fatal misdiagnoses or induce autonomous driving systems to make dangerous decisions, with potential risks that have transcended traditional algorithmic fault tolerance boundaries. Although adversarial training techniques have enhanced model robustness by proactively generating attack samples and have mitigated malicious input threats to some extent, their defense mechanisms still exhibit significant uncertainty and cannot provide verifiable robustness guarantees. Against this backdrop, Conformal Prediction (CP) has emerged as a promising paradigm, with its core advantages lying in distribution-free properties and robust uncertainty quantification capabilities. This data-driven approach generates statistically rigorous confidence intervals and precisely calibrated probability estimates, demonstrating excellent performance in both regression and classification tasks, thereby providing a novel solution for constructing trustworthy deep learning systems.

However, the application of Conformal Prediction (CP) methods faces fundamental limitations, as their effectiveness strictly depends on the exchangeability assumption between training and test sets. This assumption is often difficult to satisfy in real-world application scenarios, particularly under adversarial attack conditions where attackers may employ unknown attack strategies to maliciously perturb input data. Existing research efforts primarily attempt to mitigate the impact of adversarial samples on test data distribution by reconstructing non-conformity scoring functions, but these approaches still face significant methodological bottlenecks. First, the strategy of expanding prediction set sizes to ensure statistical validity directly leads to a significant reduction in the decision utility of prediction results. Second, existing improvement

schemes are mostly optimized for specific attack paradigms and struggle to guarantee coverage generalizability when confronting unknown attack vectors. Furthermore, the engineering implementation of complex non-conformity scoring functions presents technical obstacles, and existing methods exhibit systematic deficiencies in both computational complexity and cross-scenario generalization capabilities, which severely constrains the practical application value of CP methods in adversarial environments.

Therefore, developing a universal conformal prediction framework is essential. Such a framework needs to address three critical challenges: first, establishing a theoretical guarantee system that ensures arbitrary standard non-conformity scoring functions maintain preset coverage probabilities in open-world scenarios; second, optimizing uncertainty quantification strategies to minimize prediction set sizes while preserving prediction reliability; third, enabling the framework to simultaneously achieve effective coverage maintenance and prediction set size minimization when confronting arbitrary unknown attack vectors in adversarial environments.

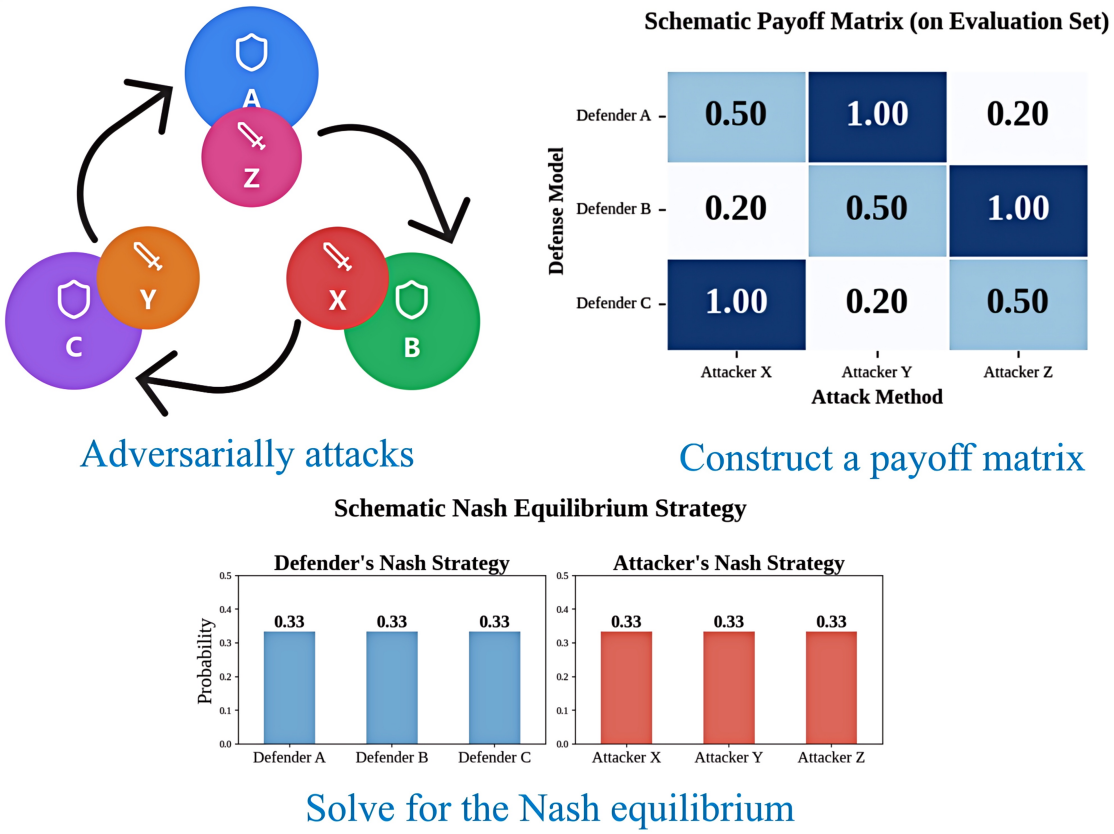

Figure 1: Address adversarial attacks in conformal prediction by constructing a game-theoretic model.

2 Preliminary and Problem Setup

Consider a classifier $f : \mathcal{X} \to \mathbb{R}^K$, which maps an input image $\mathrm{x} \in \mathcal{X}$ to a probability vector $f(\mathrm{x}) \in \mathbb{R}^K$. Here, $\mathcal{X} = [0,1]^d$ represents the image space, where pixel values are normalized and d denotes the image dimension. The integer K signifies the total number of classes, and $y \in [K] := \{1, \ldots, K\}$ denotes the ground truth label for the input x.

2.1 Conformal Prediction

We first train the classifier f on the training set $\mathcal{D}_{\text{train}}$, and then compute nonconformity scores on the calibration set $\mathcal{D}_{\text{cal}}$. On the test set $\mathcal{D}_{\text{test}}$, for each input x, we construct a prediction set $\mathcal{C}(\text{x}) \subseteq [K]$ such that the true label y is contained in $\mathcal{C}(\text{x})$ with probability at least $1-\alpha$, where $\alpha \in (0,1)$ is a user-specified significance level. Formally, we guarantee

$$\mathbb{P}\left(Y \in \mathcal{C}(\text{x})\right) \geq 1-\alpha. \tag{1}$$

In conformal prediction, different nonconformity score functions $s_f(\text{x}, y)$ can lead to varying outcomes, resulting in prediction sets of different sizes. For instance, a choice for the nonconformity score is $s(\text{x}, k) = 1 - f_k(\text{x})$, where k is all the classes. We compute the nonconformity scores $s(\text{x}, k)$ in $\mathcal{D}_{\text{cal}}$. Subsequently, the conformal quantile $\hat{q}$ is determined as the $(1-\alpha)$-th quantile of these scores:

$$q_{1-\alpha} \;=\; \lceil (1+|\mathcal{D}_{\text{cal}}|)(1-\alpha) \rceil\text{-th largest value in } \{s(x_i, y_i)\}_{i \in \mathcal{D}_{\text{cal}}}. \tag{2}$$

This quantile provides the precise value for our prediction threshold, such that $\tau = q_{1-\alpha}$. For a new test input $\text{x} \in \mathcal{D}_{\text{test}}$, the prediction set $\mathcal{C}(\cdot)$ is constructed using this threshold:

$$\mathcal{C}(\text{x}; f, \tau) = \{k \in [K] : s_f(\text{x}, y) \leq q_{1-\alpha}\}. \tag{3}$$

This guarantee fundamentally relies on the assumption that calibration data $\mathcal{D}_{\text{cal}}$ and test data $\mathcal{D}_{\text{test}}$ are exchangeable.

2.2 Violation of Exchangeable Assumption

In adversarial scenarios, attackers manipulate the test set images $\mathcal{D}_{\text{test}}$, transforming the original input samples x into adversarial examples x^{adv}. This transformation clearly disrupts the exchangeability property that underpins conformal prediction, thereby rendering standard conformal prediction unreliable in the presence of adversaries. To address this, there is a pressing need for more robust calibration and model selection strategies that explicitly account for adversarial perturbations.

2.3 Problem Setup

We consider a problem where an attacker selects an optimal attack $g_n^* \in \{g_n\}_{n=1}^m$ from a set of possible attacks, and a defender aims to choose a best classification model $f_j^* \in \{f_j\}_{j=1}^p$ and a threshold $\tau \in \mathbb{R}$ to minimize the expected size of the prediction set for adversarially perturbed test samples. The optimization must ensure that the true label is included in the prediction set with a probability of at least $1-\alpha$. The attacker, within this robust framework, seeks to maximize the prediction set size. The defender formulates the problem as:

$$\begin{aligned} &\min_{f_j^*, \tau \in \mathbb{R}} \max_{g_n^*} \sum_{i \in \mathcal{D}\text{test}} |\mathcal{C}(\text{x}_i^{adv}; f_j^*, \tau)|, \\ &\text{subject to:} \quad \mathbb{P}\left(y_i \in \mathcal{C}(\text{x}_i^{adv}; f_j^*, \tau)\right) \geq 1-\alpha. \end{aligned} \tag{4}$$

3 Game-Theoretic Framework for Adversarially Robust CP

As shown in Figure 1, the optimization problem can be naturally framed as a two-player zero-sum game. As described in the problem setup, the defender aims to minimize the prediction set size while maintaining the coverage guarantee, whereas the attacker seeks to maximize it. This section first addresses how we ensure the critical coverage constraint, $\mathbb{P}\left(y_i \in \mathcal{C}(\text{x}_i^{adv}; f_j^*, \tau)\right) \geq 1-\alpha$, under adversarial conditions, before formulating the full game.

3.1 Robust Coverage Guarantee under Known Adversarial Attacks

To satisfy the coverage constraint in Eq. 4 under adversarial perturbations, we employ a robust calibration strategy. The robust conformal quantile $q^j_{1-\alpha}$ for model f_j is then set as the maximum across all attacks:

$$q^j_{1-\alpha} = \max_{n=1,\ldots,m} q^{j_n}_{1-\alpha} \tag{5}$$

Prediction sets $\mathcal{C}(\mathrm{x}_i^{adv}; f_j, q^j_{1-\alpha})$ are constructed using $q^j_{1-\alpha}$ as the threshold, consistent with Section 2.1. Additionally, we present Theorem 1, with its corresponding proof provided in Appendix A.1.

Theorem 1 (Robust Coverage Guarantee)**.** *For any classifier $f_j \in \{f_j\}_{j=1}^p$ and any adversarial attack $g_n \in \{g_n\}_{n=1}^m$ that was included in the calibration process used to compute $q^j_{1-\alpha}$, the prediction set $\mathcal{C}(\mathrm{x}_i^{adv}; f_j, q^j_{1-\alpha})$ provides a coverage probability of at least $1-\alpha$ for adversarially perturbed test samples x_i^{adv}:*

$$\mathbb{P}\left(y_i \in \mathcal{C}(\mathrm{x}_i^{adv}; f_j, q^j_{1-\alpha})\right) \geq 1-\alpha. \tag{6}$$

3.2 Game Formulation and Nash Equilibrium

Building upon the robust coverage guarantee established in the preceding section, the constrained optimization problem presented in Eq. 4 can be formulated as a two-player zero-sum game. In this game, the defender aims to minimize the prediction set size, while the attacker seeks to maximize it.

3.2.1 Participants and Strategies

This game involves two participants:

- **The Defender:** The defender's strategy is to select a classifier f_j from a finite set of models $\mathcal{F} = \{f_1, \ldots, f_p\}$. For each chosen f_j, its robust conformal quantile $q^j_{1-\alpha}$ will serve as the threshold for constructing the prediction set.
- **The Attacker:** The attacker's strategy is to select an adversarial attack g_n from a finite set of known attack methods $\mathcal{G} = \{g_1, \ldots, g_m\}$.

3.2.2 Payoff Function

The payoff of the game is defined as the expected size of the prediction set. For a defender's chosen f_j and an attacker's chosen g_n, the payoff (denoted as $\mathrm{Payoff}(f_j, g_n)$) is the average prediction set size on adversarially perturbed test samples:

$$\mathrm{Payoff}(f_j, g_n) = \mathbb{E}_{x \sim P_{\mathcal{X}}}\left[|\mathcal{C}(\mathbf{x}^{g_n}; f_j, q^j_{1-\alpha})|\right] \tag{7}$$

where $P_{\mathcal{X}}$ is the distribution of the unperturbed x, $\mathbf{x}^{g_n}$ denotes the adversarial sample generated by applying attack g_n to the original sample $\mathbf{x}$. The defender seeks to minimize this payoff, while the attacker seeks to maximize it.

3.2.3 Nash Equilibrium

Given that this is a finite two-player zero-sum game (with finite strategy sets $\mathcal{F}$ and $\mathcal{G}$), the existence of a Nash Equilibrium in mixed strategies is guaranteed by the fundamental Minimax Theorem v. Neumann (1928).

Let $\mathbf{d}$ be a mixed strategy for the defender (a probability distribution over $\mathcal{F}$) and $\mathbf{a}$ be a mixed strategy for the attacker (a probability distribution over $\mathcal{G}$). The expected payoff under these mixed strategies is $\mathbb{E}[\mathrm{Payoff}(\mathbf{d}, \mathbf{a})]$.

