ATTACK-FREE ADVERSARIAL VULNERABILITY ASSESSMENT VIA SECOND-ORDER STATISTICS OF PREDICTIVE PROBABILITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are widely deployed in safety-critical applications such as medical image analysis and autonomous driving, where adversarial robustness is essential. Existing robustness evaluations primarily measure overall performance degradation under attack, offering limited insight into class-wise or directional vulnerability patterns. We propose a second-order statistical framework that treats class-wise softmax outputs as random variables and analyzes their conditional variance–covariance structure to characterize structural vulnerability. Using only clean validation data-without generating adversarial examples or relying on specific attack algorithms-the framework enables attack-agnostic evaluation while significantly reducing computational cost, serving as a practical pre-screening tool for rapid model assessment. The conditional variance of the correct-class probability quantifies relative class-wise adversarial vulnerability, while conditional covariance reveals directional misclassification tendencies. Experiments on standard benchmarks under diverse untargeted and targeted attack settings demonstrate strong alignment between the proposed metrics and class-wise attack success rates, suggesting that adversarial vulnerability can be effectively characterized without explicit attack generation.

1 INTRODUCTION

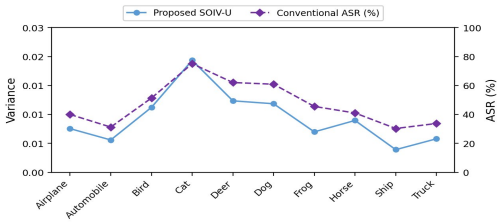
Deep classification models are vulnerable to small adversarial perturbations that can significantly alter prediction outcomes (Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2017; Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Kurakin et al., 2018; Andriushchenko et al., 2020; Eykholt et al., 2018; Yuan et al., 2019; Chakraborty et al., 2018). This raises fundamental questions about how robustness should be properly evaluated and interpreted beyond simple performance comparison.

Existing adversarial robustness evaluations primarily rely on aggregated metrics such as accuracy degradation or attack success rate (ASR) to assess overall vulnerability (Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Andriushchenko et al., 2020; Yuan et al., 2019; Chakraborty et al., 2018; Sharma & Chen, 2024; Croce & Hein, 2020; Papernot et al., 2017; Croce et al., 2020). While effective for overall comparison, such metrics obscure structural characteristics, including how vulnerabilities are distributed across classes and how misclassifications are directionally structured.

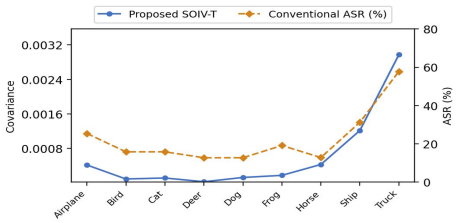
Many robustness analyses rely on generating adversarial examples and evaluating performance under specific threat models (Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Andriushchenko et al., 2020; Croce & Hein, 2020; Croce et al., 2020; Zhang et al., 2019; Engstrom et al., 2017). However, the results can vary depending on implementation details or hyperparameter settings and often require substantial computational cost. A notable exception is CLEVER (Weng et al., 2018), which avoids adversarial example generation but provides a single robustness score per input rather than class-wise or directional vulnerability structure.

We propose a second-order statistical framework that models class-wise softmax outputs as random variables and analyzes their conditional variance–covariance structure using clean validation data. The conditional variance of the correct-class probability quantifies relative class-wise vulnerability, while conditional inter-class covariance captures directional misclassification tendencies.

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.



Untargeted (FGSM)



Targeted (PGD, Source class: Automobile)

Figure 1: Class-wise adversarial vulnerability estimated by the proposed SOIV metrics and conventional ASR on CIFAR-10 using ResNet-18. SOIV-U (variance-based) and SOIV-T (covariance-based) are computed solely from clean validation data, without generating adversarial examples. (Left) Untargeted attack setting (FGSM): proposed SOIV-U shows consistent alignment with class-wise ASR. (Right) Targeted attack setting (PGD, source class: Automobile): SOIV-T exhibits strong agreement with class-wise ASR. Despite requiring no adversarial optimization, both metrics effectively characterize class-wise adversarial vulnerability without explicit attack generation.

The framework provides an attack-agnostic and computationally efficient approach for robustness diagnosis, enabling fine-grained analysis of structural vulnerability patterns.

The proposed metrics show strong alignment with class-wise attack success rates across both untargeted and targeted attack settings, without requiring adversarial example generation.

Our contributions are summarized as follows:

1. We propose an attack-agnostic framework that models class-wise softmax outputs as random variables in clean settings to diagnose adversarial vulnerability without generating adversarial examples.
2. We quantify class-wise vulnerability using correct-class variance and analyze directional misclassification tendencies via inter-class covariance.
3. We demonstrate strong alignment between the proposed metrics and class-wise attack success rates across diverse attacks, supporting the framework as an efficient tool for adversarial vulnerability diagnosis.

2 Related Works

2.1 Adversarial Attack Methods

Since the adversarial vulnerability of deep neural networks was reported by Szegedy et al. (2013), various adversarial attack methods have been proposed (Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Kurakin et al., 2018; Andriushchenko et al., 2020; Croce & Hein, 2020; Kurakin et al., 2016; Tramèr et al., 2017; Schmidt et al., 2018; Rice et al., 2020; Moosavi-Dezfooli et al., 2017a; Su et al., 2019; Narodytska & Kasiviswanathan, 2016; Brendel et al., 2017). Major adversarial attack methods include FGSM (Goodfellow et al., 2014; Kurakin et al., 2016; Tramèr et al., 2017; Narodytska & Kasiviswanathan, 2016), PGD (Madry et al., 2017; Croce & Hein, 2020; Schmidt et al., 2018; Rice et al., 2020), BIM (Kurakin et al., 2018; 2016), C&W (Carlini & Wagner, 2017; Croce & Hein, 2020) , and DeepFool (Moosavi-Dezfooli et al., 2016; 2017a). All of these are white-box attacks, meaning the attacker has full knowledge of the model, including its internal architecture, hyperparameters, and gradients of the loss function.

In contrast, black-box attacks (Andriushchenko et al., 2020; Papernot et al., 2017; Tramèr et al., 2017; Su et al., 2019; Narodytska & Kasiviswanathan, 2016; Brendel et al., 2017; Liu et al., 2016; Chen et al., 2017; Ilyas et al., 2018; Chen et al., 2020; Cheng et al., 2019) assume limited access to the model, where the attacker can only rely on inputs and final outputs. Depending on the type of information available to an attacker, black-box attacks are largely divided into transfer-based attack (Papernot et al., 2017; Tramèr et al., 2017; Liu et al., 2016), score-based attack (Andriushchenko et al., 2020; Su et al., 2019; Narodytska & Kasiviswanathan, 2016; Chen et al., 2017; Ilyas et al., 2018), and decision-based attack (Brendel et al., 2017; Chen et al., 2020; Cheng et al., 2019). Examples of black-box attacks include Square Attack (Andriushchenko et al., 2020) and One-Pixel Attack

(Su et al., 2019; Narodytska & Kasiviswanathan, 2016). These diverse attack methods have been widely used in research to analyze and evaluate the vulnerability of deep learning models (Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Andriushchenko et al., 2020; Croce & Hein, 2020; Croce et al., 2020; Zhang et al., 2019; Tramèr et al., 2017; Schmidt et al., 2018; Rice et al., 2020; Moosavi-Dezfooli et al., 2017a; Su et al., 2019; Narodytska & Kasiviswanathan, 2016; Athalye et al., 2018; Uesato et al., 2018; Brendel et al., 2017; Tsipras et al., 2018; Carlini et al., 2019).

2.2 Attack-based Robustness Evaluation and Its Limitations

Adversarial attacks have been extensively studied as a primary means for evaluating model robustness, with various evaluation metrics derived from attack-based experiments (Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Andriushchenko et al., 2020; Croce & Hein, 2020; Croce et al., 2020; Moosavi-Dezfooli et al., 2017a; Athalye et al., 2018; Uesato et al., 2018; Brendel et al., 2017; Carlini et al., 2019). However, such approaches typically require substantial computational resources and time, as adversarial examples must be generated (Gao et al., 2022; Shafahi et al., 2019). When diagnosing class-wise vulnerabilities, attacks must be performed repeatedly for each class, which is inefficient due to repetitive gradient computations and optimization procedures. This problem becomes more severe as the dataset becomes larger, the number of classes increases, and the attack methods become more complex, as seen in methods such as C&W and DeepFool (Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Rony et al., 2019). Also, results may vary depending on the attack configuration or the hyperparameter settings (Croce & Hein, 2020; Wei et al., 2018; Zimmer & Karame, 2025). This raises the need for an efficient and attack-agnostic approach that can evaluate vulnerability in advance, without generating adversarial examples.

2.3 Class-wise Vulnerability Analysis

Class-wise characteristics have been widely studied, as they often reveal important structural properties that aggregate metrics obscure. For example, research on long-tailed data learning (Zhang et al., 2023) and noisy label problems (Natarajan et al., 2013) demonstrates that class-wise analysis can uncover class-specific weaknesses that overall performance metrics fail to capture. This insight is directly applicable to adversarial robustness evaluation.

On the other hand, relatively few studies in adversarial robustness focus on class-wise characteristics beyond aggregate metrics. Instead, many studies have used aggregate indicators, such as average accuracy and average ASR, as representative measures of vulnerability (Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Andriushchenko et al., 2020; Croce & Hein, 2020; Croce et al., 2020; Zhang et al., 2019; Schmidt et al., 2018; Rice et al., 2020; Moosavi-Dezfooli et al., 2017a; Su et al., 2019; Narodytska & Kasiviswanathan, 2016; Athalye et al., 2018; Uesato et al., 2018; Brendel et al., 2017; Tsipras et al., 2018). This approach has a limitation in that it cannot sufficiently explain class-wise differences in vulnerability or the structure of the model's vulnerability distribution. This is because even if vulnerabilities occur in certain classes, they can be averaged and hidden by other robust classes.

Furthermore, although the number of samples per class is balanced, the inter-class distances or the density of distributions in the feature space may not be uniform (Moosavi-Dezfooli et al., 2016; Papyan et al., 2020). In CIFAR-10, visually similar class pairs (e.g., *cat*/*dog*, *automobile*/*truck*) may have closely positioned decision boundaries, making them easier to cross. Therefore, a dedicated class-wise analysis is essential for a complete understanding of a model's vulnerability structure, and motivates the need for evaluation frameworks that go beyond class-averaged metrics.

3 Method

3.1 Problem Setting

In this work, we consider the multi-class classification problem consisting of C classes. For the input image $\mathbf{x} \in \mathcal{X}$, the pre-trained classification model f_θ returns the softmax output probability vector $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_C(\mathbf{x}))$, where $p_c(\mathbf{x})$ denotes the predicted probability that the input $\mathbf{x}$ belongs to the class c. Throughout this paper, we fix the model parameters θ and focus on analyzing the statistical properties of the output probability distribution observed under clean inputs.

For the analysis, we use a validation dataset $\mathcal{D}_{\text{val}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, which is a portion of the original dataset held out from training (typically reserved for model selection, hyperparameter tuning, and related purposes), where $y_i \in \{1, \ldots, C\}$ is the ground-truth label of the input $\mathbf{x}_i$. For each sample, we define the softmax probability corresponding to the true class as $p_{\text{true}}(\mathbf{x}_i) = p_{y_i}(\mathbf{x}_i)$.

Throughout the paper, we use uppercase X to denote a random input and lowercase $\mathbf{x}$ (or $\mathbf{x}_i$) to denote a realized input sample.

Our analysis focuses on how unstable the model's output probabilities are across samples in the clean setting. Such variability reflects how consistently the model maintains confidence for a given class, and, when predictions become unstable, in which direction misclassification is likely to occur. Accordingly, we treat the output probabilities observed on the validation dataset as random variables, and characterize class-wise vulnerability through their conditional variance and covariance structure without relying on any specific adversarial attack method.