Algorithm 1 Game-Theoretic Framework for Adversarially Robust Conformal Prediction

Input: labeled data $\mathcal{D}$, unlabel test data $\mathcal{D}_{\text{test}}$, set of defensive models $f_j, j = 1, \ldots, p$, set of potential attack functions $g_n, n = 1, \ldots, m$, Coverage probability $1 - \alpha$.
Randomly split $\mathcal{D}$ into $\{(x_i, y_i)\}_{i \in \mathcal{D}_{\text{train}}}$, $\{(x_i, y_i)\}_{i \in \mathcal{D}_{\text{cal}}}$, and $\{(x_i, y_i)\}_{i \in \mathcal{D}_{\text{eval}}}$.
▷ ***Step 1: Train all defensive models.***
for $j = 1, \ldots, p$ **do**
  Train defensive model f_j on the training set $\mathcal{D}_{\text{train}}$.
end for
▷ ***Step 2: Compute robust conformal quantiles for each model.***
for each defensive model $f_j, j = 1, \ldots, p$ **do**
  for each attack methods $g_n, n = 1, \ldots, m$ **do**
    Generate adversarial calibration set: $\mathrm{x}_i^{g_n} = g_n(\mathrm{x}_i, y_i, f_j)_{i \in \mathcal{D}_{\text{cal}}}$.
    Use $f_j(\mathrm{x}_i^{g_n})$ to compute non-conformity scores $\{s_j^n(\mathrm{x}_i^{g_n}, y_i)\}_{i \in \mathcal{D}_{\text{cal}}}$.
    Compute the quantile threshold as: $q_{1-\alpha}^{j_n}$ as the $\lceil (1 + |\mathcal{D}_{\text{cal}}|)(1 - \alpha) \rceil$-th largest score in $\{s_j^n(x_i^{g_n}, y_i)\}_{i \in \mathcal{D}_{\text{cal}}}$.
  end for
  Determine the maximum quantile threshold $q_{1-\alpha}^{j} = \max_{n=1,\ldots,m} q_{1-\alpha}^{j_n}$ for defensive model f_j.
end for
▷ ***Step 3: Estimate the prediction set sizes on the evaluation set.***
for each defensive model $f_j, j = 1, \ldots, p$ **do**
  for each attack methods $g_n, n = 1, \ldots, m$ **do**
    Generate adversarial evaluation set: $\mathrm{x}_i^{g_n} = g_n(\mathrm{x}_i, y_i, f_j)_{i \in \mathcal{D}_{\text{eval}}}$.
    Construct the prediction set using the model's robust quantile: $\mathcal{C}(\mathrm{x}_i^{g_n}; f_j, q_{1-\alpha}^{j})_{i \in \mathcal{D}_{\text{eval}}}$.
    Estimate the payoff by averaging the prediction set sizes over the adversarial evaluation set:
    $\text{Payoff}(f_j, g_n) = \frac{1}{|\mathcal{D}\text{eval}|} \sum i \in \mathcal{D}_{\text{eval}} |\mathcal{C}(\mathrm{x}_i^{g_n}; f_j, q_{1-\alpha}^{j})|$.
  end for
end for
▷ ***Step 4: Find Nash Equilibrium & Construct Robust Conformal Predictor.***
Compute the Nash Equilibrium by solving a Linear Program: $\text{Payoff}(f_j, g_n)$.
Output the Nash Defense Strategy $\mathbf{d}^* = (d_1, \ldots, d_p)$ and Nash Attack Strategy $\mathbf{a}^* = (a_1, \ldots, a_m)$.
Use the Nash Defense Strategy $\mathbf{d}^*$ to define the final robust predictor.
For a new test point x_{new} (which may be subjected to an unknown adversarial attack $g \in \mathcal{G}$ to become $\mathrm{x}_{\text{new}}^{adv}$):
  Randomly sample a defensive model f_j according to the Nash Equilibrium strategy $\mathbf{d}^*$.
  Construct and output the prediction set using the sampled model f_j and its corresponding pre-computed robust quantile $q_{1-\alpha}^{j}$: $C(\mathrm{x}_{\text{new}}^{adv}) = y' : s_j(\mathrm{x}_{\text{new}}^{adv}, y') \leq q_{1-\alpha}^{j}$.

Definition (Nash Equilibrium in Mixed Strategies): A pair of strategies $(\mathbf{d}^*, \mathbf{a}^*)$ constitutes a Nash Equilibrium if, for all other possible mixed strategies $\mathbf{d}$ and $\mathbf{a}$:

$$\mathbb{E}[\text{Payoff}(\mathbf{d}^*, \mathbf{a})] \leq \mathbb{E}[\text{Payoff}(\mathbf{d}^*, \mathbf{a}^*)] \leq \mathbb{E}[\text{Payoff}(\mathbf{d}, \mathbf{a}^*)] \tag{8}$$

In the context of our zero-sum game, the defender $\mathbf{d}$ seeks to minimize the expected payoff (prediction set size), and the attacker $\mathbf{a}$ seeks to maximize it. The Nash Equilibrium $(\mathbf{d}^*, \mathbf{a}^*)$ is the stable point where no player can gain an advantage by unilaterally deviating.

The Minimax Theorem states that the minmax value (the best the defender can do) is equal to the maxmin value (the best the attacker can do), and this common value, V, is the value of the game. The optimal mixed strategies $(\mathbf{d}^*, \mathbf{a}^*)$ achieve this value V. Formally, the value of the game V is defined by the equality:

$$\begin{aligned} V = \mathbb{E}[\text{Payoff}(\mathbf{d}^*, \mathbf{a}^*)] &= \min_{\mathbf{d}} \max_{\mathbf{a}} \mathbb{E}[\text{Payoff}(\mathbf{d}, \mathbf{a})] \\ &= \max_{\mathbf{a}} \min_{\mathbf{d}} \mathbb{E}[\text{Payoff}(\mathbf{d}, \mathbf{a})] \end{aligned} \tag{9}$$

To compute this equilibrium, we first construct a payoff matrix where rows correspond to defender's pure strategies $f_j \in \mathcal{F}$ and columns correspond to attacker's pure strategies $g_n \in \mathcal{G}$, with each entry being Payoff(f_j, g_n). We then solve for the optimal mixed strategies $(\mathbf{d}^*, \mathbf{a}^*)$ by solving a Linear Program derived from the Minimax equality, which yields a solution that is robust and realistic, especially when pure strategy Nash Equilibria do not exist or when players wish to introduce unpredictability. Algorithm 1 outlines the entire process of our framework.

Remark 1. Our framework finds a Nash Equilibrium, defined by optimal mixed strategies for the defender ($\mathbf{d}^*$) and attacker ($\mathbf{a}^*$). This creates a stable state where no player can gain an advantage by unilaterally deviating, as any such move would be met by the opponent's optimal counter-play. The defender's strategy $\mathbf{d}^*$, an unpredictable probabilistic ensemble, is thus robust against the attacker's best possible strategic response, with the comprehensiveness of the attack set $\mathcal{G}$ defining the strict boundary and scope of this guarantee.

4 Related Work

The advent of adversarial examples Goodfellow et al. (2014); Zhang et al. (2022) has posed substantial security challenges within the field of machine learning. In this context, uncertainty quantification emerges as a critical factor for enhancing the resilience of deep learning models. Conformal Prediction (CP) Papadopoulos et al. (2002); Vovk et al. (2005), renowned for its ability to provide distribution-independent coverage guarantees, faces notable obstacles when confronted with data poisoning and adversarial perturbations. Empirical investigations, such as those presented by Liu et al. (2024), reveal that conventional adversarial attack strategies—including the Projected Gradient Descent (PGD) method Madry et al. (2017)—can significantly undermine the reliability of conformal prediction frameworks.

Adversarially Robust Conformal Prediction To alleviate the adverse effects on conformal prediction (CP) under adversarial settings, a number of research efforts have been dedicated to tackling this challenge without relying on model retraining.

Adversarially Robust Conformal Prediction (ARCP) Gendler et al. (2021) integrates conformal prediction with randomized smoothing to provide finite-sample coverage guarantees in the presence of L_2-norm-bounded adversarial noise. By incorporating Gaussian noise, it constrains the Lipschitz constant of the non-conformity score, thereby handling unknown adversarial perturbations without the need for training adjustments. Probabilistically Robust Conformal Prediction (PRCP) Ghosh et al. (2023) employs a quantile-of-quantile approach to adapt to perturbations, establishing thresholds for both data samples and perturbations. It utilizes adversarial attacks to compute empirical robust quantiles, independently of the model training process.

Yan et al. (2024) put forward two methods, Post-Training Transformation (PTT) and Robust Conformal Training (RCT), to enhance the efficiency of robust conformal prediction. They revise RSCP into RSCP+ to provide certified guarantees and incorporate it into the training phase. Zargarbashi et al. (2024) develop robust prediction sets by bounding the worst-case variations in conformity scores under adversarial evasion and poisoning attacks. They employ CDF-based bounds to calculate conservative prediction sets and corresponding thresholds for these scenarios. Jeary et al. (2024) introduce Verifiable Robust Conformal Prediction (VRCP), which utilizes neural network verification techniques to uphold coverage guarantees amid adversarial attacks. VRCP accommodates arbitrary norm-bounded perturbations and can be extended to regression tasks.

While these approaches offer partial alleviation of the adversarial impact, they either sacrifice the compactness of the prediction sets or struggle to sustain robustness against diverse attack types and varying perturbation magnitudes.

Adversarial Training for Conformal Prediction To bolster adversarial robustness in conformal prediction (CP), a straightforward and intuitive strategy involves integrating adversarial training techniques into the CP framework.

Bao et al. (2025) put forward a method that enhances adversarial robustness under the CP setting by training on samples with uncertain attack models and incorporating conformal training with a hard threshold. More-

over, Liu et al. (2024) introduced Uncertainty-Reducing Adversarial Training (AT-UR) aimed at elevating both the efficiency and adversarial robustness of CP. This is achieved by minimizing predictive entropy and employing a weighted loss function grounded in True Class Probability Ranking. They further combined AT-UR with several established adversarial training methods, namely Adversarial Training (AT) Madry et al. (2017), Friendly Adversarial Training (FAT) Zhang et al. (2020), and Trade-off-inspired Adversarial Defense via Surrogate-loss Minimization (TRADES) Zhang et al. (2019). For generating adversarial examples, they used the Projected Gradient Descent (PGD) approach.

5 Experiment Results

Attacks	Indicator	APS	RANK	RAPS	SAPS	TOPK
Clean	Coverage	0.997 (0.001)	0.998 (0.001)	0.997 (0.001)	0.997 (0.001)	0.997 (0.001)
	Size	4.826 (0.092)	5.263 (0.070)	4.833 (0.095)	4.666 (0.105)	5.655 (0.093)
	SSCV	0.100 (0.000)	0.098 (0.001)	0.100 (0.000)	0.100 (0.000)	0.097 (0.001)
FGSM	Coverage	0.982 (0.003)	0.982 (0.002)	0.982 (0.003)	0.981 (0.003)	0.980 (0.003)
	Size	5.237 (0.086)	5.356 (0.076)	5.241 (0.090)	5.069 (0.096)	5.653 (0.091)
	SSCV	0.099 (0.002)	0.082 (0.002)	0.099 (0.000)	0.098 (0.001)	0.080 (0.003)
PGD	Coverage	0.983 (0.003)	0.984 (0.002)	0.983 (0.002)	0.982 (0.002)	0.981 (0.003)
	Size	5.189 (0.086)	5.342 (0.075)	5.193 (0.088)	5.040 (0.099)	5.655 (0.089)
	SSCV	0.099 (0.002)	0.084 (0.002)	0.099 (0.001)	0.098 (0.002)	0.081 (0.003)
APGD	Coverage	0.995 (0.001)	0.989 (0.002)	0.995 (0.001)	0.991 (0.002)	0.986 (0.002)
	Size	4.879 (0.092)	5.287 (0.066)	4.887 (0.093)	4.762 (0.103)	5.654 (0.091)
	SSCV	0.100 (0.000)	0.089 (0.002)	0.100 (0.001)	0.100 (0.000)	0.086 (0.002)
CW	Coverage	0.997 (0.001)	0.998 (0.001)	0.997 (0.001)	0.997 (0.001)	0.997 (0.001)
	Size	4.917 (0.090)	5.291 (0.069)	4.924 (0.091)	5.124 (0.095)	**5.658 (0.089)**
	SSCV	0.100 (0.000)	0.098 (0.001)	0.100 (0.001)	0.100 (0.000)	0.097 (0.001)
PIFGSM	Coverage	0.898 (0.008)	0.898 (0.007)	0.899 (0.008)	0.898 (0.007)	0.900 (0.007)
	Size	5.553 (0.085)	5.416 (0.083)	5.563 (0.088)	5.282 (0.094)	5.656 (0.088)
	SSCV	0.099 (0.002)	0.005 (0.005)	0.099 (0.002)	0.066 (0.013)	0.006 (0.004)
GN	Coverage	0.995 (0.001)	0.989 (0.002)	0.995 (0.001)	0.992 (0.001)	0.986 (0.002)
	Size	**5.562 (0.091)**	**5.532 (0.072)**	**5.570 (0.086)**	**5.305 (0.078)**	5.653 (0.092)
	SSCV	0.100 (0.000)	0.089 (0.002)	0.100 (0.000)	0.100 (0.000)	0.086 (0.002)
Square	Coverage	0.997 (0.001)	0.998 (0.001)	0.997 (0.001)	0.998 (0.001)	0.997 (0.001)
	Size	4.833 (0.088)	5.269 (0.068)	4.840 (0.094)	4.756 (0.099)	5.655 (0.088)
	SSCV	0.100 (0.000)	0.098 (0.001)	0.100 (0.000)	0.100 (0.000)	0.097 (0.001)
SPSA	Coverage	0.996 (0.001)	0.998 (0.001)	0.997 (0.001)	0.997 (0.001)	0.997 (0.001)
	Size	4.852 (0.091)	5.266 (0.070)	4.858 (0.094)	4.699 (0.107)	5.658 (0.088)
	SSCV	0.100 (0.000)	0.098 (0.001)	0.100 (0.001)	0.100 (0.000)	0.097 (0.001)
Nash	Coverage	0.949 (0.013)	0.989 (0.001)	0.951 (0.012)	0.968 (0.021)	0.982 (0.029)
	Size	**5.558 (0.088)**	**5.532 (0.072)**	**5.566 (0.087)**	**5.304 (0.078)**	**5.657 (0.092)**
	SSCV	0.099 (0.001)	0.089 (0.001)	0.100 (0.001)	0.093 (0.006)	0.083 (0.028)

Table 1: Performance of various Conformal Prediction (CP) methods under the Nash defense strategy on the CIFAR-10 dataset. The table presents the mean and standard deviation of coverage, size, and SSCV against a range of attack strategies. All results are averaged over 20 independent random splits.