3.2 Class-wise Variance for Untargeted Adversarial Vulnerability

We define a class-wise conditional variance measure to quantify the variability of the true-class probability under clean inputs. Even for samples of the same class, the assigned confidence may vary, reflecting prediction stability. In the untargeted attack setting, where the objective is to reduce the true-class probability, larger clean confidence fluctuations can be associated with stronger local probability sensitivity. Thus, we use the variance of the true-class probability as a clean-data indicator of relative untargeted vulnerability.

Let $\mathcal{D}_{\text{val}}^{(c)}$ denote the subset of validation samples whose ground-truth class is c. We define the conditional sample variance of the true-class probability for class c as

$$\text{Var}_c^{\text{sample}} = \frac{1}{|\mathcal{D}_{\text{val}}^{(c)}| - 1} \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{val}}^{(c)}} \left(p_c(\mathbf{x}_i) - \overline{p}_c^{(c)} \right)^2, \tag{1}$$

where $\overline{p}_{\text{c}}^{(c)}$ is the sample mean of the true-class probability for class c, defined as $\overline{p}_{\text{c}}^{(c)} = \frac{1}{|\mathcal{D}_{\text{val}}^{(c)}|} \sum_{\mathbf{x}_i \in \mathcal{D}_{val}^{(c)}} p_c(\mathbf{x}_i)$.

Beyond variability around the empirical mean, we consider deviation from the ideal confidence value of 1. The ideal-centered conditional variance is defined as

$$\text{Var}_c^{\text{ideal}} = \frac{1}{|\mathcal{D}_{\text{val}}^{(c)}|} \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{val}}^{(c)}} \left(p_{\text{c}}(\mathbf{x}_i) - 1 \right)^2. \tag{2}$$

This quantity captures the extent to which the true-class probability deviates from perfect confidence under clean inputs. We denote this measure as the Second-Order Intrinsic Vulnerability for Untargeted settings (SOIV-U). That is,

$$\text{SOIV-U}(c) \triangleq \text{Var}_c^{\text{ideal}}. \tag{3}$$

Larger SOIV-U values indicate structurally unstable predictions for class c, implying higher intrinsic vulnerability in the untargeted attack setting.

3.3 Covariance-based Directional Adversarial Vulnerability

In the targeted attack setting, the adversary seeks to misclassify inputs from a source class into a designated target class. Beyond overall instability, it is therefore necessary to analyze how probability mass shifts when the source-class confidence decreases.

To quantify this directional behavior, we analyze the relationship between the true-class probability $p_s(\mathbf{x}_i)$ and the target-class probability $p_t(\mathbf{x}_i)$, for samples whose ground-truth class is s, where $t \neq s$.

For a source class s, we consider the set of validation samples whose ground-truth label is s, denoted by $\mathcal{D}_{\text{val}}^{(s)}$. We first define the conditional sample covariance between these two probabilities as

$$\text{Cov}_{s \to t}^{\text{sample}} = \frac{1}{|\mathcal{D}_{\text{val}}^{(s)}| - 1} \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{val}}^{(s)}} \left(p_s(\mathbf{x}_i) - \overline{p}_s^{(s)} \right) \left(p_t(\mathbf{x}_i) - \overline{p}_t^{(s)} \right). \tag{4}$$

where $\overline{p}_s^{(s)}$ and $\overline{p}_t^{(s)}$ are the mean source-class and target-class probabilities over $\mathcal{D}_{\text{val}}^{(s)}$, respectively.

Beyond mean-centered covariance, we consider deviation from the ideal confidence value of 1 for the source class and the baseline value of 0 for non-ground-truth classes. Specifically, we define the ideal-centered conditional covariance as

$$\text{Cov}_{s\to t}^{\text{ideal}} = \frac{1}{|\mathcal{D}_{\text{val}}^{(s)}|} \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{val}}^{(s)}} (p_s(\mathbf{x}_i) - 1)\, p_t(\mathbf{x}_i). \tag{5}$$

This quantity captures the extent to which, even in the clean setting, a decrease in confidence for the source class is structurally associated with an increase in probability for a particular target class. In other words, it reflects whether misclassification tendencies are directionally structured in the probability landscape.

We interpret the ideal-centered covariance $\text{Cov}_{s\to t}^{\text{ideal}}$ as a second-order intrinsic vulnerability measure for the targeted attack setting, and denote it as the Second-Order Intrinsic Vulnerability for Targeted settings (SOIV-T). That is,

$$\text{SOIV-T}(s \to t) \triangleq \text{Cov}_{s\to t}^{\text{ideal}}. \tag{6}$$

This indicates whether a decrease in the source-class confidence is systematically accompanied by an increase in the target-class probability. Since $p_s(\mathbf{x}_i)$ and $p_t(\mathbf{x}_i)$ lie in $[0, 1]$, SOIV-T always yields a non-positive value. A larger magnitude indicates stronger directional probability shifts from the source class to the target class, implying higher intrinsic directional vulnerability.

3.4 Motivation: From Clean Probability Variability to Local Adversarial Sensitivity

We next provide theoretical motivation for why the clean-data statistics defined in Sections 3.2 and 3.3 can be informative about adversarial vulnerability. Recent studies have associated adversarial vulnerability with the sharpness or local sensitivity of the loss and output-probability landscape (Su et al., 2019; Ye et al., 2024; Zheng et al., 2023; Zhang et al., 2024; Foret et al., 2020; Moosavi-Dezfooli et al., 2017b). Our central premise is that signatures of such sensitivity and directional structure can, to some extent, already be reflected in the distribution of model outputs on clean data.

Untargeted vulnerability For class c, let $f_c(\mathbf{x}) = p_c(\mathbf{x})$ denote the true-class probability, and let $X, X' \overset{\text{i.i.d.}}{\sim} P(X \mid Y = c)$ be two independent clean samples from the same class. The clean class-probability variance admits the exact pairwise representation

$$\text{Var}_{X|Y=c}[f_c(X)] = \frac{1}{2}\mathbb{E}_{X,X'|Y=c}\left[\left(f_c(X') - f_c(X)\right)^2\right]. \tag{7}$$

Under a local first-order approximation,

$$f_c(X') - f_c(X) \approx \nabla f_c(X)^\top (X' - X). \tag{8}$$

Thus, when natural within-class variations sufficiently explore locally probability-sensitive directions, greater clean class-probability variability reflects stronger local probability sensitivity.

The same sample-wise gradient governs the first-order response to an adversarial perturbation δ_X:

$$f_c(X + \delta_X) - f_c(X) \approx \nabla f_c(X)^\top \delta_X. \tag{9}$$

Moreover, for the cross-entropy loss $\ell(X, c) = -\log f_c(X)$,

$$\nabla_X \ell(X, c) = -\frac{1}{f_c(X)} \nabla f_c(X), \tag{10}$$

showing that the true-class probability gradient is directly related to the local loss sensitivity exploited by gradient-based untargeted attacks.

These observations motivate the variance component of SOIV-U as an attack-free signature of local adversarial probability sensitivity. The connection is conditional rather than universal: it depends on local smoothness and sufficient coverage of probability-sensitive directions by natural within-class variation, and it does not imply a direct equivalence with attack success rate. Appendix Section A.1 formalizes these conditions, derives the corresponding clean-to-adversarial sensitivity bounds, and discusses higher-order effects and different perturbation norms.

Targeted vulnerability For targeted vulnerability, the relevant structure is source–target specific. The magnitude of the sample-wise SOIV-T term can be written as

$$q_{s\to t}^{\text{clean}}(X) = (1 - p_s(X))\, p_t(X). \tag{11}$$

This quantity is large when low source-class confidence co-occurs with high probability assigned to a particular target class. Since $q_{s\to t} = (1 - p_s)p_t$ decreases monotonically with p_s and increases monotonically with p_t, the characteristic source-to-target probability response, $p_s \downarrow$ and $p_t \uparrow$, necessarily increases $q_{s\to t}$. Thus, $q_{s\to t}$ provides a continuous measure of target-directed structure in the model output.

To establish the clean-to-adversarial relationship, consider a clean sample $X \sim P(X \mid Y = s)$ and its adversarial counterpart $X^{\text{adv}} = X + \delta_X$. We evaluate the same directional statistic after perturbation as

$$q_{s\to t}^{\text{adv}}(X) = \left(1 - p_s(X^{\text{adv}})\right) p_t(X^{\text{adv}}). \tag{12}$$

For the characteristic source-to-target response,

$$p_s(X^{\text{adv}}) \le p_s(X), \qquad p_t(X^{\text{adv}}) \ge p_t(X), \tag{13}$$

the monotonicity of $q_{s\to t}$ gives

$$q_{s\to t}^{\text{adv}}(X) \ge q_{s\to t}^{\text{clean}}(X). \tag{14}$$

This pointwise relation compares the clean and adversarial responses of the same underlying sample. Consequently, averaging over $X \sim P(X \mid Y = s)$ preserves the inequality at the population level:

$$\mathbb{E}_{X|Y=s}\left[q_{s\to t}^{\text{adv}}(X)\right] \ge \mathbb{E}_{X|Y=s}\left[q_{s\to t}^{\text{clean}}(X)\right]. \tag{15}$$

The clean expectation corresponds to the population-level counterpart of the magnitude of SOIV-T,

$$\mathbb{E}_{X|Y=s}\left[q_{s\to t}^{\text{clean}}(X)\right] = \left|\text{SOIV-T}(s \to t)\right|, \tag{16}$$

and therefore represents the clean-data version of the same target-directed statistic measured under adversarial perturbation.

This provides the intuition for using SOIV-T as an attack-free indicator of pair-specific directional vulnerability: under the characteristic source-to-target response, adversarial perturbation strengthens the same source–target structure that is already measured from clean outputs. Appendix Section A.2 provides the exact clean-to-adversarial decomposition and clarifies the scope of this relationship.

Importantly, the relation above is a population-level statement in which adversarial examples are generated from clean inputs drawn from the same underlying source-class distribution. In practice, the clean statistic and its adversarial counterpart may be estimated from finite and disjoint evaluation sets whose underlying clean samples follow the same source-class distribution, so their empirical values need not satisfy the population-level relation exactly. Moreover, the relative source–target structure is not guaranteed to be preserved under every attack. We therefore treat its preservation as an empirical hypothesis and validate the correspondence between the clean and adversarial directional statistics, together with attack-based vulnerability, in our experiments.

4 Experiments

4.1 SOIV-U Aligns with Untargeted Adversarial Vulnerability

We analyze how SOIV-U computed from clean validation data corresponds to class-wise vulnerability under untargeted attacks. Table 1 compares the class-wise rankings of SOIV-U and ASR across FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED.

Under FGSM, the most vulnerable classes, including *cat*, *deer*, and *dog*, also appear at the top of the SOIV-U ranking, while *automobile* and *ship* are ranked among the least vulnerable. SOIV-U shows strong alignment with FGSM ASR, with Pearson, Spearman, and Kendall correlations of 0.959, 0.964, and 0.911, respectively.

This trend remains consistent across the other untargeted attacks, with Pearson, Spearman, and Kendall correlation coefficients of 0.878–0.936, 0.915–0.939, and 0.778–0.822, respectively, indicating strong consistency between the two metrics.