5.1 Experimental Setup

To comprehensively evaluate the effectiveness and robustness of our proposed game-theoretic framework across datasets of varying complexity and scale, we have designed a series of detailed experiments. This section will provide a thorough introduction to the datasets used in the experiments, the defense models, the attack methods, as well as the specific parameter configurations.

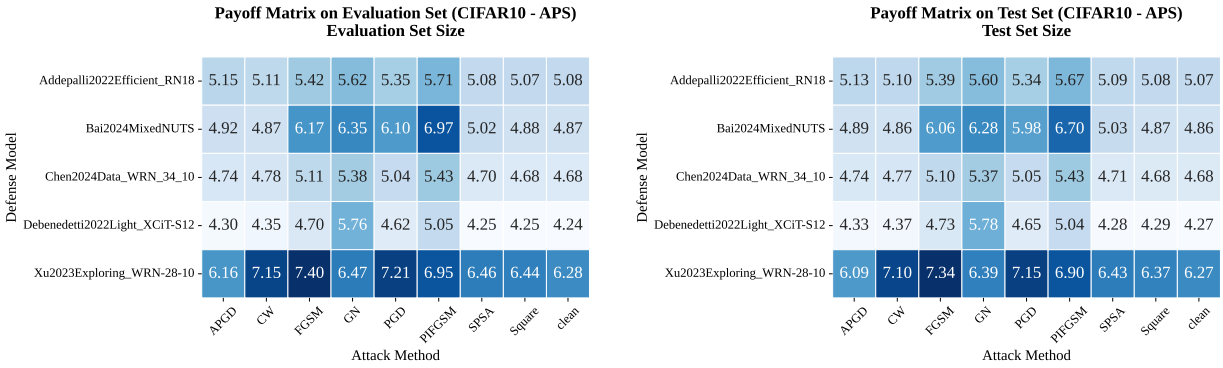

Figure 2: Evaluate the payoff matrix on both the CIFAR-10 dataset's evaluation set and test set using the APS method.

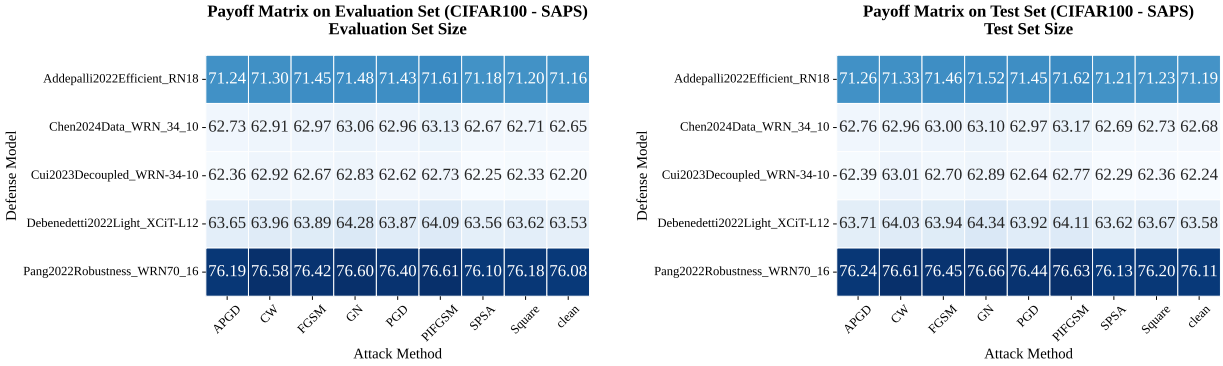

Figure 3: Evaluate the payoff matrix on both the CIFAR-100 dataset's evaluation set and test set using the SAPS method.

5.1.1 Datasets

Our experiments are conducted on three standard computer vision benchmarks: CIFAR-10, CIFAR-100, and ImageNet. For CIFAR-10 and CIFAR-100, we utilize their official test sets. For ImageNet, we use the official validation set. For each dataset, we randomly split the corresponding set into three disjoint subsets: 30% for the calibration set ($\mathcal{D}_{\text{cal}}$), 30% for the evaluation set ($\mathcal{D}_{\text{eval}}$), and the remaining 40% for the final test set ($\mathcal{D}_{\text{test}}$).

5.1.2 Adversarial Attacks

We generate adversarial examples using the torchattacks library Kim (2020), with the perturbation for all attacks constrained within an L_∞-norm ball of radius $\epsilon = 8/255$. As a baseline, we use original, unperturbed images. The specific attack parameters are set as follows: the FGSM (Fast Gradient Sign Method) attack Goodfellow et al. (2014) uses the default epsilon of $\epsilon = 8/255$; PGD (Projected Gradient Descent) Madry et al. (2017) is configured with 10 iteration steps, a step size (α) of $1/255$, and a random starting point within

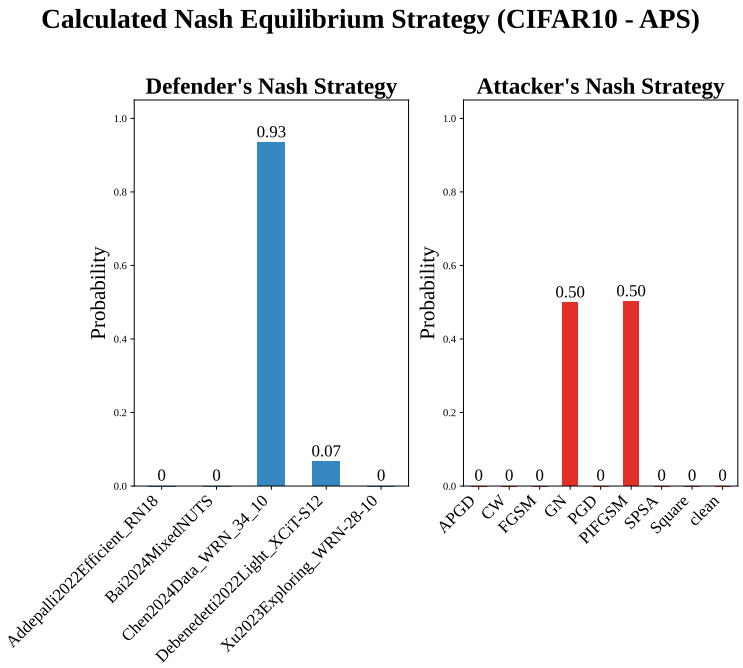

Figure 4: Obtain the Nash attack and Nash defense strategies on the CIFAR-10 dataset using the APS method.

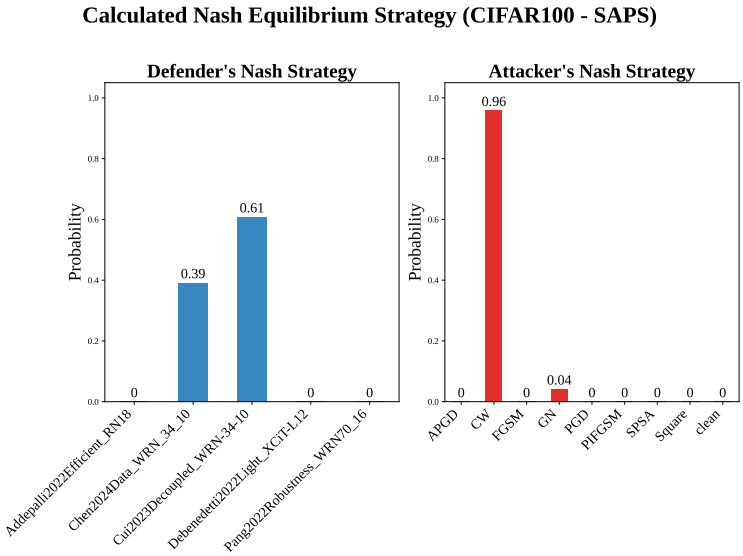

Figure 5: Obtain the Nash attack and Nash defense strategies on the CIFAR-100 dataset using the SAPS method.

the ϵ-ball; APGD (Auto-PGD) Croce & Hein (2020) is set to run for 10 steps utilizing the cross-entropy loss function. Furthermore, the CW (Carlini & Wagner) attack Carlini & Wagner (2017) is configured with 50 optimization steps, a learning rate of 0.01, and a confidence parameter (c) of 1; PIFGSM (Projected Iterative FGSM) Gao et al. (2020) is executed for 10 iterations; and the default library implementation is used for Gaussian Noise (GN). For the CIFAR-10 and CIFAR-100 datasets, we also incorporate two additional black-box methods: the Square attack Andriushchenko et al. (2020), which is limited to a maximum of 1000 queries

per sample, and SPSA (Simultaneous Perturbation Stochastic Approximation) Uesato et al. (2018), for which we employ the default library implementation.

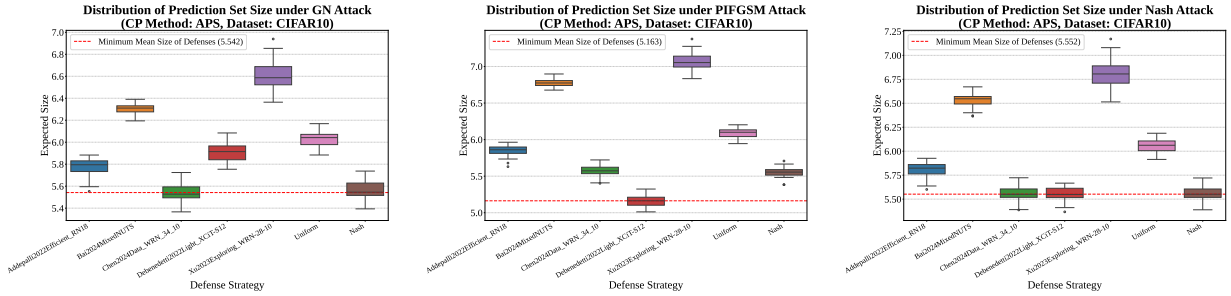

Figure 6: Box plots of size for different defense models using the APS method on the CIFAR-10 dataset under various attacks.

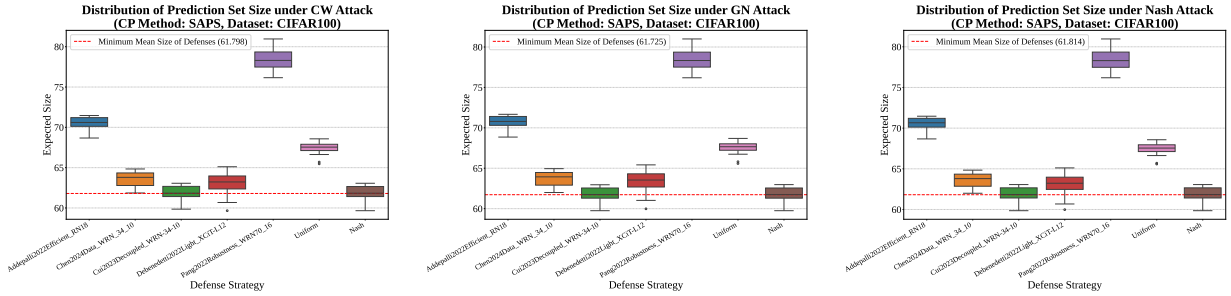

Figure 7: Box plots of size for different defense models using the SAPS method on the CIFAR-100 dataset under various attacks.

5.1.3 Defensive Models

To ensure fairness and reproducibility, all defensive models, denoted as f_j, are sourced from the RobustBench benchmark Croce et al. (2021). RobustBench provides a standardized library of publicly available, state-of-the-art models. Since there are inconsistencies in training models across different datasets, we mainly selected these models Addepalli et al. (2022); Bai et al. (2024); Chen & Lee (2024); Debenedetti et al. (2023); Xu et al. (2023); Cui et al. (2024); Addepalli et al. (2022); Pang et al. (2022); Peng et al. (2023). These models have verified adversarial robustness and serve as a recognized reference for progress in the field.

In our experiments, the significance level α is consistently set to 0.1 across all trials. We employ five widely recognized CP methods Huang et al. (2024a) to evaluate the performance of our framework, namely APS Romano et al. (2020), RAPS Angelopoulos et al. (2021), TOPK Angelopoulos et al. (2021), SAPS Huang et al. (2024b), and RANK Luo & Zhou (2024) approaches. We use coverage, size, and SSCV Angelopoulos et al. (2021) as evaluation metrics.

A core finding of our game-theoretic framework is the exceptional robustness of the Nash defense strategy, irrespective of the Conformal Prediction (CP) method. As shown in Table 1, our Nash defense guarantees that the expected prediction set size remains at or below the threshold set by the attacker's own Nash strategy. This upper bound holds consistently, with minor deviations attributable to the inherent generalization gap between the evaluation and test sets.

Additionally, as shown in Figures 6 and 7, when confronted with Nash attacks, our Nash defense strategy consistently outperforms many standalone defense models. While some defense models demonstrate comparable performance, their vulnerabilities are also exposed (as illustrated in the left and middle panels of Figures 6 and 7). The key advantage of Nash defense lies in its guaranteed performance ceiling: it ensures

that the size of the prediction set does not exceed the upper bound defined by the Nash attack value, a guarantee that other individual models cannot provide against their respective worst-case attacks. This is the essence of its robustness: it prevents adversaries from gaining an advantage by unilaterally altering their strategies, thereby ensuring reliable worst-case performance.