Table 1: Comparison between SOIV-U ranking and ASR ranking under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED). The second column lists classes sorted in descending order of SOIV-U (larger values indicate stronger untargeted vulnerability), formatted as *class name (SOIV-U value)*. Each subsequent column reports the ASR under the corresponding untargeted attack, sorted in descending order of ASR, and formatted as *class name (ASR value)*. The MIXED setting aggregates multiple untargeted attacks to construct a combined threat environment and evaluate whether SOIV-U consistently reflects the overall vulnerability structure across diverse attack strategies. The bottom rows provide the Pearson, Spearman, and Kendall correlation coefficients between SOIV-U ranking and ASR ranking for each attack. Attack success rates (ASRs) are reported as percentages. For the bottom correlation rows, values in parentheses denote two-sided p-values for testing the significance of the corresponding correlation coefficients.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	AUTOATTACK	MIXED
1	cat (1.93e-2)	cat (75.07)	cat (78.25)	cat (75.60)	cat (86.60)	cat (81.20)	cat (79.34)
2	deer (1.23e-2)	deer (62.04)	deer (67.55)	deer (73.28)	deer (78.10)	deer (71.10)	deer (70.41)
3	dog (1.18e-2)	dog (60.87)	dog (62.58)	dog (62.09)	dog (71.81)	dog (63.80)	dog (64.23)
4	bird (1.12e-2)	bird (51.32)	bird (55.76)	bird (58.87)	bird (66.67)	bird (58.80)	bird (58.28)
5	horse (8.96e-3)	frog (45.51)	frog (50.81)	frog (53.41)	frog (60.76)	frog (52.60)	frog (52.62)
6	airplane (7.54e-3)	horse (40.99)	airplane (43.04)	airplane (44.36)	airplane (55.31)	airplane (45.20)	airplane (45.58)
7	frog (6.99e-3)	airplane (39.98)	horse (42.42)	horse (40.77)	horse (52.31)	horse (44.20)	horse (44.14)
8	truck (5.78e-3)	truck (33.72)	truck (33.95)	ship (35.71)	ship (46.59)	ship (36.50)	ship (36.58)
9	automobile (5.59e-3)	automobile (31.24)	ship (33.80)	truck (32.10)	truck (43.96)	truck (35.70)	truck (35.89)
10	ship (3.91e-3)	ship (30.28)	automobile (30.38)	automobile (28.15)	automobile (37.31)	automobile (31.80)	automobile (31.78)
Pearson		0.959 (1.2e-05)	0.936 (6.8e-05)	0.878 (8.3e-04)	0.912 (2.4e-04)	0.933 (8.1e-05)	0.927 (1.1e-04)
Spearman		0.964 (7.0e-06)	0.939 (5.6e-05)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)
Kendall		0.911 (3.0e-05)	0.822 (3.6e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)

The consistency also holds under MIXED, which aggregates multiple untargeted attacks. SOIV-U achieves Pearson, Spearman, and Kendall correlations of 0.927, 0.915, and 0.778, respectively, indicating that it captures stable class-wise vulnerability patterns across diverse attack settings.

4.2 SOIV-T Aligns with Targeted Adversarial Vulnerability

We evaluate whether SOIV-T computed from clean validation data aligns with directional vulnerability under targeted attacks. Since targeted attacks concern specific source–target class pairs, we analyze a representative case with source = *automobile* and compare SOIV-T with targeted ASR across target classes.

Under the C&W targeted attack, *truck* is the most vulnerable target and also has the largest-magnitude SOIV-T value. SOIV-T and C&W ASR show strong alignment, with Pearson, Spearman, and Kendall correlations of 0.968, 0.700, and 0.611, respectively. This agreement indicates that the directional relationship observed under attack is also reflected in clean softmax covariance.

As summarized in Table 2, similar trends hold across FGSM, PGD, C&W, and MIXED, with *truck* remaining the top-ranked target in all settings. SOIV-T maintains high Pearson correlations of 0.968–0.975 together with consistently positive rank-based correlations (Spearman and Kendall), indicating stable source-target vulnerability patterns across different targeted attacks.

4.3 Unified SOIV Matrix for Structural Adversarial Vulnerability Analysis

In this section, we integrate SOIV-U and SOIV-T within a unified variance–covariance matrix constructed from clean validation data, as illustrated in Fig. 2. The diagonal entries correspond to SOIV-U values and represent class-wise vulnerability under untargeted settings, while the off-diagonal entries correspond to SOIV-T values and capture directional vulnerability between source–target class pairs. The matrix is visualized as a heatmap, where color intensity reflects the magnitude of each entry, enabling intuitive identification of vulnerable classes and high-risk source-target class pairs at a glance. This unified structure enables simultaneous analysis of class-level and directional vulnerabilities within a single statistical framework.

Specifically, the diagonal entries reveal clear class-wise vulnerability patterns, with *cat*, *deer*, and *dog* exhibiting relatively large SOIV-U values, indicating higher vulnerability, while *ship*, *truck*, and *automobile* show smaller values indicating comparatively stable prediction confidence. The off-diagonal entries describe directional vulnerability patterns, where larger-magnitude SOIV-T values indicate stronger directional misclassification tendencies toward a particular target class. For instance, larger-magnitude values are observed for *automobile* $\rightarrow$ *truck* and *automobile* $\rightarrow$ *ship*, as

Table 2: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (FGSM, PGD, C&W, and MIXED) when the source class is *automobile*. The second column lists target classes sorted in ascending order of SOIV-T (more negative values indicate stronger directional vulnerability), formatted as *class name (SOIV-T value)*. Each subsequent column reports the ASR under the corresponding targeted attack, formatted as *class name (ASR value)*. The MIXED setting aggregates multiple targeted attacks to simulate a combined threat environment and evaluate whether SOIV-T consistently reflects directional vulnerability across diverse attack strategies. For correlation analysis, the absolute value of SOIV-T is used since SOIV-T is inherently non-positive while ASR is non-negative. The bottom rows provide the Pearson, Spearman, and Kendall correlation coefficients between SOIV-T and each ASR. Values in parentheses in the bottom rows are two-sided p-values for testing the significance of the corresponding correlation coefficients.

Rank	SOIV-T	FGSM	PGD	C&W	MIXED
1	truck (-0.002972)	truck (65.78)	truck (57.78)	truck (64.93)	truck (62.83)
2	ship (-0.001217)	ship (25.27)	ship (31.24)	ship (37.42)	ship (31.31)
3	horse (-0.000430)	frog (17.38)	airplane (25.37)	airplane (23.67)	airplane (22.00)
4	airplane (-0.000418)	airplane (16.95)	frog (19.08)	frog (23.35)	frog (19.94)
5	frog (-0.000178)	bird (10.77)	bird (15.78)	cat (14.82)	bird (13.58)
6	dog (-0.000129)	horse (7.89)	cat (15.78)	bird (14.18)	cat (12.65)
7	cat (-0.000117)	cat (7.36)	horse (12.79)	dog (13.01)	horse (10.98)
8	bird (-0.000094)	deer (7.25)	deer (12.69)	horse (12.26)	dog (10.84)
9	deer (-0.000030)	dog (6.82)	dog (12.69)	deer (10.45)	deer (10.13)
Pearson		0.973757 (9.4e-06)	0.970146 (1.5e-05)	0.968151 (1.8e-05)	0.975086 (7.8e-06)
Spearman		0.733333 (2.5e-02)	0.722715 (2.8e-02)	0.700000 (3.6e-02)	0.733333 (2.5e-02)
Kendall		0.555556 (4.5e-02)	0.628828 (1.8e-02)	0.611111 (2.5e-02)	0.611111 (2.5e-02)

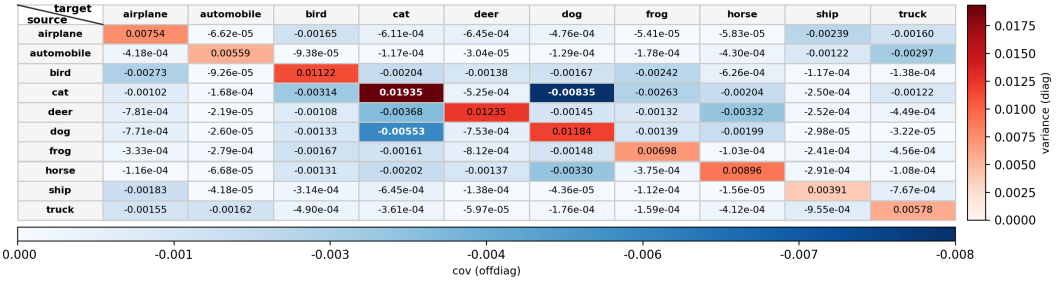


Figure 2: Unified SOIV matrix computed from the clean validation data. The diagonal entries represent the SOIV-U values for each class, while the off-diagonal entries represent the SOIV-T values between class pairs. Color intensity reflects the magnitude of each entry, enabling intuitive identification of vulnerable classes and high-risk source-target pairs. This matrix can be interpreted as an integrated representation that simultaneously summarizes class-level vulnerability and directional vulnerability structure within a single formulation.

well as semantically related pairs such as *cat* $\rightarrow$ *dog* and *cat* $\rightarrow$ *bird*. Taken together, the unified SOIV matrix provides a structural map of both class-wise instability and inter-class vulnerability directions. Additional experimental results corresponding to Sections 4.1–4.3 are provided in Appendix Section C.

Overall, this unified representation demonstrates that adversarial vulnerability can be interpreted as an intrinsic property of the model's output probability structure. By jointly encoding SOIV-U and SOIV-T, the matrix serves as a multi-layered summary of structural vulnerability derived solely from clean validation data. The heatmap visualization allows practitioners to readily identify vulnerable classes and high-risk source-target pairs, providing insights for designing efficient adversarial attacks or developing class-aware defense strategies.

5 Discussion

5.1 Computational Efficiency of SOIV-based Evaluation

To confirm the computational efficiency of the proposed framework, we compare the total execution time of the attack-based evaluation and the SOIV-based analysis in the C&W untargeted setting using VGG-16 on CIFAR-10. The attack-based runtime includes adversarial example generation,

Table 3: Class-wise ASR under AutoAttack for the Baseline, AT-Uniform, and AT-SOIV models. AutoAttack is evaluated under the L_∞ norm with $\epsilon = 8/255$.

Class	SOIV-U	Baseline	AT-Uniform	AT-SOIV
cat	0.019349	100.00%	85.12%	74.22%
deer	0.012349	100.00%	90.82%	74.09%
dog	0.011844	100.00%	74.65%	65.61%
bird	0.011219	100.00%	63.10%	53.86%
horse	0.008957	100.00%	50.57%	54.43%
airplane	0.007544	99.89%	60.67%	58.45%
frog	0.006985	100.00%	63.75%	61.84%
truck	0.005784	100.00%	55.71%	49.67%
automobile	0.005586	100.00%	57.09%	61.09%
ship	0.003910	100.00%	44.23%	48.28%
Overall ASR	–	99.99%	61.43%	58.17%
Clean Accuracy	–	88.24%	82.43%	85.17%

model inference on adversarial samples, and ASR computation, while the SOIV-based runtime consists of clean inference and SOIV metric computation. Each experiment is repeated 10 times and cumulative execution time is reported. As a result, the attack-based evaluation took 13.03 minutes, whereas the SOIV-based analysis required only 0.08 minutes, achieving approximately 162× speedup without generating adversarial examples. Although this experiment is conducted on a relatively small scale, the computational gap between the two approaches is expected to scale with dataset size, number of classes, and attack complexity, making the efficiency advantage of the proposed framework increasingly significant in practical large-scale settings.

5.2 A Practical Application: SOIV-guided Adversarial Training

The proposed SOIV metrics can also provide practical guidance for downstream robustness improvement. As one such example, we apply SOIV-U to class-aware adversarial training. We compare three configurations: Baseline (trained only on clean data), AT-Uniform (standard adversarial training with uniform class allocation), and AT-SOIV (SOIV-U-guided adversarial training with class-wise allocation proportional to SOIV-U values). For both AT-Uniform and AT-SOIV, adversarial examples are generated online using PGD during training with $\epsilon = 8/255$, step size $2/255$, and 5 steps, using the same clean-to-adversarial ratio of 1:5 and the same overall training budget. The only difference is the class-wise allocation of adversarial samples, such that AT-SOIV assigns more training resources to classes with higher estimated vulnerability.

As shown in Table 3, we evaluate the resulting models using standard AutoAttack under the L_∞ norm with $\epsilon = 8/255$. AT-SOIV achieves lower ASR than AT-Uniform in 7 of the 10 classes. In particular, for the four classes with the highest SOIV-U values—*cat*, *deer*, *dog*, and *bird*—ASR decreases by 10.90, 16.73, 9.04, and 9.24 percentage points, respectively. The overall ASR also decreases from 61.43% to 58.17%, while clean accuracy increases from 82.43% to 85.17%. Although the improvement is not consistent across all classes, these results suggest that SOIV-U can serve as a practical signal for prioritizing adversarial-training resources toward vulnerable classes. Additional evaluation under a different attack setting is provided in Appendix D.1.