As depicted in figure 2-3, it illustrates the errors generated by Nash strategies on the evaluation set, while figure 4-5 showcases the attack and defense strategies under Nash equilibrium. Additionally, other results are presented in Appendix 3-5.

6 Conclusion

In this work, we proposed a novel game-theoretic framework to address the critical failure of conformal prediction under adversarial attacks, which violate the core exchangeability assumption. By modeling the interaction between defender and attacker as a zero-sum game, we shift the defense paradigm from selecting a static model to achieving a dynamic, strategic equilibrium. Our theoretical analysis and empirical results on CIFAR-10, CIFAR-100, and ImageNet demonstrate that the Nash Equilibrium solution yields a mixed defense strategy that is provably robust. This strategy not only maintains valid coverage guarantees with standard non-conformity scores but also minimizes prediction set size against an optimal, mixed-strategy attacker. The framework provides a verifiable defense that guarantees a worst-case performance, ensuring that no adversary can gain an advantage by unilaterally deviating from their optimal strategy. By bridging adversarial robustness and uncertainty quantification, our work offers a principled and powerful path toward deploying reliable and trustworthy AI systems in safety-critical domains. Future work could explore extending this framework to a continuous strategy space, or integrating certified robustness techniques, to provide guarantees against adversarial perturbations $g \notin G$.

References

Sravanti Addepalli, Samyak Jain, et al. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, 35:1488–1501, 2022.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

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

Yatong Bai, Mo Zhou, Vishal M Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *Transactions on Machine Learning Research*, 2024.

Jie Bao, Chuangyin Dang, Rui Luo, Hanwei Zhang, and Zhixin Zhou. Enhancing adversarial robustness with conformal prediction: A framework for guaranteed model reliability. In *Proceedings of the Forty-second International Conference on Machine Learning*. PMLR, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Erh-Chung Chen and Che-Rung Lee. Data filtering for efficient adversarial training. *Pattern Recognition*, 151:110394, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37:74461–74486, 2024.

Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 225–253. IEEE, 2023.

Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision*, pp. 307–322. Springer, 2020.

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*, pp. 681–690. PMLR, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jianguo Huang, Jianqing Song, Xuanning Zhou, Bingyi Jing, and Hongxin Wei. Torchcp: A python library for conformal prediction, 2024a.

Jianguo Huang, HuaJun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Forty-first International Conference on Machine Learning*, 2024b.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. *arXiv preprint arXiv:2405.18942*, 2024.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

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. *arXiv preprint arXiv:2405.08886*, 2024.

Rui Luo and Zhixin Zhou. Trustworthy classification through rank-based conformal prediction sets. *arXiv preprint arXiv:2407.04407*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International conference on machine learning*, pp. 17258–17277. PMLR, 2022.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer, 2002.

ShengYun Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, Mansi Phute, Jason Martin, and Duen Horng Chau. Robust principles: Architectural design principles for adversarially robust cnns. *arXiv preprint arXiv:2308.16258*, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International conference on machine learning*, pp. 5025–5034. PMLR, 2018.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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

Yuancheng Xu, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang. Exploring and exploiting decision boundary dynamics for adversarial robustness. *arXiv preprint arXiv:2302.03015*, 2023.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. *arXiv preprint arXiv:2404.19651*, 2024.

Soroush H Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. *arXiv preprint arXiv:2407.09165*, 2024.

Hanwei Zhang, Teddy Furon, Laurent Amsaleg, and Yannis Avrithis. Deep neural network attacks and defense: The case of image classification. *Multimedia Security*, 1:41–75, 2022.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pp. 11278–11287. PMLR, 2020.

A Appendix

A.1 Proof for Theorem 1

Proof. Let f_j be a given classifier and let $\mathrm{x}_i^{\mathrm{adv}}$ be a test sample generated by an arbitrary attack $g_n \in \{g_n\}_{n=1}^m$.

By the definition of the conformal quantile, for the specific attack g_n, the quantile $q_{1-\alpha}^{j_n}$ computed on the calibration set $\mathcal{D}_{\mathrm{cal}}$ guarantees:

$$\mathbb{P}\left(s(g_n(\mathrm{x}_i, y_i, f_j), y_i) \leq q_{1-\alpha}^{j_n}\right) \geq 1-\alpha. \tag{10}$$

The robust quantile $q_{1-\alpha}^{j}$ for model f_j is defined as:

$$q_{1-\alpha}^{j} = \max_{k \in 1,\ldots,m} q_{1-\alpha}^{j_k}. \tag{11}$$

From this definition, it directly follows that for our chosen attack g_n:

$$q_{1-\alpha}^{j} \geq q_{1-\alpha}^{j_n}. \tag{12}$$

The event $s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j_n}$ is a subset of the event $s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j}$. Therefore, the probability of the latter is at least as large as the probability of the former:

$$\mathbb{P}\left(s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j}\right) \geq \mathbb{P}\left(s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j_n}\right). \tag{13}$$

The condition for the true label y_i to be in the prediction set $\mathcal{C}(\mathrm{x}_i^{\mathrm{adv}}; f_j, q_{1-\alpha}^{j})$ is precisely $s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j}$. Combining these steps, we have:

$$\begin{aligned}
\mathbb{P}\left(y_i \in \mathcal{C}(\mathrm{x}_i^{\mathrm{adv}}; f_j, q_{1-\alpha}^{j})\right) &= \mathbb{P}\left(s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j}\right) \\
&\geq \mathbb{P}\left(s(\mathrm{x}_i^{\mathrm{adv}}, y_i) \leq q_{1-\alpha}^{j_n}\right) \\
&\geq 1-\alpha.
\end{aligned}$$

This completes the proof. □

A.2 Conformal Prediction Evaluation

The Coverage measures the proportion of test instances in $\mathcal{D}_{\text{test}}$ where the true label is contained within the prediction set $\Gamma(x_i)$, and is defined as

$$\text{Coverage} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \mathbb{1}\left(y_i \in \Gamma(x_i)\right). \tag{14}$$

A higher coverage indicates that the prediction sets reliably contain the true labels.

The Size metric calculates the average number of labels in the prediction sets across all test instances,

$$\text{Size} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} |\Gamma(x_i)|, \tag{15}$$

where smaller sizes denote more precise and informative predictions.

The Size-Stratified Coverage Violation (SSCV) evaluates the consistency of coverage across different prediction set sizes. It is defined as

$$\text{SSCV}(\Gamma, \{S_j\}_{j=1}^{s}) = \sup_{j \in [s]} \left| \frac{|\{i \in J_j : y_i \in \Gamma(x_i)\}|}{|J_j|} - (1-\alpha) \right|, \tag{16}$$

where $\{S_j\}_{j=1}^{s}$ partitions the possible prediction set sizes, and $J_j = \{i \in \mathcal{D}_{\text{test}} : |\Gamma(x_i)| \in S_j\}$. A smaller SSCV indicates more stable coverage across different set sizes.

A.3 CIFAR10 results

In this section, we provide supplementary payoff matrices (on the evaluation set) and payoff matrices (on the test set) for other methods applied to the CIFAR-10 dataset, along with corresponding results such as box plots.

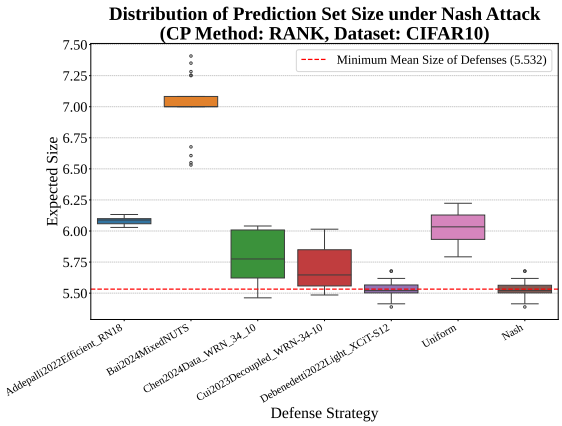

Figure 8: Box plots of size for different defense models using the RANK method on the CIFAR-10 dataset.

A.4 CIFAR100 results

In this section, we provide supplementary payoff matrices (on the evaluation set) and payoff matrices (on the test set) for other methods applied to the CIFAR-100 dataset, along with corresponding results such as box plots.

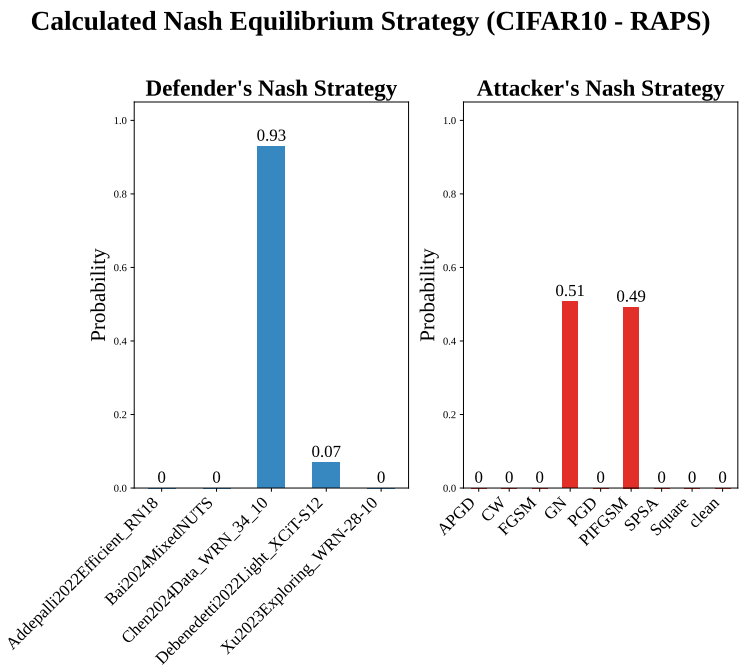

Figure 9: Obtain the Nash attack and Nash defense strategies on the CIFAR-10 dataset using the RAPS method.

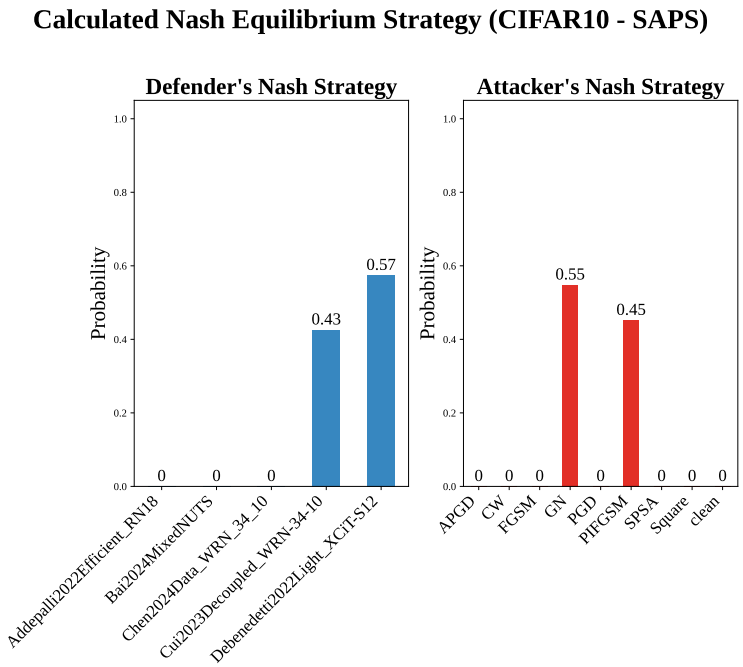

Figure 10: Obtain the Nash attack and Nash defense strategies on the CIFAR-10 dataset using the RANK method.

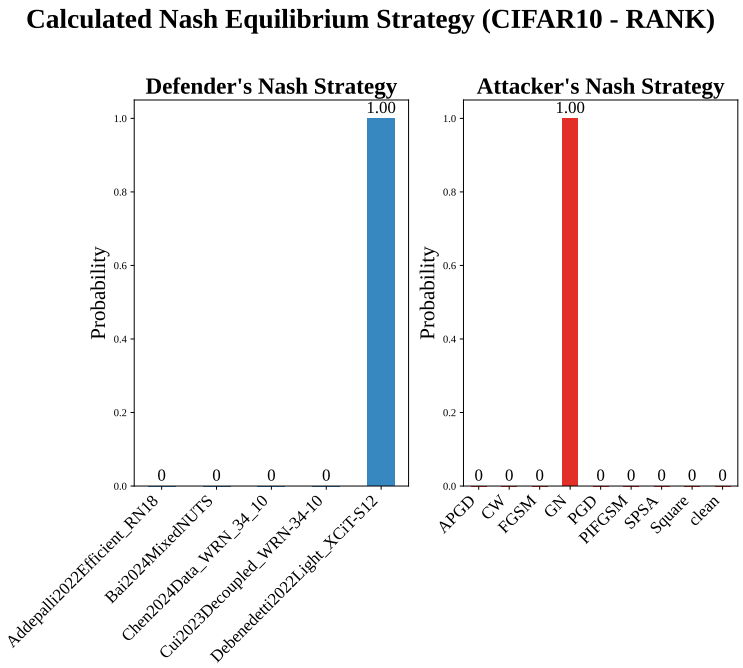

Figure 11: Obtain the Nash attack and Nash defense strategies on the CIFAR-10 dataset using the RANK method.