6 Conclusion

Adversarial robustness evaluation has traditionally relied on attack-based methods, which are computationally expensive and sensitive to specific attack configurations. In this work, we proposed a second-order statistical framework that characterizes adversarial vulnerability through conditional variance and covariance computed solely from clean validation data, without requiring adversarial example generation. The proposed SOIV-U and SOIV-T quantify class-wise instability and directional misclassification tendencies, and exhibit strong alignment with class-wise ASR across diverse attack settings and model architectures. The resulting attack-agnostic framework is computationally efficient and can help inform the design of adversarial defense strategies. Overall, the proposed approach provides a practical pre-screening tool for adversarial robustness assessment. Future work may extend SOIV beyond closed-set classification settings with a predefined label space to open-set classifiers, such as vision-language models (VLMs).

AI USE STATEMENT

In this work, we used generative AI tools to assist with identifying and correcting errors in experimental code, as well as to improve grammar and wording during English manuscript preparation. All AI-assisted code modifications and manuscript revisions were reviewed and verified by the authors. We take full responsibility for the final content of this work, including the code, experimental results, and manuscript.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or sensitive information. The experiments are conducted using standard publicly available benchmark datasets and machine learning models. SOIV has a dual-use aspect, since the same information can guide defensive diagnosis or help an adversary prioritize targets. It generates no adversarial examples and no attack procedure, and it is intended for robustness auditing in access-controlled settings.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the datasets and data splits, model architectures and training configurations, adversarial attack settings, and the computation of the proposed SOIV metrics. Additional experimental settings and implementation details are provided in the Appendix B. These details are intended to enable reproduction of the main experimental results reported in this work.

REFERENCES

Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, pp. 301–316. Springer, 2022a.

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

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.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

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.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019.

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. *arXiv preprint arXiv:2010.09670*, 2020.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Ruize Gao, Jiongxiao Wang, Kaiwen Zhou, Feng Liu, Binghui Xie, Gang Niu, Bo Han, and James Cheng. Fast and reliable evaluation of adversarial robustness with minimum-margin attack. In *International Conference on Machine Learning*, pp. 7144–7163. PMLR, 2022.

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

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pp. 2137–2146. PMLR, 2018.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017a.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: A geometric perspective. *arXiv preprint arXiv:1705.09554*, 2017b.

Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M. Molloy, and Ben Edwards. Adversarial robustness toolbox v1.0.0. *arXiv preprint arXiv:1807.01069*, 2018.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pp. 8093–8104. PMLR, 2020.

Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4322–4330, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Ali Shafahi, Amin Ghiasi, Furong Huang, and Tom Goldstein. Label smoothing and logit squeezing: A replacement for adversarial training? *arXiv preprint arXiv:1910.11585*, 2019.

Sanidhya Sharma and Zesheng Chen. A systematic study of adversarial attacks against network intrusion detection systems. *Electronics*, 13(24):5030, 2024.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

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.

Wenqi Wei, Ling Liu, Margaret Loper, Stacey Truex, Lei Yu, Mehmet Emre Gursoy, and Yanzhao Wu. Adversarial examples in deep learning: Characterization and divergence. *arXiv preprint arXiv:1807.00051*, 2018.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

Muchao Ye, Xiang Xu, Qin Zhang, and Jonathan Wu. Sharpness-aware optimization for real-world adversarial attacks for diverse compute platforms with enhanced transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2937–2946, 2024.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

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.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. On the duality between sharpness-aware minimization and adversarial training. *arXiv preprint arXiv:2402.15152*, 2024.

Rui Zheng, Shihan Dou, Yuhao Zhou, Qin Liu, Tao Gui, Qi Zhang, Zhongyu Wei, Xuan-Jing Huang, and Menghan Zhang. Detecting adversarial samples through sharpness of loss landscape. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11282–11298, 2023.

Pascal Zimmer and Ghassan Karame. Tuning for two adversaries: Enhancing the robustness against transfer and query-based attacks using hyperparameter tuning. *arXiv preprint arXiv:2511.13654*, 2025.

APPENDIX

A THEORETICAL INTERPRETATION OF SOIV

A.1 THEORETICAL INTERPRETATION OF SOIV-U

We provide a theoretical interpretation of why the clean class-probability variance underlying SOIV-U can reflect local adversarial sensitivity. The key observation is that the class-wise variance of the true-class probability can be expressed exactly in terms of pairwise probability differences between independent clean samples from the same class. This representation allows both natural clean-sample variation and adversarial perturbations to be analyzed around the same clean sample X, and therefore through the same sample-wise local probability gradient $\nabla f_c(X)$.

Pairwise Representation of Clean Probability Variance For class c, let

$$f_c(\mathbf{x}) = p_c(\mathbf{x}) \tag{17}$$

denote the true-class probability produced by the classifier, and let

$$X, X' \overset{\text{i.i.d.}}{\sim} P(X \mid Y = c) \tag{18}$$

be two independent clean samples from the same class.

The clean class-probability variance underlying SOIV-U is

$$V_{\text{clean},c} = \operatorname{Var}_{X|Y=c}\left[f_c(X)\right]. \tag{19}$$

For two independent and identically distributed random variables, the variance admits the exact pairwise representation

$$V_{\text{clean},c} = \frac{1}{2}\mathbb{E}_{X,X'|Y=c}\left[\left(f_c(X') - f_c(X)\right)^2\right]. \tag{20}$$

To see this, let $Z = f_c(X)$ and $Z' = f_c(X')$. Since Z and Z' are independent and identically distributed,

$$\begin{aligned}\mathbb{E}\left[(Z - Z')^2\right] &= \mathbb{E}[Z^2] + \mathbb{E}[{Z'}^2] - 2\mathbb{E}[ZZ'] \\ &= 2\mathbb{E}[Z^2] - 2\mathbb{E}[Z]^2 \\ &= 2\operatorname{Var}[Z].\end{aligned} \tag{21}$$

Thus, the pairwise expression does not introduce a new statistic; it is an exact reformulation of the clean class-probability variance. Its advantage is that it expresses the variance in terms of probability changes between two clean samples, which can be analyzed through the local probability gradient.

Clean Probability Variability and Local Sensitivity Define the natural within-class displacement from X to X' as

$$D_{X,X'} = X' - X. \tag{22}$$

We assume that f_c is β-smooth over a region containing the relevant same-class samples, the line segments connecting them, and the adversarial neighborhoods considered below, i.e.,

$$\left\|\nabla f_c(x) - \nabla f_c(x')\right\|_2 \le \beta \left\|x - x'\right\|_2. \tag{23}$$

A first-order Taylor expansion of $f_c(X')$ around the actual clean sample X gives

$$f_c(X') - f_c(X) = \nabla f_c(X)^\top D_{X,X'} + r_{\text{clean}}(X, X'), \tag{24}$$

where the smoothness assumption gives

$$|r_{\text{clean}}(X, X')| \le \frac{\beta}{2}\left\|D_{X,X'}\right\|_2^2. \tag{25}$$

Neglecting the higher-order remainder momentarily, the first-order approximation of the clean probability variance becomes

$$V_{\mathrm{clean},c}^{\mathrm{lin}} = \frac{1}{2}\mathbb{E}_{X,X'|Y=c}\left[\left(\nabla f_c(X)^\top D_{X,X'}\right)^2\right]. \tag{26}$$

Importantly, the gradient in this expression is evaluated at the actual clean sample X, rather than at a class-average reference point.

For a fixed clean sample X, define the conditional second moment of natural within-class displacements as

$$M_{\mathrm{nat},c}(X) = \mathbb{E}_{X'|Y=c}\left[D_{X,X'}D_{X,X'}^\top \mid X\right]. \tag{27}$$

Then,

$$V_{\mathrm{clean},c}^{\mathrm{lin}} = \frac{1}{2}\mathbb{E}_{X|Y=c}\left[\nabla f_c(X)^\top M_{\mathrm{nat},c}(X)\nabla f_c(X)\right]. \tag{28}$$

Let

$$u_c(X) = \frac{\nabla f_c(X)}{\|\nabla f_c(X)\|_2} \tag{29}$$

denote the local probability-sensitive direction whenever $\nabla f_c(X) \neq 0$, and define

$$\kappa_c(X) = u_c(X)^\top M_{\mathrm{nat},c}(X)u_c(X). \tag{30}$$

The quantity $\kappa_c(X)$ measures how strongly the natural within-class variations explore the local probability-sensitive direction at X. We can therefore write

$$V_{\mathrm{clean},c}^{\mathrm{lin}} = \frac{1}{2}\mathbb{E}_{X|Y=c}\left[\kappa_c(X)\|\nabla f_c(X)\|_2^2\right]. \tag{31}$$

This expression directly connects the clean class-probability variance to the sample-wise local probability-gradient magnitude. In particular, suppose that the natural within-class distribution has non-negligible variation along the locally sensitive direction such that

$$\kappa_c(X) \geq \underline{\kappa}_c > 0 \tag{32}$$

where $\underline{\kappa}_c > 0$ denotes a lower bound on $\kappa_c(X)$ over the relevant region of class c. It then follows that

$$V_{\mathrm{clean},c}^{\mathrm{lin}} \geq \frac{\underline{\kappa}_c}{2}\mathbb{E}_{X|Y=c}\left[\|\nabla f_c(X)\|_2^2\right]. \tag{33}$$

Defining the class-wise RMS local probability sensitivity as

$$G_c = \sqrt{\mathbb{E}_{X|Y=c}\left[\|\nabla f_c(X)\|_2^2\right]}, \tag{34}$$

we obtain

$$V_{\mathrm{clean},c}^{\mathrm{lin}} \geq \frac{\underline{\kappa}_c}{2}G_c^2. \tag{35}$$

Thus, when natural within-class variation sufficiently covers the locally probability-sensitive directions, larger local probability-gradient energy induces a larger lower bound on the clean class-probability variability in the first-order approximation.

Connection to Adversarial Probability Sensitivity We next consider an adversarial perturbation δ_X applied to the same clean sample X. For an ℓ_2-bounded perturbation satisfying

$$\|\delta_X\|_2 \leq \epsilon, \tag{36}$$

a first-order expansion gives

$$f_c(X+\delta_X) - f_c(X) = \nabla f_c(X)^\top \delta_X + r_{\mathrm{adv}}(X), \tag{37}$$

with

$$|r_{\mathrm{adv}}(X)| \leq \frac{\beta}{2}\epsilon^2. \tag{38}$$

Define the perturbation-induced probability instability as

$$V_{\mathrm{adv},c} = \mathbb{E}_{X|Y=c}\left[\left(f_c(X+\delta_X) - f_c(X)\right)^2\right]. \tag{39}$$

Using the L_2 triangle inequality and the Cauchy–Schwarz inequality,

$$\sqrt{V_{\mathrm{adv},c}} \le \epsilon\sqrt{\mathbb{E}_{X|Y=c}\left[\|\nabla f_c(X)\|_2^2\right]} + \frac{\beta}{2}\epsilon^2. \tag{40}$$

Using the definition of G_c,

$$\sqrt{V_{\mathrm{adv},c}} \le \epsilon G_c + \frac{\beta}{2}\epsilon^2. \tag{41}$$

Therefore,

$$G_c \ge \left[\frac{\sqrt{V_{\mathrm{adv},c}}}{\epsilon} - \frac{\beta\epsilon}{2}\right]_+, \tag{42}$$

where $[z]_+ = \max(z, 0)$.