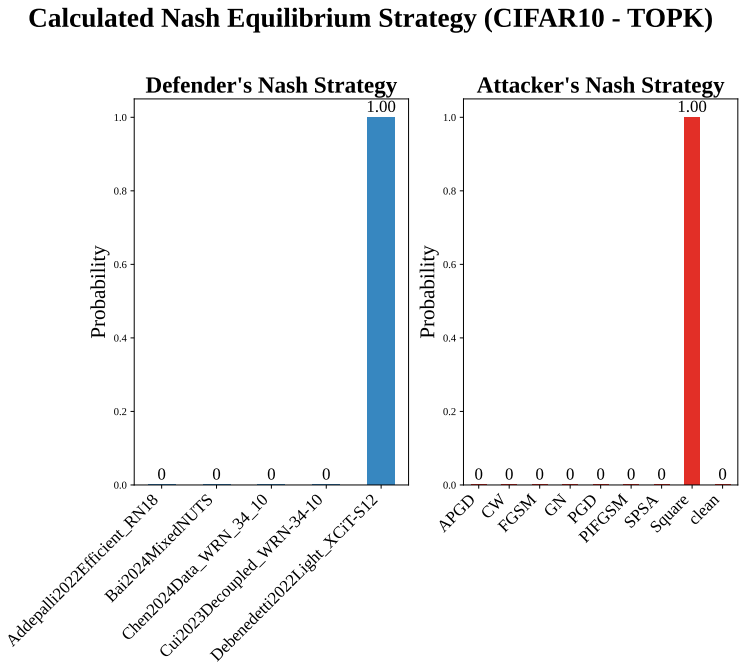

Figure 12: Obtain the Nash attack and Nash defense strategies on the CIFAR-10 dataset using the TOPK method.

A.5 ImageNet results

In this section, we provide supplementary payoff matrices (on the evaluation set) and payoff matrices (on the test set) for other methods applied to the ImageNet dataset, along with corresponding results such as box plots.

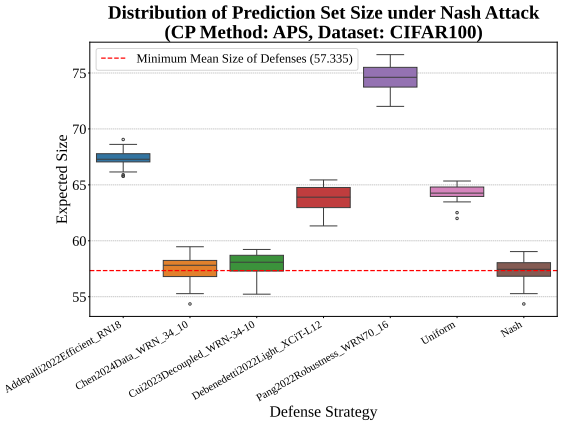

Figure 13: Box plots of size for different defense models using the APS method on the CIFAR-100 dataset.

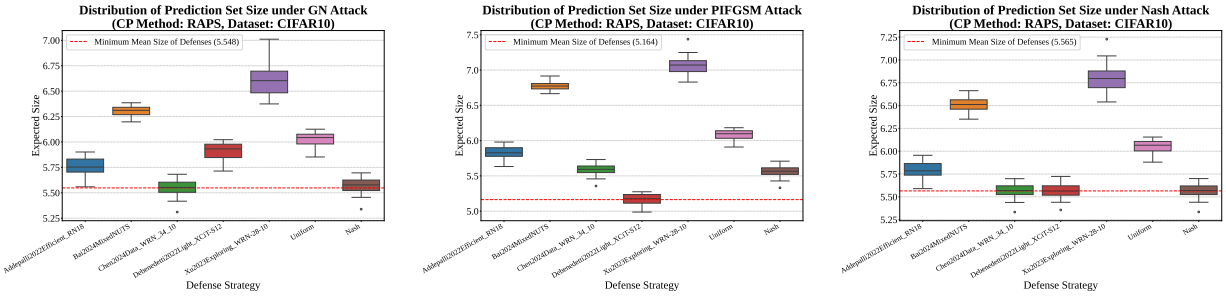

Figure 14: Box plots of size for different defense models using the RAPS method on the CIFAR-10 dataset under various attacks.

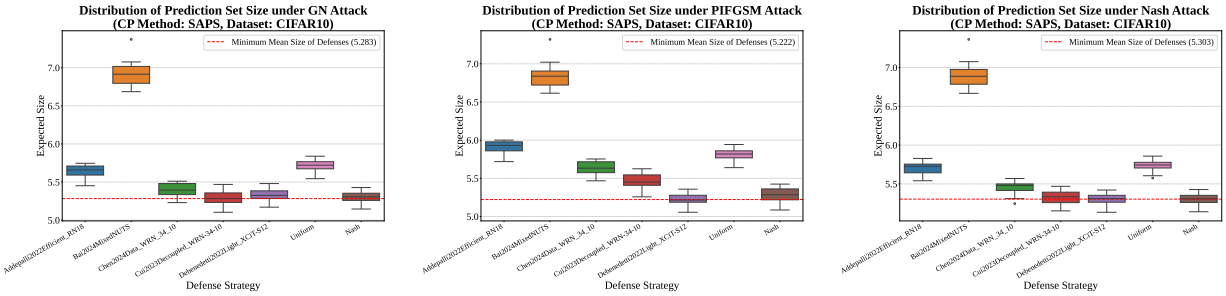

Figure 15: Box plots of size for different defense models using the SAPS method on the CIFAR-10 dataset under various attacks.

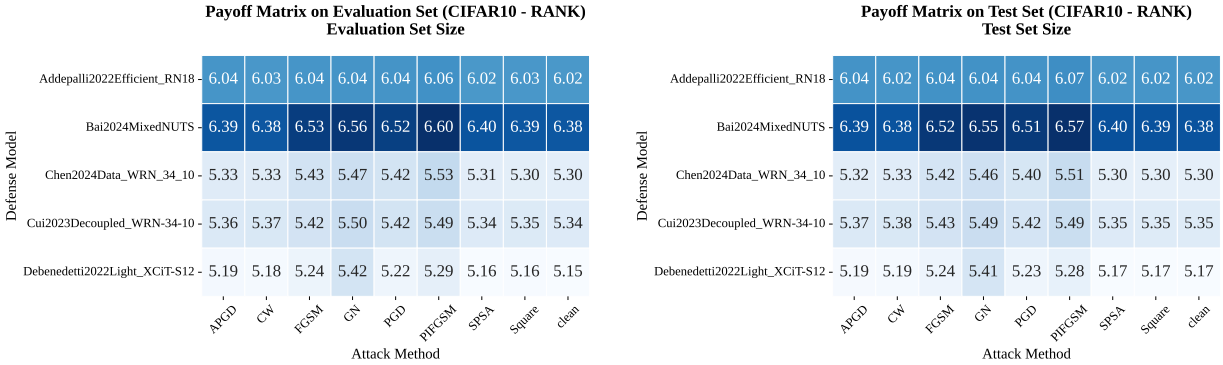

Figure 16: Evaluate the payoff matrix on both the CIFAR-10 dataset's evaluation set and test set using the RANK method.

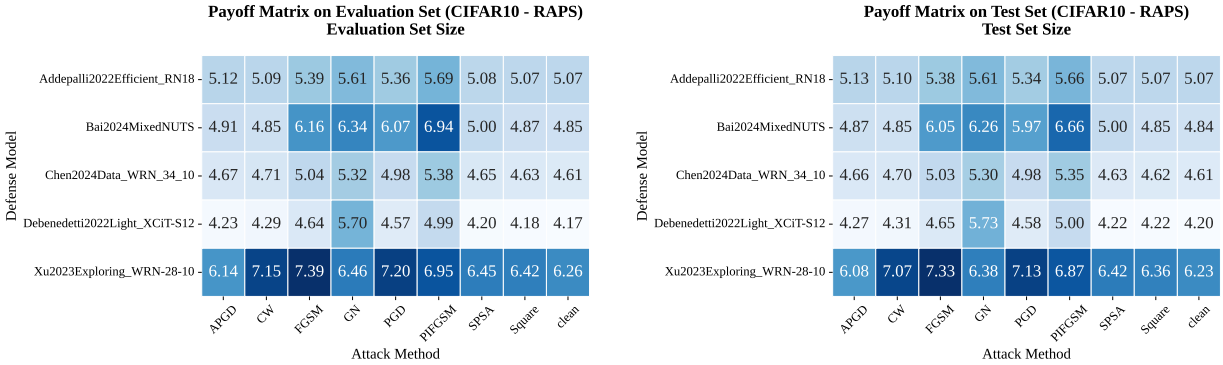

Figure 17: Evaluate the payoff matrix on both the CIFAR-10 dataset's evaluation set and test set using the RAPS method.

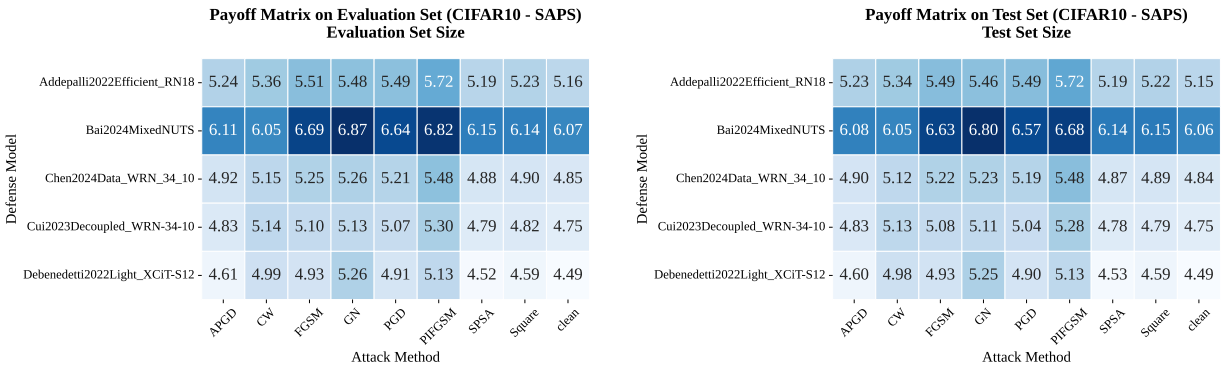

Figure 18: Evaluate the payoff matrix on both the CIFAR-10 dataset's evaluation set and test set using the SAPS method.

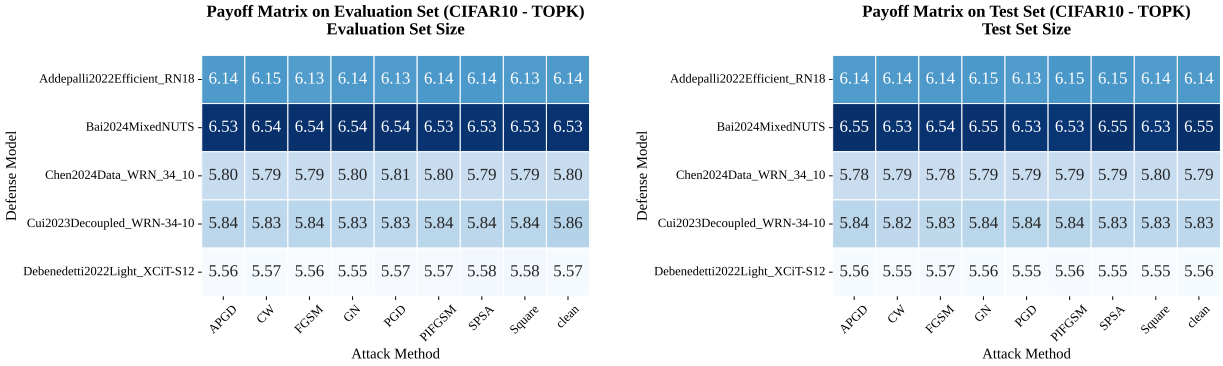

Figure 19: Evaluate the payoff matrix on both the CIFAR-10 dataset's evaluation set and test set using the TOPK method.

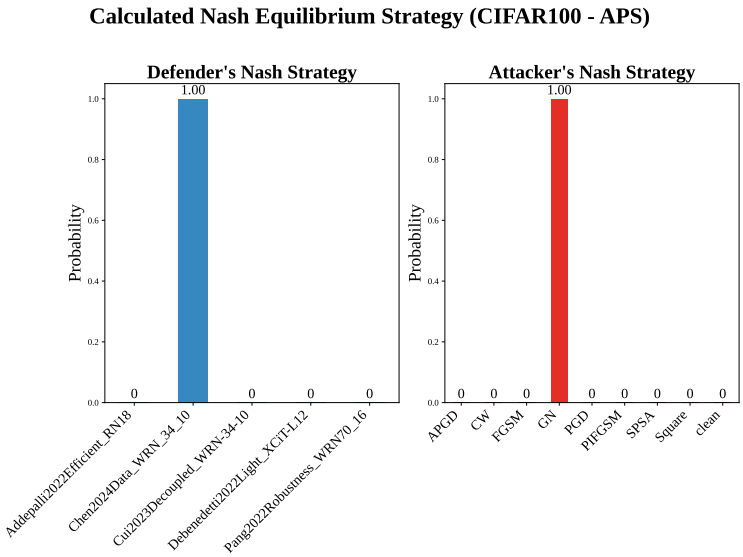

Figure 20: Obtain the Nash attack and Nash defense strategies on the CIFAR-100 dataset using the APS method.

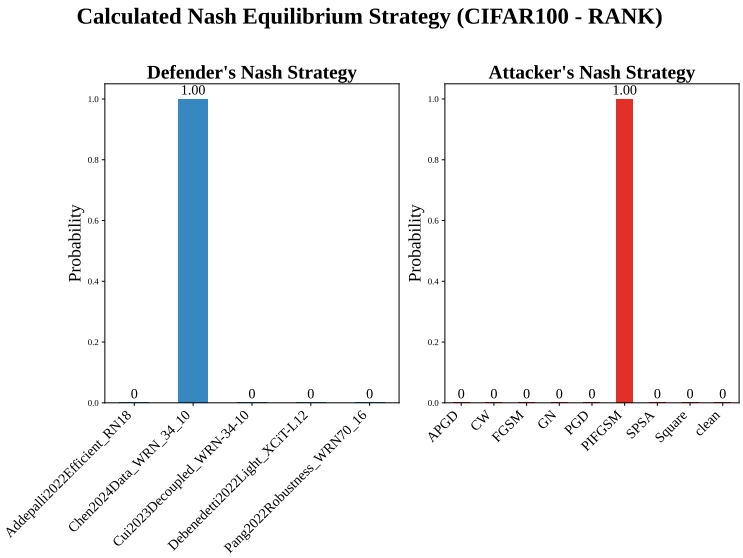

Figure 21: Obtain the Nash attack and Nash defense strategies on the CIFAR-100 dataset using the RANK method.

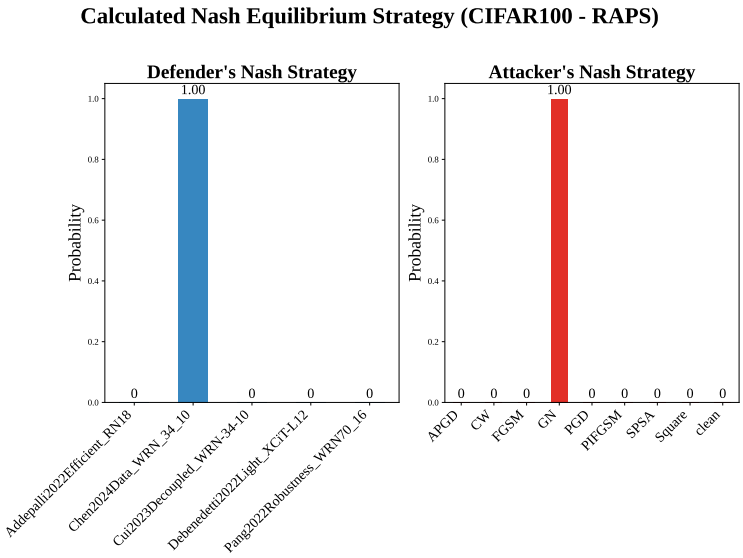

Figure 22: Obtain the Nash attack and Nash defense strategies on the CIFAR-100 dataset using the RAPS method.

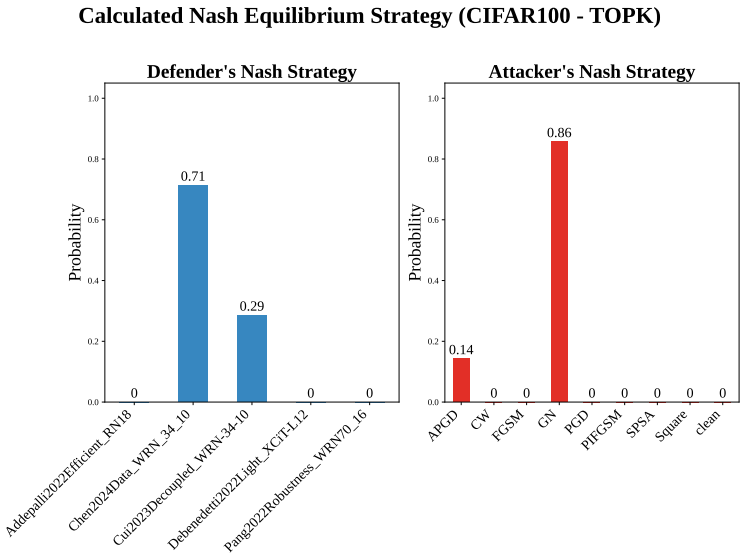

Figure 23: Obtain the Nash attack and Nash defense strategies on the CIFAR-100 dataset using the TOPK method.

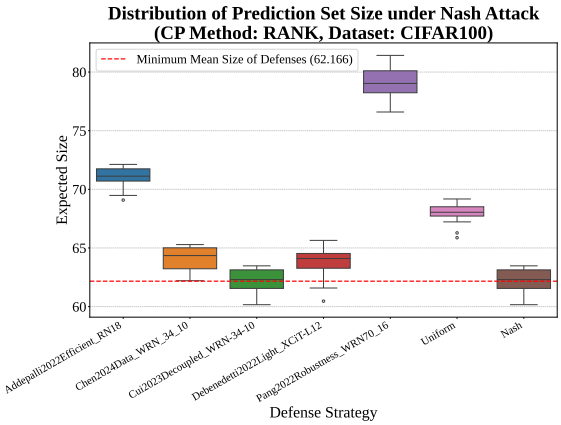

Figure 24: Box plots of size for different defense models using the RANK method on the CIFAR-100 dataset.

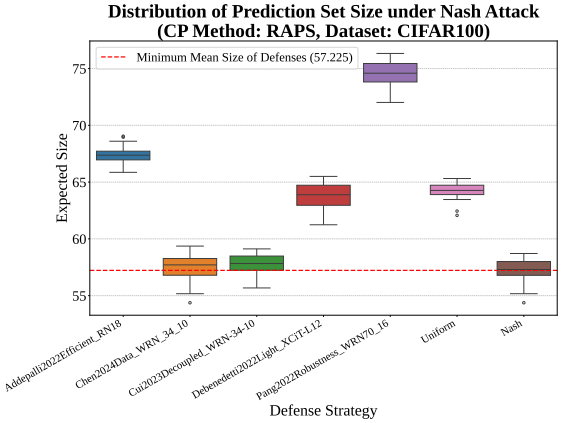

Figure 25: Box plots of size for different defense models using the RAPS method on the CIFAR-100 dataset.

Attacks	Indicator	APS	RANK	RAPS	SAPS	TOPK
Clean	Coverage	0.996 (0.003)	0.997 (0.001)	0.996 (0.003)	0.997 (0.001)	0.997 (0.001)
	Size	49.198 (2.038)	62.059 (1.023)	48.881 (1.960)	61.097 (0.985)	62.304 (1.010)
	SSCV	0.100 (0.001)	0.097 (0.001)	0.100 (0.001)	0.097 (0.001)	0.097 (0.001)
FGSM	Coverage	0.980 (0.003)	0.977 (0.002)	0.979 (0.002)	0.977 (0.002)	0.978 (0.003)
	Size	53.423 (1.117)	62.123 (1.021)	53.238 (1.045)	61.535 (0.994)	62.305 (1.005)
	SSCV	0.094 (0.008)	0.077 (0.002)	0.093 (0.008)	0.077 (0.002)	0.078 (0.003)
PGD	Coverage	0.981 (0.003)	0.978 (0.002)	0.981 (0.003)	0.978 (0.002)	0.979 (0.003)
	Size	52.955 (1.181)	62.113 (1.022)	52.757 (1.106)	61.496 (0.991)	62.302 (1.008)
	SSCV	0.095 (0.007)	0.078 (0.002)	0.095 (0.007)	0.078 (0.002)	0.079 (0.003)
APGD	Coverage	0.996 (0.003)	0.997 (0.001)	0.996 (0.003)	0.997 (0.001)	0.997 (0.001)
	Size	49.999 (1.845)	62.092 (1.022)	49.706 (1.763)	61.233 (0.985)	62.302 (1.010)
	SSCV	0.100 (0.001)	0.097 (0.001)	0.100 (0.001)	0.097 (0.001)	0.097 (0.001)
CW	Coverage	0.996 (0.003)	0.997 (0.001)	0.996 (0.002)	0.997 (0.001)	0.997 (0.001)
	Size	52.307 (1.283)	62.099 (1.020)	52.218 (1.265)	61.798 (1.001)	62.303 (1.007)
	SSCV	0.100 (0.001)	0.097 (0.001)	0.100 (0.001)	0.097 (0.001)	0.097 (0.001)
PIFGSM	Coverage	0.898 (0.009)	0.898 (0.006)	0.897 (0.008)	0.898 (0.008)	0.896 (0.008)
	Size	55.500 (1.156)	62.133 (1.026)	55.291 (1.099)	61.610 (0.990)	62.305 (1.008)
	SSCV	0.143 (0.091)	0.005 (0.004)	0.127 (0.068)	0.006 (0.005)	0.007 (0.006)
GN	Coverage	0.991 (0.002)	0.988 (0.001)	0.991 (0.002)	0.988 (0.001)	0.988 (0.001)
	Size	57.335 (1.154)	62.166 (1.022)	57.225 (1.113)	61.725 (0.970)	62.302 (1.008)
	SSCV	0.100 (0.000)	0.088 (0.001)	0.100 (0.000)	0.088 (0.001)	0.088 (0.001)
Square	Coverage	0.996 (0.003)	0.997 (0.001)	0.996 (0.003)	0.997 (0.001)	0.997 (0.001)
	Size	49.391 (1.940)	62.068 (1.022)	49.086 (1.862)	61.198 (0.988)	62.305 (1.010)
	SSCV	0.100 (0.001)	0.097 (0.001)	0.100 (0.001)	0.097 (0.001)	0.097 (0.001)
SPSA	Coverage	0.995 (0.003)	0.996 (0.001)	0.995 (0.003)	0.996 (0.001)	0.996 (0.001)
	Size	49.487 (1.909)	62.060 (1.022)	49.184 (1.829)	61.131 (0.986)	62.308 (1.006)
	SSCV	0.100 (0.001)	0.096 (0.001)	0.099 (0.001)	0.096 (0.001)	0.096 (0.001)
Nash	Coverage	0.991 (0.002)	0.979 (0.029)	0.991 (0.002)	0.996 (0.004)	0.985 (0.024)
	Size	57.335 (1.154)	62.166 (1.022)	57.225 (1.113)	61.814 (0.967)	62.302 (1.006)
	SSCV	0.100 (0.000)	0.080 (0.025)	0.100 (0.000)	0.096 (0.004)	0.086 (0.019)

Table 2: Performance of various Conformal Prediction (CP) methods under the Nash defense strategy on the CIFAR-100 dataset. The table presents the mean and standard deviation of coverage, size, and SSCV against a range of attack strategies. All results are averaged over 20 independent random splits.

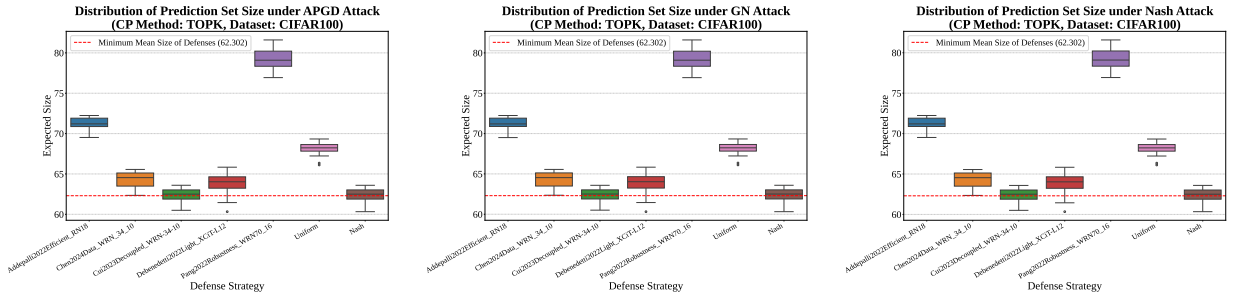

Figure 26: Box plots of size for different defense models using the TOPK method on the CIFAR-100 dataset under various attacks.

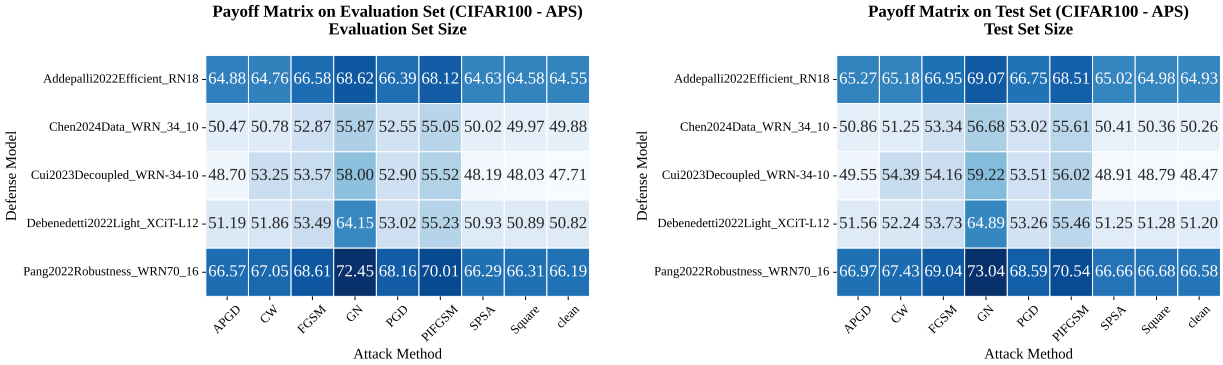

Figure 27: Evaluate the payoff matrix on both the CIFAR-100 dataset's evaluation set and test set using the APS method.

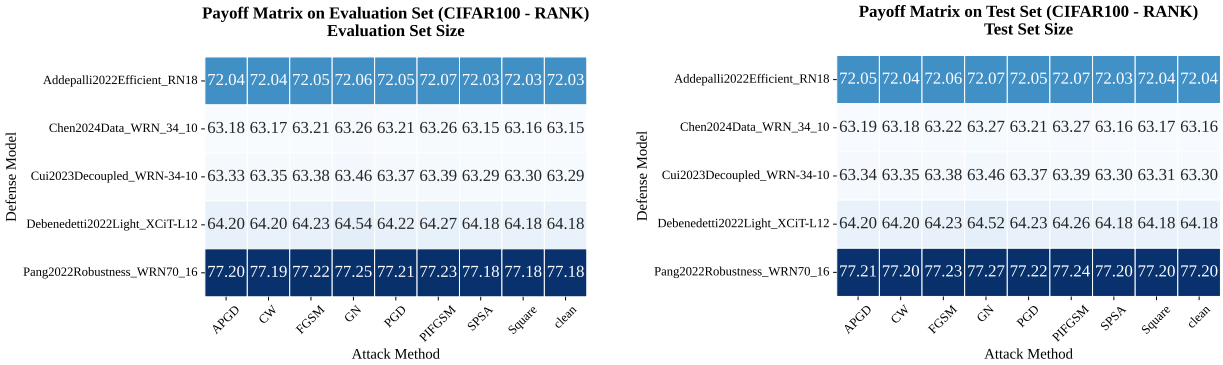

Figure 28: Evaluate the payoff matrix on both the CIFAR-100 dataset's evaluation set and test set using the RANK method.