This result shows that large perturbation-induced probability instability, as quantified by $V_{\mathrm{adv},c}$, requires sufficiently large local probability sensitivity, as quantified by G_c, up to the second-order smoothness correction. This local probability sensitivity is also directly related to the gradient exploited by gradient-based untargeted adversarial attacks. For the cross-entropy loss with respect to the true class c,

$$\ell(X, c) = -\log f_c(X), \tag{43}$$

and its input gradient is

$$\nabla_X \ell(X, c) = -\frac{1}{f_c(X)}\nabla f_c(X). \tag{44}$$

Therefore,

$$\|\nabla_X \ell(X, c)\|_2 = \frac{1}{f_c(X)}\|\nabla f_c(X)\|_2. \tag{45}$$

Thus, at each clean sample X, the true-class probability gradient and the untargeted cross-entropy loss gradient are oppositely directed and their magnitudes are proportional through the positive factor $1/f_c(X)$. Since gradient-based untargeted attacks locally seek directions that increase the classification loss, the local probability sensitivity quantified by $\|\nabla f_c(X)\|_2$ is directly related to the first-order sensitivity exploited by such attacks.

Connecting Clean and Adversarial Probability Sensitivity Combining

$$V^{\mathrm{lin}}_{\mathrm{clean},c} \ge \frac{\kappa_c}{2} G_c^2 \tag{46}$$

with

$$G_c \ge \left[\frac{\sqrt{V_{\mathrm{adv},c}}}{\epsilon} - \frac{\beta\epsilon}{2}\right]_+ \tag{47}$$

gives

$$V^{\mathrm{lin}}_{\mathrm{clean},c} \ge \frac{\kappa_c}{2}\left[\frac{\sqrt{V_{\mathrm{adv},c}}}{\epsilon} - \frac{\beta\epsilon}{2}\right]_+^2. \tag{48}$$

This bound provides a conditional theoretical connection between the linearized clean class-probability variability and perturbation-induced probability instability. Both quantities depend, to first order, on the same sample-wise local probability gradient $\nabla f_c(X)$, while the clean-side relationship additionally requires that natural within-class variation sufficiently covers the locally probability-sensitive directions.

Effect of Higher-Order Terms The preceding relationship is based on the first-order approximation of the clean probability variance. The effect of the Taylor remainder can also be bounded.

From

$$f_c(X') - f_c(X) = \nabla f_c(X)^\top D_{X,X'} + r_{\text{clean}}(X, X'), \tag{49}$$

the reverse triangle inequality in L_2 gives

$$\sqrt{2V_{\text{clean},c}} \geq \sqrt{\mathbb{E}_{X,X'|Y=c}\left[\left(\nabla f_c(X)^\top D_{X,X'}\right)^2\right]} - \sqrt{\mathbb{E}_{X,X'|Y=c}\left[r_{\text{clean}}(X, X')^2\right]}. \tag{50}$$

Define

$$q_c = \sqrt{\mathbb{E}_{X,X'|Y=c}\left[\|X' - X\|_2^4\right]}. \tag{51}$$

Since

$$|r_{\text{clean}}(X, X')| \leq \frac{\beta}{2}\|X' - X\|_2^2, \tag{52}$$

we have

$$\sqrt{\mathbb{E}\left[r_{\text{clean}}(X, X')^2\right]} \leq \frac{\beta}{2} q_c. \tag{53}$$

Furthermore, under the directional-coverage condition,

$$\sqrt{\mathbb{E}\left[\left(\nabla f_c(X)^\top D_{X,X'}\right)^2\right]} \geq \sqrt{\underline{\kappa}_c}\, G_c. \tag{54}$$

Consequently,

$$\sqrt{2V_{\text{clean},c}} \geq \left[\sqrt{\underline{\kappa}_c}\left(\frac{\sqrt{V_{\text{adv},c}}}{\epsilon} - \frac{\beta\epsilon}{2}\right)_+ - \frac{\beta}{2} q_c\right]_+. \tag{55}$$

The higher-order terms therefore weaken the relationship when the probability landscape has high local curvature or when same-class samples are widely separated in input space.

Extension to ℓ_∞-Bounded Perturbations The preceding derivation is expressed for an ℓ_2-bounded perturbation. Because our experiments also include ℓ_∞-bounded attacks, we additionally provide the corresponding extension while retaining the same local sensitivity quantity G_c defined in Eq. 34.

Let the input dimension be d and consider an adversarial perturbation satisfying

$$\|\delta_X\|_\infty \leq \epsilon. \tag{56}$$

By norm equivalence,

$$\|\delta_X\|_2 \leq \sqrt{d}\|\delta_X\|_\infty \leq \sqrt{d}\epsilon. \tag{57}$$

Furthermore, by Hölder's inequality,

$$\left|\nabla f_c(X)^\top \delta_X\right| \leq \|\nabla f_c(X)\|_1 \|\delta_X\|_\infty. \tag{58}$$

Using

$$\|\nabla f_c(X)\|_1 \leq \sqrt{d}\|\nabla f_c(X)\|_2, \tag{59}$$

we obtain

$$\left|\nabla f_c(X)^\top \delta_X\right| \leq \sqrt{d}\epsilon\|\nabla f_c(X)\|_2. \tag{60}$$

Under the same β-smoothness assumption used above, the Taylor remainder satisfies

$$|r_{\text{adv}}(X)| \leq \frac{\beta}{2}\|\delta_X\|_2^2 \leq \frac{\beta d}{2}\epsilon^2. \tag{61}$$

Therefore, the perturbation-induced probability instability satisfies

$$\sqrt{V_{\mathrm{adv},c}} \le \sqrt{d}\epsilon\sqrt{\mathbb{E}_{X|Y=c}\left[\|\nabla f_c(X)\|_2^2\right]} + \frac{\beta d}{2}\epsilon^2. \tag{62}$$

Using the definition of G_c in Eq. 34,

$$\sqrt{V_{\mathrm{adv},c}} \le \sqrt{d}\epsilon G_c + \frac{\beta d}{2}\epsilon^2, \tag{63}$$

and hence

$$G_c \ge \left[\frac{\sqrt{V_{\mathrm{adv},c}}}{\sqrt{d}\epsilon} - \frac{\beta\sqrt{d}}{2}\epsilon\right]_+. \tag{64}$$

Combining this result with Eq. 35 gives

$$V^{\mathrm{lin}}_{\mathrm{clean},c} \ge \frac{\kappa_c}{2}\left[\frac{\sqrt{V_{\mathrm{adv},c}}}{\sqrt{d}\epsilon} - \frac{\beta\sqrt{d}}{2}\epsilon\right]_+^2. \tag{65}$$

Thus, the same qualitative connection between clean class-probability variability and adversarial probability sensitivity also holds for ℓ_∞-bounded perturbations. The additional dimension-dependent factor results from relating the ℓ_∞ perturbation constraint to the ℓ_2 local-gradient formulation used in the preceding analysis.

Connection to SOIV-U The proposed SOIV-U metric adopts an ideal-centered second-order formulation,

$$\text{SOIV-U}_c = \mathbb{E}_{X|Y=c}\left[(f_c(X)-1)^2\right]. \tag{66}$$

By the bias–variance decomposition,

$$\text{SOIV-U}_c = V_{\mathrm{clean},c} + \left(\mathbb{E}_{X|Y=c}[f_c(X)] - 1\right)^2. \tag{67}$$

This decomposition provides two complementary interpretations of SOIV-U. As shown above, the variance term $V_{\mathrm{clean},c}$ is linked, under the stated local smoothness and directional-coverage conditions, to the sample-wise local probability-gradient energy, and therefore reflects local probability sensitivity. The second term measures the deviation of the mean true-class probability from the ideal confidence of one, capturing the deviation of the average true-class confidence from its ideal value.

Thus, SOIV-U combines a sensitivity-related component with an ideal-confidence deviation term. A large variance indicates stronger variation of the true-class probability across clean samples, while a large deviation term indicates that the average true-class confidence is farther from its ideal value.

Interpretation and Scope The analysis above provides a theoretical interpretation of SOIV-U in terms of local probability sensitivity. The class-probability variance underlying SOIV-U is exactly equivalent to the expected squared probability difference between two independent clean samples from the same class. Under a local first-order approximation, this probability difference is governed by the sample-wise local gradient $\nabla f_c(X)$.

The same local gradient also governs the first-order response to an adversarial perturbation. Therefore, clean cross-sample probability variability and adversarially induced probability changes can be viewed as two manifestations of the same underlying local probability geometry, evaluated along natural and adversarial input directions, respectively.

This interpretation is conditional rather than universal. In particular, the relationship can weaken when natural within-class variations have little projection onto locally probability-sensitive directions, when the probability landscape exhibits substantial curvature, or when the relevant geometric quantities vary strongly across classes.

Importantly, the analysis establishes a connection to local adversarial probability sensitivity rather than directly to attack success rate. Whether a perturbation-induced probability change results in misclassification additionally depends on the decision margin and competing-class probabilities. Accordingly, the relationship between SOIV-U and class-wise attack success rate is evaluated empirically in our experiments.

A.2 Theoretical Interpretation of SOIV-T

SOIV-T is computed exclusively from clean validation samples and is intended to characterize pair-specific directional structure without generating adversarial examples. To motivate its use as an attack-free indicator, we consider the corresponding statistic evaluated on adversarial outputs, denoted SOAV-T. SOIV-T and SOAV-T evaluate the same ideal-centered source–target statistic on clean and adversarial outputs, respectively. We show below how adversarial perturbation modifies this common statistic and clarify the role of SOIV-T as its clean-data counterpart.

Target-directed probability structure. For a source class s and a target class $t \neq s$, consider the sample-wise quantity

$$q_{s\to t}(\mathbf{x}) = (1 - p_s(\mathbf{x}))\, p_t(\mathbf{x}). \tag{68}$$

This quantity jointly reflects low source-class confidence and high target-class probability. Its value increases when the source-class probability decreases and/or the target-class probability increases. Accordingly, $q_{s\to t}$ provides a continuous measure of target-directed structure in the model output.

For an adversarially perturbed input

$$X^{\mathrm{adv}} = X + \delta_X, \tag{69}$$

the corresponding quantity is

$$q^{\mathrm{adv}}_{s\to t}(X) = \big(1 - p_s(X^{\mathrm{adv}})\big)\, p_t(X^{\mathrm{adv}}). \tag{70}$$

Thus, $q^{\mathrm{adv}}_{s\to t}(X)$ characterizes the source–target structure present in the adversarial output for a given sample.

SOAV-T as the adversarial-data counterpart. The ideal-centered source–target statistic evaluated over adversarial samples is

$$\text{SOAV-T}_{s\to t} = \mathbb{E}_{X|Y=s}\left[\big(p_s(X^{\mathrm{adv}}) - 1\big)\, p_t(X^{\mathrm{adv}})\right]. \tag{71}$$

Since

$$p_s(X^{\mathrm{adv}}) - 1 \leq 0, \qquad p_t(X^{\mathrm{adv}}) \geq 0, \tag{72}$$

SOAV-T is non-positive, and its magnitude satisfies

$$|\text{SOAV-T}_{s\to t}| = \mathbb{E}_{X|Y=s}\left[q^{\mathrm{adv}}_{s\to t}(X)\right]. \tag{73}$$

Hence, the magnitude of SOAV-T summarizes, over adversarial source-class samples, the extent to which low source-class confidence co-occurs with high probability assigned to a particular target class. It therefore provides a dataset-level summary of target-directed structure in the adversarial outputs.

SOIV-T as the clean-data counterpart. SOIV-T applies the same ideal-centered source–target statistic to clean samples:

$$\text{SOIV-T}_{s\to t} = \mathbb{E}_{X|Y=s}\left[(p_s(X) - 1)\, p_t(X)\right]. \tag{74}$$

Likewise,

$$|\text{SOIV-T}_{s\to t}| = \mathbb{E}_{X|Y=s}\left[q^{\mathrm{clean}}_{s\to t}(X)\right], \tag{75}$$

where

$$q^{\mathrm{clean}}_{s\to t}(X) = (1 - p_s(X))\, p_t(X). \tag{76}$$

Thus, SOIV-T and SOAV-T evaluate the same source–target directional statistic on clean and adversarial outputs, respectively. SOIV-T measures the baseline source–target structure present in the clean output distribution, whereas SOAV-T measures the corresponding structure after adversarial perturbation.