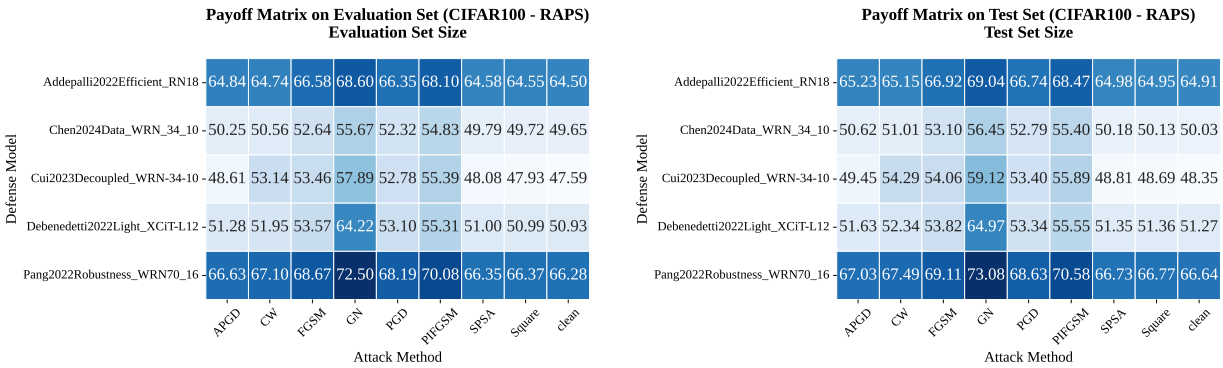

Figure 29: Evaluate the payoff matrix on both the CIFAR-100 dataset's evaluation set and test set using the RAPS method.

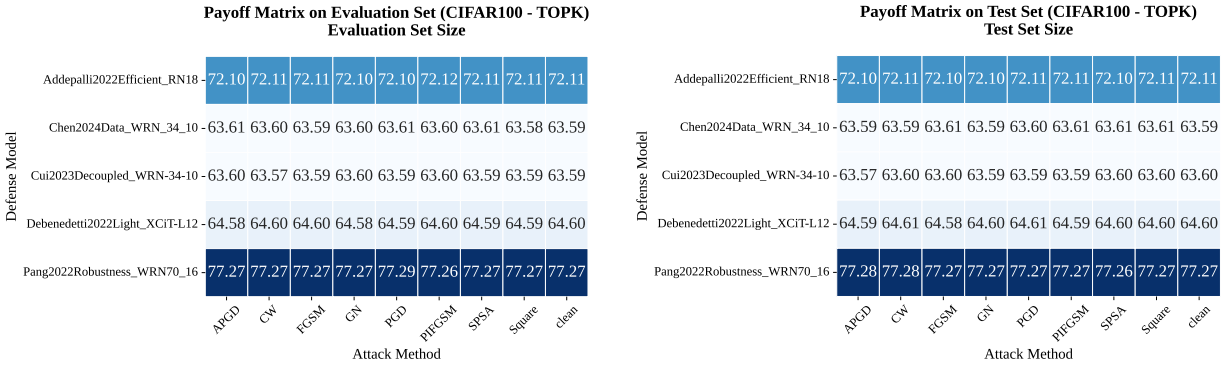

Figure 30: Evaluate the payoff matrix on both the CIFAR-100 dataset's evaluation set and test set using the TOPK method.

Attacks	Indicator	APS	RANK	RAPS	SAPS	TOPK
Clean	Coverage	0.955 (0.001)	0.937 (0.011)	0.955 (0.002)	0.938 (0.008)	0.937 (0.011)
	Size	269.403 (4.520)	257.710 (4.929)	269.351 (4.577)	257.284 (4.756)	257.565 (4.827)
	SSCV	0.086 (0.003)	0.037 (0.011)	0.086 (0.003)	0.038 (0.008)	0.037 (0.011)
FGSM	Coverage	0.977 (0.001)	0.963 (0.005)	0.977 (0.001)	0.962 (0.004)	0.963 (0.005)
	Size	269.117 (4.453)	257.692 (4.942)	269.065 (4.537)	257.111 (4.755)	257.564 (4.826)
	SSCV	0.081 (0.002)	0.063 (0.005)	0.081 (0.003)	0.062 (0.004)	0.063 (0.005)
PGD	Coverage	0.964 (0.001)	0.951 (0.003)	0.964 (0.001)	0.951 (0.003)	0.951 (0.003)
	Size	266.248 (4.478)	257.689 (4.944)	266.194 (4.567)	257.077 (4.754)	257.565 (4.826)
	SSCV	0.072 (0.004)	0.051 (0.003)	0.072 (0.005)	0.051 (0.003)	0.051 (0.003)
APGD	Coverage	0.955 (0.002)	0.936 (0.011)	0.955 (0.002)	0.938 (0.008)	0.936 (0.011)
	Size	273.920 (4.421)	257.713 (4.928)	273.868 (4.478)	257.329 (4.757)	257.564 (4.826)
	SSCV	0.086 (0.003)	0.036 (0.011)	0.086 (0.003)	0.038 (0.008)	0.036 (0.011)
CW	Coverage	0.986 (0.001)	0.974 (0.004)	0.986 (0.001)	0.974 (0.003)	0.974 (0.004)
	Size	278.106 (4.541)	257.706 (4.934)	278.054 (4.597)	257.318 (4.756)	257.565 (4.826)
	SSCV	0.090 (0.003)	0.074 (0.004)	0.090 (0.003)	0.074 (0.003)	0.074 (0.004)
PIFGSM	Coverage	0.901 (0.003)	0.900 (0.003)	0.901 (0.003)	0.900 (0.003)	0.900 (0.003)
	Size	271.377 (4.892)	257.689 (4.944)	271.325 (5.014)	257.065 (4.752)	257.566 (4.827)
	SSCV	0.138 (0.017)	0.003 (0.002)	0.143 (0.019)	0.002 (0.001)	0.003 (0.002)
GN	Coverage	0.987 (0.001)	0.976 (0.003)	0.987 (0.001)	0.975 (0.002)	0.975 (0.003)
	Size	253.404 (4.189)	257.686 (4.947)	253.355 (4.246)	257.015 (4.757)	257.566 (4.826)
	SSCV	0.092 (0.002)	0.076 (0.003)	0.092 (0.002)	0.075 (0.002)	0.075 (0.003)
Nash	Coverage	0.986 (0.001)	0.942 (0.018)	0.986 (0.001)	0.949 (0.020)	0.944 (0.026)
	Size	278.106 (4.541)	257.715 (4.929)	278.054 (4.597)	257.328 (4.758)	257.565 (4.827)
	SSCV	0.090 (0.003)	0.042 (0.018)	0.090 (0.003)	0.049 (0.020)	0.044 (0.026)

Table 3: Performance of various Conformal Prediction (CP) methods under the Nash defense strategy on the ImageNet dataset. The table presents the mean and standard deviation of coverage, size, and SSCV against a range of attack strategies. All results are averaged over 20 independent random splits.

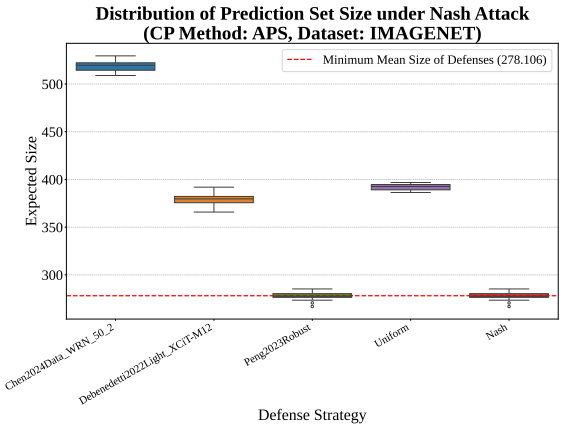

Figure 31: Box plots of size for different defense models using the APS method on the ImageNet dataset.

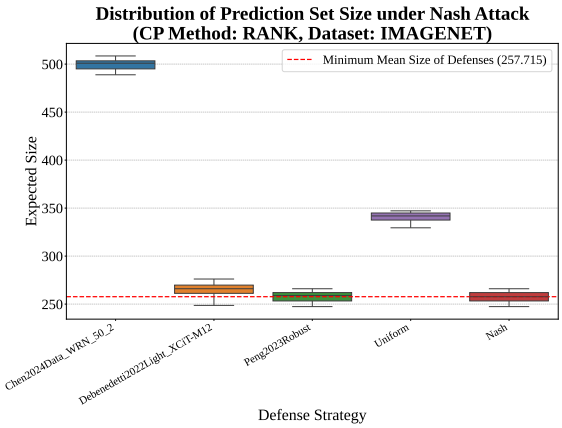

Figure 32: Box plots of size for different defense models using the RNAK method on the ImageNet dataset.

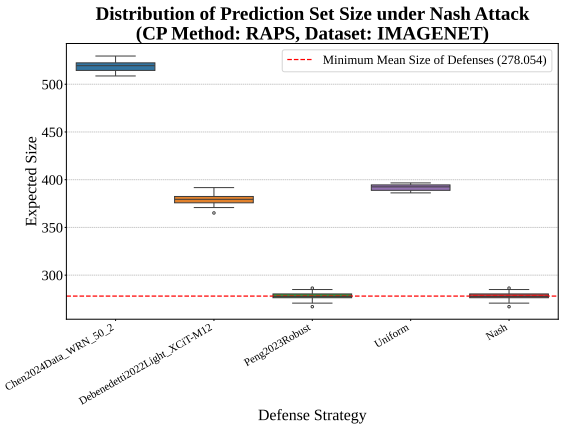

Figure 33: Box plots of size for different defense models using the RAPS method on the ImageNet dataset.

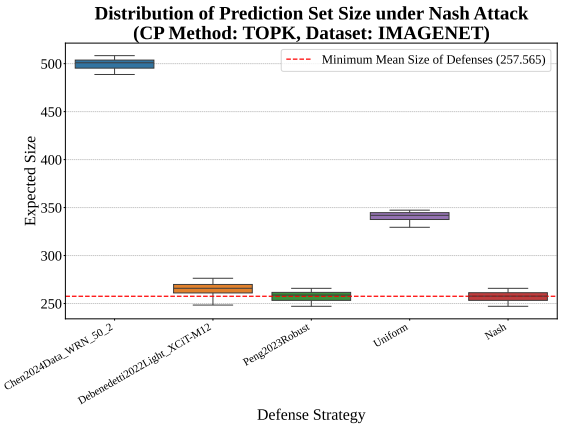

Figure 34: Box plots of size for different defense models using the TOPK method on the ImageNet dataset.

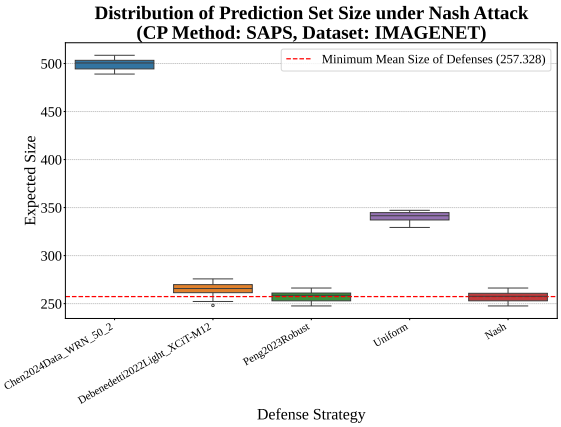

Figure 35: Box plots of size for different defense models using the SAPS method on the ImageNet dataset.

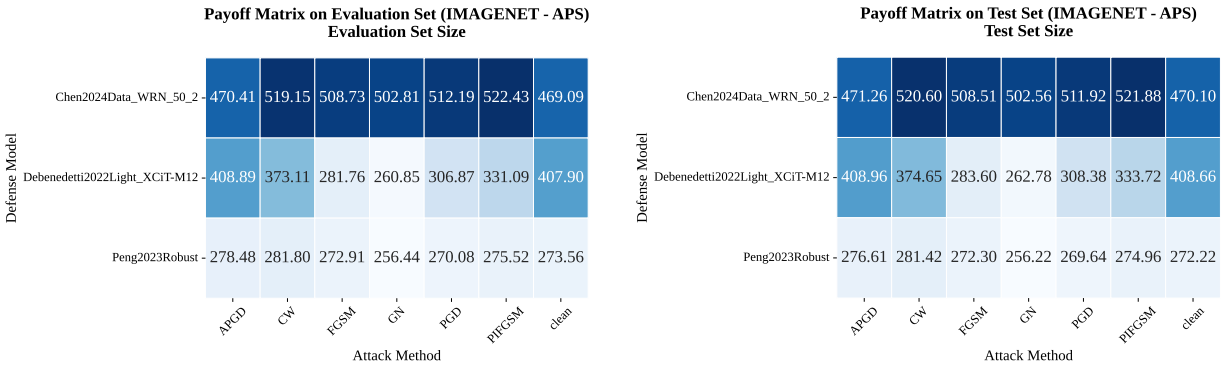

Figure 36: Evaluate the payoff matrix on both the ImageNet dataset's evaluation set and test set using the APS method.