Effect of adversarial perturbation. To relate the clean and adversarial quantities explicitly, define the perturbation-induced probability changes as

$$\Delta_s(X) = p_s(X^{\mathrm{adv}}) - p_s(X), \qquad \Delta_t(X) = p_t(X^{\mathrm{adv}}) - p_t(X). \tag{77}$$

Then,

$$p_s(X^{\mathrm{adv}}) = p_s(X) + \Delta_s(X), \qquad p_t(X^{\mathrm{adv}}) = p_t(X) + \Delta_t(X). \tag{78}$$

Substituting these expressions into $q^{\mathrm{adv}}_{s\to t}(X)$ gives

$$\begin{aligned} q^{\mathrm{adv}}_{s\to t}(X) &= (1 - p_s(X) - \Delta_s(X))\,(p_t(X) + \Delta_t(X)) \\ &= q^{\mathrm{clean}}_{s\to t}(X) + (1 - p_s(X))\,\Delta_t(X) \\ &\quad - p_t(X)\Delta_s(X) - \Delta_s(X)\Delta_t(X). \end{aligned} \tag{79}$$

Define

$$\Gamma_{s\to t}(X) = (1 - p_s(X))\,\Delta_t(X) - p_t(X)\Delta_s(X) - \Delta_s(X)\Delta_t(X) \tag{80}$$

as the attack-induced directional change. Then,

$$q^{\mathrm{adv}}_{s\to t}(X) = q^{\mathrm{clean}}_{s\to t}(X) + \Gamma_{s\to t}(X). \tag{81}$$

For the characteristic source-to-target probability response,

$$\Delta_s(X) \le 0, \qquad \Delta_t(X) \ge 0, \tag{82}$$

all terms in $\Gamma_{s\to t}(X)$ are non-negative. Therefore,

$$\Gamma_{s\to t}(X) \ge 0 \tag{83}$$

and

$$q^{\mathrm{adv}}_{s\to t}(X) \ge q^{\mathrm{clean}}_{s\to t}(X). \tag{84}$$

Thus, when an adversarial perturbation decreases the source-class probability and increases the target-class probability, it strengthens the same source–target directional statistic that is already defined on the clean output.

Averaging the exact decomposition over source-class samples gives

$$|\text{SOAV-T}_{s\to t}| = |\text{SOIV-T}_{s\to t}| + \mathbb{E}_{X|Y=s}\left[\Gamma_{s\to t}(X)\right]. \tag{85}$$

Hence, the magnitude of SOAV-T can be interpreted as the clean-data SOIV-T baseline together with the average directional change induced by the attack.

SOIV-T as an attack-free proxy for SOAV-T. The relation above does not imply that SOIV-T and SOAV-T are numerically equal. In particular, the attack-induced term $\Gamma_{s\to t}(X)$ may vary across source–target pairs, model architectures, perturbation budgets, and attack methods. Therefore, the decomposition alone does not guarantee preservation of the relative ordering of source–target pairs from SOIV-T to SOAV-T.

Nevertheless, it provides a structural motivation for using SOIV-T as an attack-free proxy for SOAV-T. Both quantities evaluate the same source–target statistic, with SOIV-T measuring its clean baseline and SOAV-T measuring its adversarial counterpart after the addition of attack-induced directional changes.

Our hypothesis is that the relative source–target structure observed in clean outputs is sufficiently preserved under adversarial perturbation for SOIV-T to remain informative about the corresponding SOAV-T pattern. This is an empirical hypothesis rather than a guaranteed consequence of the decomposition, and is evaluated experimentally in Appendix Section D.2.

Interpretation and scope. The roles of the quantities above can be distinguished as follows. The sample-wise quantity $q_{s\to t}^{\mathrm{clean}}(X)$ measures target-directed structure in the clean output, while $q_{s\to t}^{\mathrm{adv}}(X)$ measures the same structure after adversarial perturbation. Their difference, $\Gamma_{s\to t}(X)$, captures the corresponding attack-induced directional change. At the dataset level, the magnitudes of SOIV-T and SOAV-T average the clean and adversarial versions of this statistic, respectively.

This interpretation establishes a direct structural relationship between SOIV-T and SOAV-T without assuming that they are identical. Under the characteristic source-to-target response, the adversarial perturbation increases the magnitude of the same directional statistic measured by SOIV-T. However, the amount of this increase can vary across source–target pairs and attack configurations. Accordingly, the use of SOIV-T as a proxy for SOAV-T relies on the empirical hypothesis that the relative directional structure present in clean outputs remains informative after adversarial perturbation. The consistency between SOIV-T, SOAV-T, and attack-based vulnerability patterns is evaluated experimentally in Appendix Section D.2.

B Experimental Details

B.1 Experimental Setup

We conduct experiments on multiple datasets and model architectures, including CIFAR-10 with ResNet-18, Swin-T, and VGG-16, FashionMNIST with ResNet-18, CIFAR-100 with an adversarially trained PreActResNet-18 (Addepalli et al., 2022a) and ResNet-18 (Addepalli et al., 2022b) obtained from RobustBench (Croce et al., 2020), and ImageNet-100 with ViT-L/16 pretrained on ImageNet-1K. No training is performed in this work for the CIFAR-100 and ImageNet-100 settings. Unless otherwise specified, CIFAR-10 with ResNet-18 serves as the primary experimental setting used for the main analysis. All experiments are implemented in PyTorch and performed on a single NVIDIA RTX 5070 Ti GPU.

We first compute SOIV-U and SOIV-T from softmax outputs on the validation set. We then apply various untargeted and targeted attacks to measure class-wise and directional attack success rates (ASR) on the test set. Finally, we compare SOIV values with ASR values using Pearson, Spearman, and Kendall correlation coefficients to quantitatively assess the alignment between the proposed metrics and conventional attack-based vulnerability measures. Values in parentheses after correlation coefficients are two-sided p-values.

B.2 Datasets and Data Splits

For CIFAR-10, the dataset is split into 45,000 training samples, 5,000 validation samples, and 10,000 test samples. The validation set is used to compute the proposed statistical metrics, while adversarial attacks are evaluated on the disjoint test set. For FashionMNIST, we follow the same settings by splitting the original training set into 55,000 training samples and 5,000 validation samples, while using the official test set of 10,000 images for evaluation. For CIFAR-100, we split the official test set of 10,000 images into 5,000 validation samples and 5,000 test samples, without separate training samples since pretrained models are used. For ImageNet-100, a 100-class subset of ImageNet-1K, we split the official validation set of 5,000 images into 2,500 validation samples and 2,500 test samples. No separate training samples are used since pretrained models are used. The official ImageNet test set is not used because its labels are not publicly available.

B.3 Training Configuration

For the primary setting (CIFAR-10 + ResNet-18), the model is trained from scratch using stochastic gradient descent (SGD) with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is set to 0.1 and decayed by a factor of 10 (gamma = 0.1) at epochs 100 and 150. Training is performed for 300 epochs with a batch size of 128. Standard data augmentation including random cropping and horizontal flipping is applied. Other dataset and model combinations use different training hyperparameters depending on the architecture and dataset characteristics. The detailed training configurations for each dataset and model combination are summarized in Table 4. The normalization statistics used for each dataset are shown in Table 5. All experiments are conducted on a single NVIDIA RTX 5070 Ti GPU.

Table 4: Training configurations used for different dataset and model combinations. The CIFAR-100 and ImageNet-100 settings are omitted since pretrained checkpoints are used.

Dataset	Model	Optimizer	LR	Epochs	Batch	Weight-Decay	Augmentation
CIFAR-10	ResNet-18	SGD	0.1	300	128	5×10^{-4}	crop + flip
CIFAR-10	Swin-T	AdamW	3×10^{-4}	200	128	0.05	flip
CIFAR-10	VGG-16	Adam	1×10^{-3}	200	128	0	crop + flip + rotate
Fashion-MNIST	ResNet-18	SGD	0.05	100	128	1×10^{-4}	–

Table 5: Normalization statistics used for each dataset. No external normalization is applied to the CIFAR-100 models, which take inputs in $[0, 1]$ as provided by RobustBench.

Dataset	Mean	Std
CIFAR-10	(0.4914, 0.4822, 0.4465)	(0.2023, 0.1994, 0.2010)
Fashion-MNIST	0.1307	0.3081
CIFAR-100	–	–
ImageNet-100	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)

B.4 Computation of Second-Order Statistics

To compute the proposed statistics, we collect softmax outputs for all validation samples using the trained model without any perturbation. For each input $\mathbf{x}$ with ground-truth label c, we record the predicted probability of the correct class $p_c(\mathbf{x})$ and use these probabilities to compute the proposed second-order statistics, including SOIV-U and SOIV-T. To evaluate adversarial vulnerability, we then apply various untargeted and targeted attacks on the test set and measure class-wise and directional attack success rates (ASR). Finally, we compute Pearson, Spearman, and Kendall correlation coefficients between the proposed SOIV metrics and the observed ASR values to quantify the alignment between the proposed statistics and conventional attack-based vulnerability measures.

B.5 Attack Configurations

To evaluate adversarial robustness, we use several commonly adopted adversarial attacks. For untargeted evaluation, we consider FGSM (Goodfellow et al., 2014; Kurakin et al., 2016; Tramèr et al., 2017; Narodytska & Kasiviswanathan, 2016), PGD (Madry et al., 2017; Croce & Hein, 2020; Schmidt et al., 2018; Rice et al., 2020), C&W (Carlini & Wagner, 2017; Croce & Hein, 2020), and Square attacks (Andriushchenko et al., 2020). In some settings, the untargeted evaluation additionally includes AutoAttack (Croce & Hein, 2020). For targeted evaluation, we use FGSM, PGD, and C&W attacks. In some settings, I-FGSM (Kurakin et al., 2018) replaces FGSM, and Square is additionally included. All attacks are applied only to clean-correct test samples. The attack strengths are chosen so that class-wise ASRs remain spread out. Here ASR is the reference signal used to validate SOIV, so what matters is the ordering and spread of ASR across classes rather than its absolute level. Appendix Section D.7 reports the supporting analysis. Since the attack hyperparameters vary across dataset and model combinations, we summarize the detailed configurations separately for each setting in Tables 6, 7, 8, 9, 10, 11, and 12.

C Additional Experimental Results

We present additional experimental results to demonstrate that the proposed framework generalizes across diverse dataset-model combinations. Specifically, for each setting we report ranking comparisons between the proposed SOIV metrics and adversarial attack outcomes. First, we present class-wise ranking comparisons between SOIV-U and the class-wise untargeted ASR, illustrating how the proposed vulnerability measure aligns with observed untargeted attack behavior. Second, we report directional ranking comparisons between SOIV-T and the targeted ASR, which capture directional misclassification tendencies between source and target classes. Finally, we visualize

Table 6: Attack configurations for **CIFAR-10 + ResNet-18** used in the experiments reported in Sections 4.1, 4.2 of the main paper.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 2/255$, norm=L_∞
PGD	$\epsilon = 1.5/255$, $\alpha = 1/255$, steps=6, random start=True, norm=L_∞
C&W	$c = 0.01$, $\kappa = 3$, steps=230, lr=7×10^{-3}, norm=L_2
Square	$\epsilon = 4/255$, queries=1000, restart=True, norm=L_∞, $p_{\text{init}} = 0.8$, loss=margin
AutoAttack	$\epsilon = 1.5/255$, norm=L_∞, version=standard, seed=0, verbose=True
Targeted(source=automobile)	
FGSM	$\epsilon = 6/255$, norm=L_∞
PGD	$\epsilon = 4/255$, $\alpha = 0.4/255$, steps=10, random start=True, norm=L_∞
C&W	$c = 0.1$, $\kappa = 1$, steps=10, lr=0.01, norm=L_2

Table 7: Attack configurations for **CIFAR-10 + Swin-T** used in the experiments reported in Appendix Section C.1.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 0.6/255$, norm=L_∞
PGD	$\epsilon = 0.05/255$, $\alpha = 0.5/255$, steps=3, random start=True, norm=L_∞
C&W	$c = 1 \times 10^{-8}$, $\kappa = 0$, steps=130, lr=5×10^{-3}, norm=L_2
Square	$\epsilon = 1.5/255$, queries=750, restart=True, norm=L_∞, $p_{\text{init}} = 0.8$, loss=margin
AutoAttack	$\epsilon = 0.5/255$, norm=L_∞, version=standard, seed=0, verbose=True
Targeted(source=horse)	
FGSM	$\epsilon = 0.35/255$, norm=L_∞
PGD	$\epsilon = 0.5/255$, $\alpha = 1/255$, steps=3, random start=True, norm=L_∞
C&W	$c = 0.1$, $\kappa = 0$, steps=10, lr=1×10^{-3}, norm=L_2

Table 8: Attack configurations for **FashionMNIST + ResNet-18** used in the experiments reported in Appendix Section C.3.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 3/255$, norm=L_∞
PGD	$\epsilon = 4/255$, $\alpha = 1/255$, steps=6, random start=True, norm=L_∞
C&W	$c = 1 \times 10^{-3}$, $\kappa = 3$, steps=80, lr=3×10^{-3}, norm=L_2
Square	$\epsilon = 4/255$, queries=1000, restart=True, norm=L_∞, $p_{\text{init}} = 0.8$, loss=margin
AutoAttack	$\epsilon = 4/255$, norm=L_∞, version=standard, seed=0, verbose=True
Targeted(source=8)	
FGSM	$\epsilon = 10/255$, norm=L_∞
PGD	$\epsilon = 0.5/255$, $\alpha = 1/255$, steps=3, random start=True, norm=L_∞
C&W	$c = 0.9$, $\kappa = 2$, steps=40, lr=0.01, norm=L_2

Table 9: Attack configurations for **CIFAR-10 + VGG-16** used in the experiments reported in Appendix Section C.2.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 2/255$, norm=L_∞
PGD	$\epsilon = 1.2/255$, $\alpha = 1/255$, steps=7, random start=True, norm=L_∞
C&W	$c = 2 \times 10^{-3}$, $\kappa = 0.05$, steps=200, lr=5×10^{-3}, norm=L_2
Square	$\epsilon = 2/255$, queries=500, restart=True, norm=L_∞, $p_{\text{init}} = 0.9$, loss=margin
Targeted(source=cat)	
FGSM	$\epsilon = 1/255$, norm=L_∞
PGD	$\epsilon = 1/255$, $\alpha = 1/255$, steps=10, random start=True, norm=L_∞
C&W	$c = 0.01$, $\kappa = 0.5$, steps=1500, lr=5×10^{-3}, norm=L_2
Square	$\epsilon = 2/255$, queries=1000, restart=True, norm=L_∞, $p_{\text{init}} = 0.5$, loss=margin

Table 10: Attack configurations for **CIFAR-100 + Adversarially Trained PreActResNet-18** used in the experiments reported in Appendix Section C.4.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 8/255$, norm=L_∞
PGD	$\epsilon = 8/255$, $\alpha = 2/255$, steps= 20, random start= True, norm=L_∞
C&W	$c = 0.5$, $\kappa = 0$, steps= 100, lr= 0.01, norm=L_2
Square	$\epsilon = 8/255$, queries= 200, restarts= True, norm= L_∞, $p_{\text{init}} = 0.8$, loss=margin
Targeted(source=0)	
I-FGSM	$\epsilon = 32/255$, $\alpha = 8/255$, steps= 10, norm=L_∞
PGD	$\epsilon = 32/255$, $\alpha = 8/255$, steps= 20, random start= True, norm=L_∞
C&W	$c = 10$, $\kappa = 0$, steps= 100, lr= 0.01, norm=L_2

Table 11: Attack configurations for **CIFAR-100 + Adversarially Trained ResNet-18** used in the experiments reported in Appendix Section C.5.

Attack	Parameters
Untargeted	
FGSM	$\epsilon = 8/255$, norm=L_∞
PGD	$\epsilon = 8/255$, $\alpha = 2/255$, steps= 20, random start= True, norm=L_∞
C&W	$c = 0.5$, $\kappa = 0$, steps= 100, lr= 0.01, norm=L_2
Square	$\epsilon = 8/255$, queries= 200, restarts= True, norm= L_∞, $p_{\text{init}} = 0.8$, loss=margin
Targeted(source=0)	
I-FGSM	$\epsilon = 32/255$, $\alpha = 8/255$, steps= 10, norm=L_∞
PGD	$\epsilon = 32/255$, $\alpha = 8/255$, steps= 20, random start= True, norm=L_∞
C&W	$c = 10$, $\kappa = 0$, steps= 100, lr= 0.01, norm=L_2

Table 12: Attack configurations for **ImageNet-100 + ViT-L/16** used in the experiments reported in Appendix Section C.6.

Attack	Parameters
Untargeted	
FGSM	$\epsilon \in \{0.5, 1, 2\}/255$, norm=$L_\infty$
PGD	$\epsilon \in \{0.3, 0.5\}/255$, $\alpha = 0.5/255$, steps= 10, random start= True, norm=L_∞
C&W	$c \in \{0.01, 0.05\}$, $\kappa = 0.05$, steps= 20, lr= 5×10^{-3}, norm=L_2
Targeted(source=n03642806)	
I-FGSM	$\epsilon \in \{1.25, 1.5, 1.75, 2, 2.5, 3\}/255$, $\alpha = \epsilon/5$, steps= 5, norm=L_∞
PGD	$\epsilon \in \{0.75, 1, 1.2, 1.35, 1.5\}/255$, $\alpha = \epsilon/4$, steps= 10, random start= True, norm=L_∞
C&W	$c \in \{0.5, 1, 2, 5\}$, $\kappa = 0$, steps= 10, lr= 5×10^{-3}, norm=L_2

Table 13: Comparison between SOIV-U ranking and class-wise ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED) for **CIFAR-10 + Swin-T**. The second column lists classes sorted in descending order of SOIV-U values, formatted as *class name* (*SOIV-U value*). Each subsequent column reports the ASR under the corresponding untargeted attack, sorted in descending order of ASR and formatted as *class name* (*ASR value*). The bottom rows provide the Pearson, Spearman, and Kendall correlation coefficients between the SOIV-U ranking and the ASR ranking for each attack.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	AUTOATTACK	MIXED
1	cat (4.92e-2)	cat (43.54)	cat (40.51)	cat (70.33)	cat (89.00)	deer (100.00)	cat (68.68)
2	dog (3.80e-2)	dog (29.72)	dog (26.94)	dog (55.00)	dog (77.78)	cat (100.00)	dog (57.87)
3	bird (3.29e-2)	bird (27.17)	bird (25.27)	bird (54.35)	bird (75.27)	frog (99.90)	bird (56.35)
4	deer (2.56e-2)	deer (26.99)	deer (24.66)	deer (50.80)	deer (74.23)	dog (99.90)	deer (55.34)
5	airplane (2.00e-2)	horse (17.96)	horse (15.41)	airplane (41.94)	frog (64.77)	bird (99.70)	frog (46.88)
6	horse (1.75e-2)	airplane (16.47)	airplane (14.93)	frog (39.51)	horse (61.04)	airplane (99.60)	horse (46.18)
7	frog (1.48e-2)	frog (15.99)	frog (14.25)	horse (37.99)	truck (56.00)	ship (99.20)	airplane (45.63)
8	automobile (9.93e-3)	truck (13.28)	truck (11.89)	truck (30.72)	airplane (55.21)	truck (98.80)	truck (42.14)
9	ship (9.82e-3)	automobile (11.58)	automobile (10.13)	ship (27.00)	automobile (45.55)	horse (98.50)	ship (37.92)
10	truck (9.74e-3)	ship (11.56)	ship (9.84)	automobile (23.16)	ship (41.99)	automobile (98.30)	automobile (37.74)
Pearson		0.978 (1.0e-06)	0.977 (1.2e-06)	0.970 (3.4e-06)	0.919 (1.7e-04)	0.673 (3.3e-02)	0.971 (3.0e-06)
Spearman		0.952 (2.2e-05)	0.952 (2.2e-05)	0.939 (5.6e-05)	0.855 (1.6e-03)	0.681 (3.0e-02)	0.903 (3.4e-04)
Kendall		0.867 (1.2e-04)	0.867 (1.2e-04)	0.822 (3.6e-04)	0.733 (2.2e-03)	0.539 (3.0e-02)	0.733 (2.2e-03)

the unified SOIV variance–covariance matrix that integrates SOIV-U and SOIV-T within a single structure. In this matrix, the diagonal entries correspond to SOIV-U values representing class-wise vulnerability under untargeted settings, while the off-diagonal entries correspond to SOIV-T values capturing directional vulnerability between source–target class pairs.

C.1 CIFAR-10 + Swin-T

Table 13 compares the SOIV-U ranking with class-wise ASR rankings under multiple untargeted attacks. As a representative example, we focus on the MIXED setting, which aggregates multiple untargeted attacks to represent a combined threat environment rather than relying on a single attack type. Under this setting, vulnerable classes such as *cat*, *dog*, and *bird* appear consistently near the top of both the SOIV-U and ASR rankings, while classes such as *automobile* and *ship* remain relatively lower. Quantitatively, the SOIV-U ranking shows strong alignment with the MIXED ASR ranking (Pearson 0.971, Spearman 0.903, Kendall 0.733).

Table 14 presents the comparison between SOIV-T ranking and targeted ASR when the source class is *horse*. The SOIV-T ranking indicates that *dog* and *deer* have the largest-magnitude covariance values, suggesting stronger directional vulnerability toward these target classes. This trend is reflected in the targeted attack results, where classes with larger-magnitude SOIV-T values tend to exhibit higher ASR. Quantitatively, the correlation between SOIV-T ranking and ASR remains high across attacks.

Fig. 3 visualizes the unified SOIV variance–covariance matrix integrating SOIV-U and SOIV-T. The diagonal entries represent class-wise vulnerability measured by SOIV-U, where *cat* shows the largest variance value, indicating the highest vulnerability among the classes. The off-diagonal entries capture directional relationships between class pairs; for example, when the source class is *cat*, the covariance toward *dog* shows one of the largest magnitudes, indicating that *cat* samples tend to be misclassified as *dog* more frequently, thus exhibiting higher directional vulnerability. Such information can be utilized in designing effective adversarial attacks and defense strategies.

C.2 CIFAR-10 + VGG-16

For CIFAR-10 + VGG-16, we present ranking comparisons between the proposed SOIV metrics and the corresponding ASR under both untargeted and targeted settings. In addition, we visualize the unified SOIV variance–covariance matrix that integrates SOIV-U and SOIV-T within a single structure. The corresponding results are shown in Table 15, Table 16, and Fig. 4. As in the previous results, strong correlations are observed between SOIV and ASR, indicating that the proposed metric effectively reflects class-wise adversarial vulnerability patterns.

Table 14: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (FGSM, PGD, C&W, and MIXED) when the source class is *horse* for **CIFAR-10 + Swin-T**. The second column lists target classes sorted in ascending order of SOIV-T values, formatted as *class name* (*SOIV-T value*). Each subsequent column reports the ASR under the corresponding targeted attack, formatted as *class name* (*ASR value*). The bottom rows provide the Pearson, Spearman, and Kendall correlation coefficients between the SOIV-T ranking and each ASR.