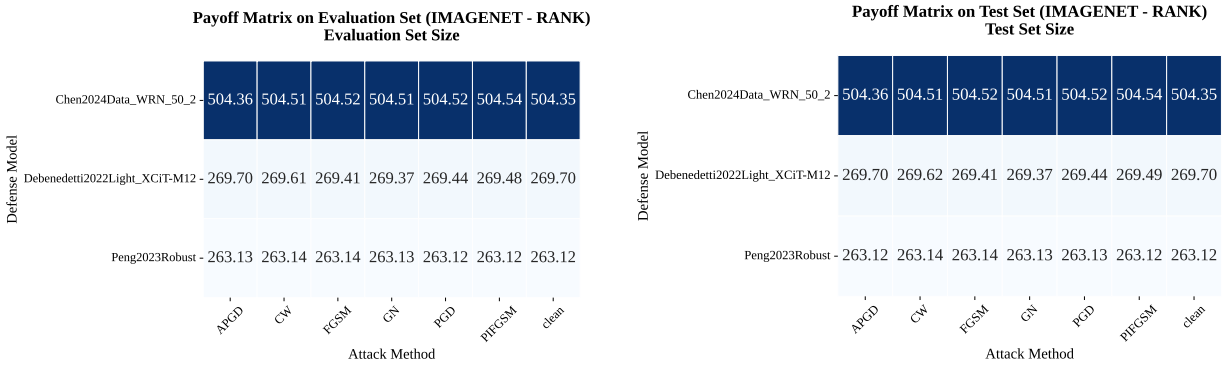

Figure 37: Evaluate the payoff matrix on both the ImageNet dataset's evaluation set and test set using the RANK method.

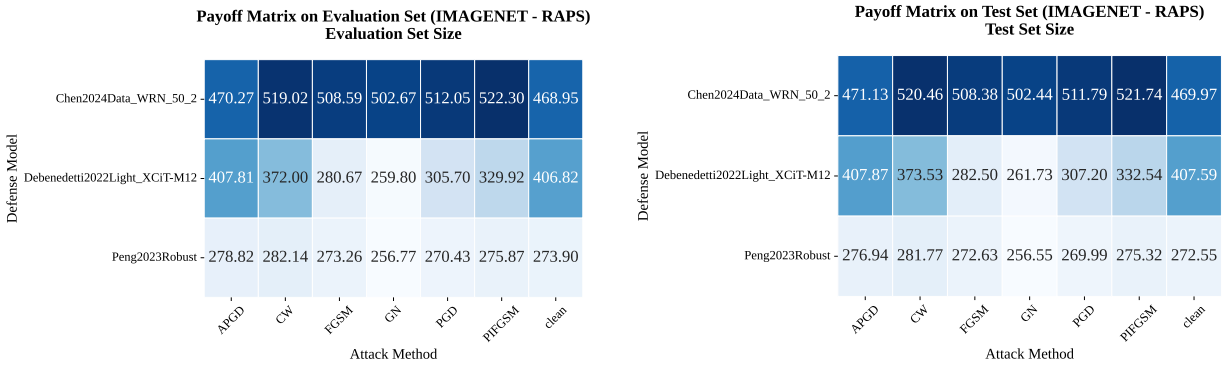

Figure 38: Evaluate the payoff matrix on both the ImageNet dataset's evaluation set and test set using the RAPS method.

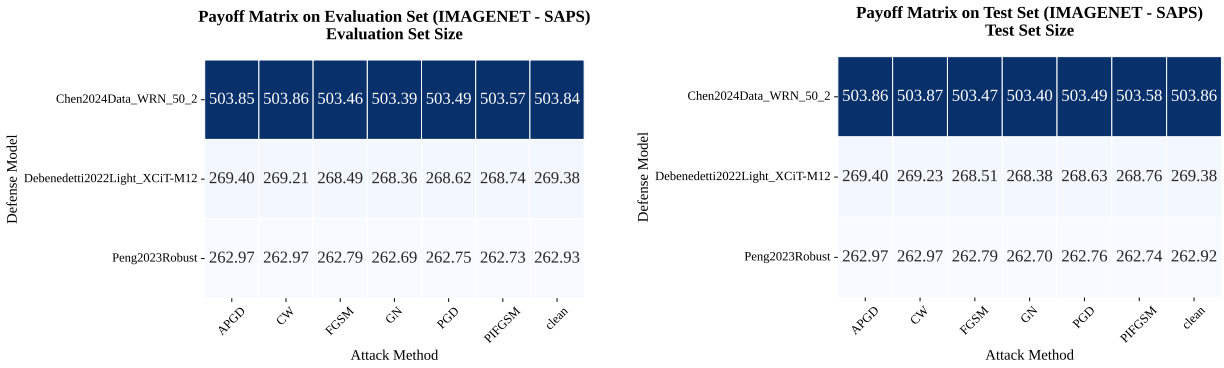

Figure 39: Evaluate the payoff matrix on both the ImageNet dataset's evaluation set and test set using the SAPS method.

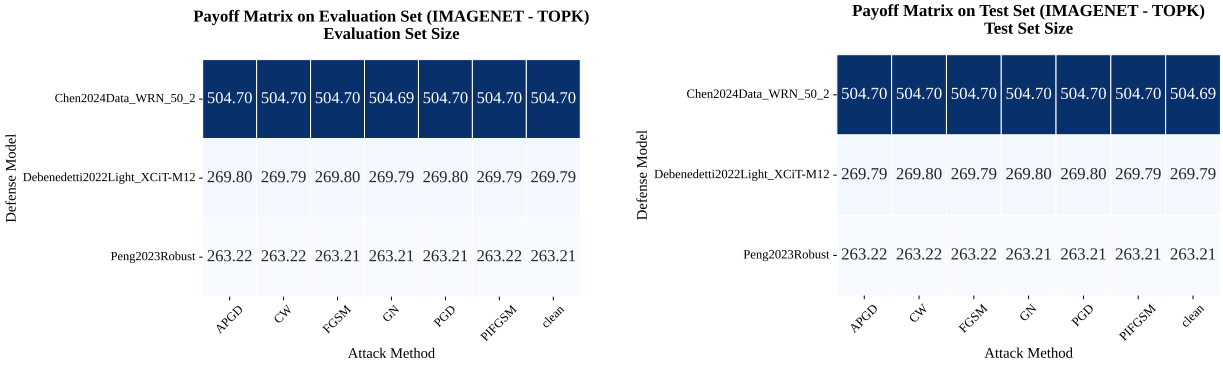

Figure 40: Evaluate the payoff matrix on both the ImageNet dataset's evaluation set and test set using the TOPK method.

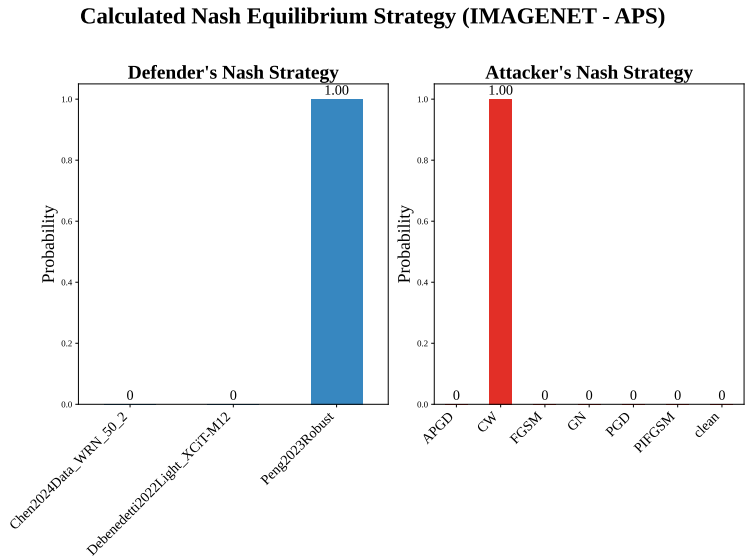

Figure 41: Obtain the Nash attack and Nash defense strategies on the ImageNet dataset using the APS method.

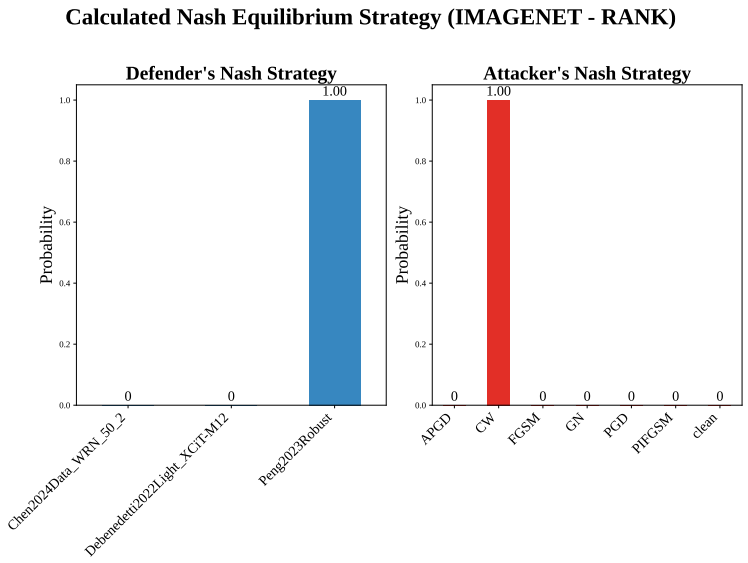

Figure 42: Obtain the Nash attack and Nash defense strategies on the ImageNet dataset using the RANK method.

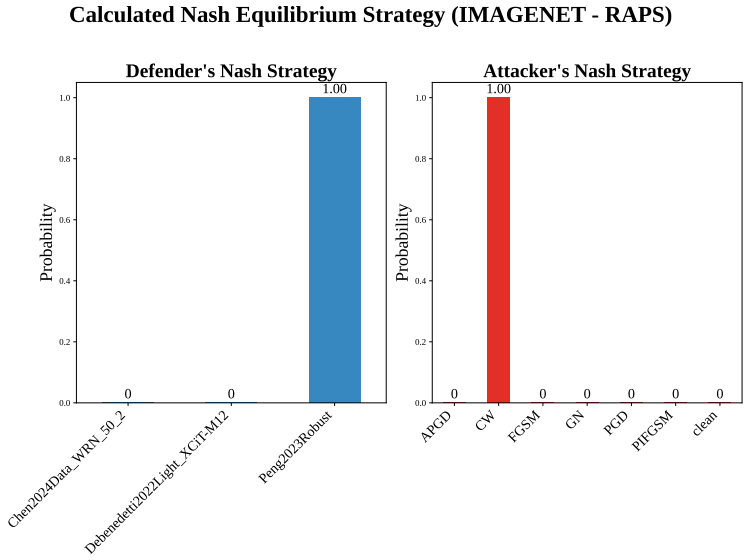

Figure 43: Obtain the Nash attack and Nash defense strategies on the ImageNet dataset using the RAPS method.

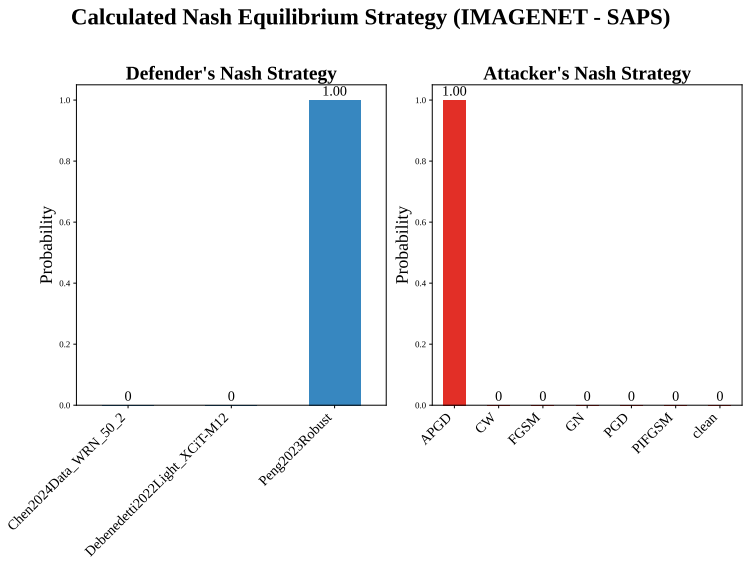

Figure 44: Obtain the Nash attack and Nash defense strategies on the ImageNet dataset using the SAPS method.

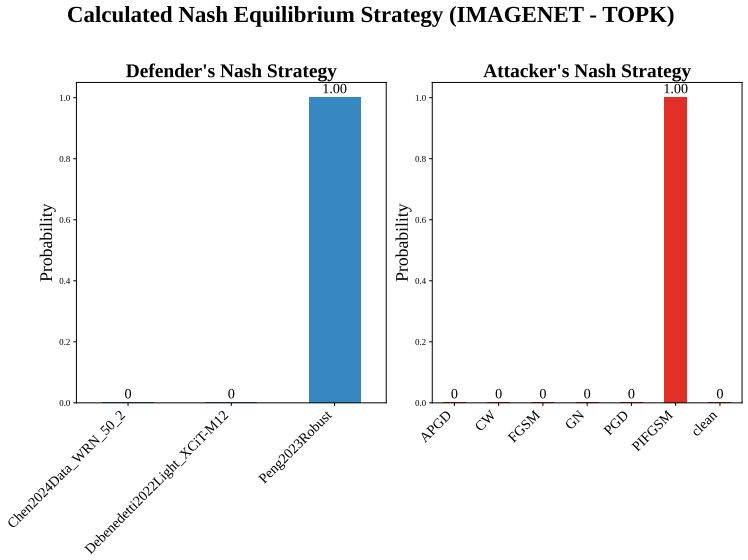

Figure 45: Obtain the Nash attack and Nash defense strategies on the ImageNet dataset using the TOPK method.