Rank	SOIV-T	FGSM	PGD	C&W	MIXED
1	dog (-0.005553)	deer (57.16)	deer (71.12)	deer (97.21)	deer (75.16)
2	deer (-0.004520)	dog (40.53)	dog (50.73)	dog (90.90)	dog (60.72)
3	cat (-0.002653)	bird (28.88)	bird (37.62)	bird (90.29)	bird (52.27)
4	bird (-0.001712)	cat (23.54)	cat (30.34)	truck (87.14)	truck (46.00)
5	airplane (-0.001314)	truck (21.36)	truck (29.49)	cat (82.65)	cat (45.51)
6	truck (-0.001222)	airplane (16.75)	airplane (19.66)	automobile (80.34)	airplane (37.78)
7	automobile (-0.000270)	automobile (10.44)	frog (17.84)	airplane (76.94)	automobile (34.99)
8	ship (-0.000176)	frog (8.50)	automobile (14.20)	frog (76.34)	frog (34.22)
9	frog (-0.000073)	ship (5.34)	ship (7.52)	ship (61.89)	ship (24.92)
Pearson		0.899882 (9.5e-04)	0.882122 (1.6e-03)	0.726233 (2.7e-02)	0.873903 (2.1e-03)
Spearman		0.933333 (2.4e-04)	0.900000 (9.4e-04)	0.850000 (3.7e-03)	0.883333 (1.6e-03)
Kendall		0.777778 (2.4e-03)	0.722222 (5.9e-03)	0.666667 (1.3e-02)	0.722222 (5.9e-03)

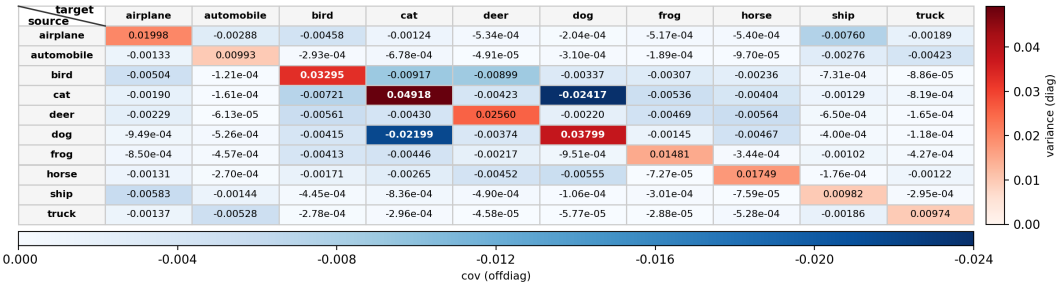


Figure 3: Unified SOIV matrix for **CIFAR-10 + Swin-T** computed from clean validation data. The diagonal entries represent SOIV-U values corresponding to class-wise vulnerability under untargeted settings (higher values indicate greater vulnerability), while the off-diagonal entries represent SOIV-T values capturing directional vulnerability between source–target class pairs (larger magnitude indicates stronger directional misclassification tendency). Color intensity reflects the magnitude of each entry, enabling intuitive identification of vulnerable classes and strong directional relationships between class pairs.

Table 15: Comparison between SOIV-U ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) for **CIFAR-10 + VGG-16**.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	MIXED
1	cat (2.12e-1)	cat (76.22)	cat (80.85)	cat (61.16)	cat (67.64)	cat (71.47)
2	dog (1.65e-1)	dog (57.29)	dog (62.56)	dog (57.41)	dog (47.86)	dog (56.28)
3	bird (1.06e-1)	deer (59.07)	deer (63.61)	bird (56.45)	deer (46.03)	deer (56.26)
4	airplane (8.09e-2)	bird (53.66)	bird (57.84)	deer (56.35)	bird (44.72)	bird (53.16)
5	frog (7.00e-2)	frog (42.87)	frog (47.37)	frog (51.02)	frog (29.47)	frog (42.68)
6	deer (6.78e-2)	airplane (41.28)	airplane (43.93)	horse (30.50)	airplane (29.58)	airplane (35.98)
7	horse (5.20e-2)	horse (37.81)	horse (37.49)	ship (28.66)	horse (25.24)	horse (32.76)
8	ship (4.72e-2)	truck (33.48)	ship (34.91)	airplane (29.14)	ship (21.01)	ship (28.93)
9	truck (4.07e-2)	ship (31.14)	truck (31.75)	truck (17.49)	truck (19.22)	truck (25.49)
10	automobile (2.09e-2)	automobile (26.11)	automobile (23.43)	automobile (16.92)	automobile (13.11)	automobile (19.89)
Pearson		0.894 (4.9e-04)	0.891 (5.4e-04)	0.766 (9.8e-03)	0.918 (1.8e-04)	0.883 (7.1e-04)
Spearman		0.855 (1.6e-03)	0.867 (1.2e-03)	0.915 (2.1e-04)	0.927 (1.1e-04)	0.915 (2.0e-04)
Kendall		0.733 (2.2e-03)	0.778 (9.5e-04)	0.822 (3.6e-04)	0.867 (1.2e-04)	0.822 (3.6e-04)

C.3 FashionMNIST + ResNet-18

For FashionMNIST + ResNet-18, we report ranking comparisons between the proposed SOIV metrics and the corresponding ASR under both untargeted and targeted attack settings, together with the unified SOIV matrix. The results are summarized in Table 17, Table 18, and Fig. 5. Despite the

Table 16: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) when the source class is *cat* for **CIFAR-10 + VGG-16**.

Rank	SOIV-T	FGSM	PGD	C&W	SQUARE	MIXED
1	dog (-0.081034)	dog (29.33)	dog (46.63)	dog (63.01)	dog (58.26)	dog (49.31)
2	deer (-0.030739)	frog (14.53)	bird (26.55)	frog (42.80)	frog (36.59)	frog (30.02)
3	frog (-0.023766)	bird (13.47)	frog (26.16)	horse (35.40)	bird (35.54)	bird (27.54)
4	airplane (-0.018932)	deer (11.76)	deer (24.70)	bird (34.61)	deer (33.69)	deer (25.92)
5	bird (-0.018262)	horse (11.10)	horse (22.99)	deer (33.55)	horse (31.57)	horse (25.26)
6	horse (-0.016610)	truck (5.42)	truck (13.08)	truck (31.57)	truck (19.55)	truck (17.40)
7	truck (-0.013484)	airplane (4.36)	airplane (10.57)	airplane (19.29)	airplane (16.91)	airplane (12.78)
8	automobile (-0.004909)	ship (3.70)	ship (8.98)	automobile (17.70)	ship (16.25)	ship (11.36)
9	ship (-0.004317)	automobile (3.30)	automobile (8.32)	ship (16.51)	automobile (13.08)	automobile (10.60)
Pearson		0.9307 (2.7e-04)	0.9036 (8.3e-04)	0.8852 (1.5e-03)	0.8917 (1.2e-03)	0.9071 (7.3e-04)
Spearman		0.8167 (7.2e-03)	0.7833 (1.3e-02)	0.7500 (2.0e-02)	0.8167 (7.2e-03)	0.8167 (7.2e-03)
Kendall		0.6667 (1.3e-02)	0.6111 (2.5e-02)	0.6111 (2.5e-02)	0.6667 (1.3e-02)	0.6667 (1.3e-02)

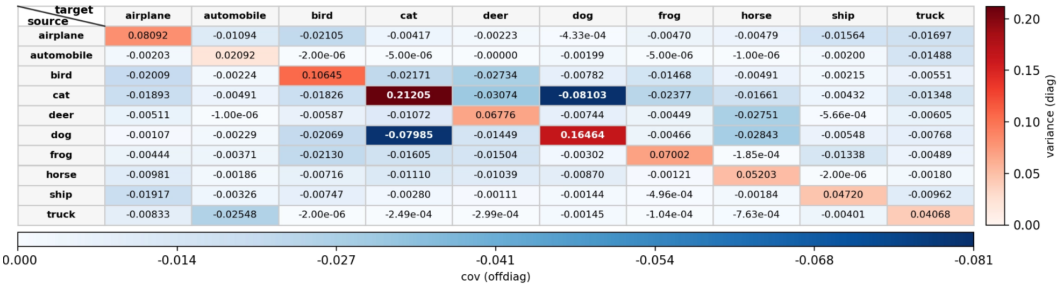


Figure 4: Unified SOIV matrix for **CIFAR-10 + VGG-16** computed from clean validation data.

Table 17: Comparison between SOIV-U ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED) for **FashionMNIST + ResNet-18**. The class indices for FashionMNIST correspond to the following labels: 0: T-shirt/top, 1: Trouser, 2: Pullover, 3: Dress, 4: Coat, 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, and 9: Ankle boot.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	AUTOATTACK	MIXED
1	6 (1.11e-2)	6 (82.6)	2 (98.6)	2 (59.1)	2 (96.5)	2 (99.9)	2 (87.1)
2	0 (5.09e-3)	2 (81.5)	6 (97.5)	6 (46.0)	6 (91.0)	6 (99.6)	6 (83.3)
3	2 (4.79e-3)	4 (75.7)	4 (96.1)	4 (45.1)	4 (88.3)	4 (99.4)	4 (80.9)
4	3 (4.39e-3)	0 (65.9)	0 (94.7)	0 (39.3)	0 (87.2)	0 (98.6)	0 (77.2)
5	4 (4.28e-3)	3 (31.9)	3 (70.7)	3 (11.9)	3 (51.3)	8 (92.2)	3 (50.5)
6	9 (1.41e-3)	7 (31.6)	7 (69.3)	7 (9.1)	7 (49.4)	9 (88.9)	7 (49.4)
7	7 (1.38e-3)	9 (20.8)	9 (66.4)	9 (6.4)	8 (47.5)	7 (87.7)	9 (45.2)
8	8 (1.00e-3)	8 (14.7)	8 (62.8)	8 (4.6)	9 (43.6)	3 (86.6)	8 (44.4)
9	5 (6.74e-4)	5 (13.0)	5 (33.6)	1 (1.3)	5 (20.6)	5 (45.9)	5 (22.7)
10	1 (4.40e-4)	1 (6.5)	1 (24.3)	5 (0.5)	1 (10.5)	1 (40.9)	1 (16.7)
Pearson		0.823 (3.4e-03)	0.740 (1.4e-02)	0.753 (1.2e-02)	0.767 (9.6e-03)	0.583 (7.7e-02)	0.774 (8.6e-03)
Spearman		0.927 (1.1e-04)	0.903 (3.4e-04)	0.891 (5.4e-04)	0.879 (8.1e-04)	0.770 (9.2e-03)	0.903 (3.4e-04)
Kendall		0.822 (3.6e-04)	0.778 (9.5e-04)	0.733 (2.2e-03)	0.733 (2.2e-03)	0.600 (1.7e-02)	0.778 (9.5e-04)

change in dataset, SOIV still exhibits positive correlations with ASR, indicating that the proposed metric captures class-wise adversarial vulnerability patterns across different data distributions.

Table 18: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (FGSM, PGD, C&W, and MIXED) when the source class is *8* for **FashionMNIST + ResNet-18**. The class indices for FashionMNIST correspond to the following labels: 0: T-shirt/top, 1: Trouser, 2: Pullover, 3: Dress, 4: Coat, 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, and 9: Ankle boot.

Rank	SOIV-T	FGSM	PGD	C&W	MIXED
1	0 (-0.000362)	6 (84.26)	0 (0.40)	2 (92.94)	6 (55.67)
2	6 (-0.000349)	4 (61.25)	6 (0.40)	4 (85.07)	4 (48.84)
3	4 (-0.000186)	0 (46.42)	2 (0.20)	6 (82.34)	2 (44.03)
4	2 (-0.000034)	3 (45.81)	4 (0.20)	0 (80.12)	0 (42.31)
5	5 (-0.000028)	2 (38.95)	7 (0.20)	3 (79.21)	3 (41.71)
6	3 (-0.000016)	7 (13.72)	3 (0.10)	1 (78.41)	1 (29.67)
7	1 (-0.000014)	5 (11.30)	5 (0.10)	5 (59.84)	5 (23.75)
8	7 (-0.000014)	1 (10.60)	1 (0.00)	9 (53.28)	7 (21.56)
9	9 (-0.000002)	9 (6.26)	9 (0.00)	7 (50.76)	9 (19.85)
Pearson		0.769772 (1.5e-02)	0.888221 (1.4e-03)	0.428324 (2.5e-01)	0.701621 (3.5e-02)
Spearman		0.833333 (5.3e-03)	0.807231 (8.5e-03)	0.766667 (1.6e-02)	0.850000 (3.7e-03)
Kendall		0.666667 (1.3e-02)	0.730297 (6.1e-03)	0.500000 (7.5e-02)	0.722222 (5.9e-03)

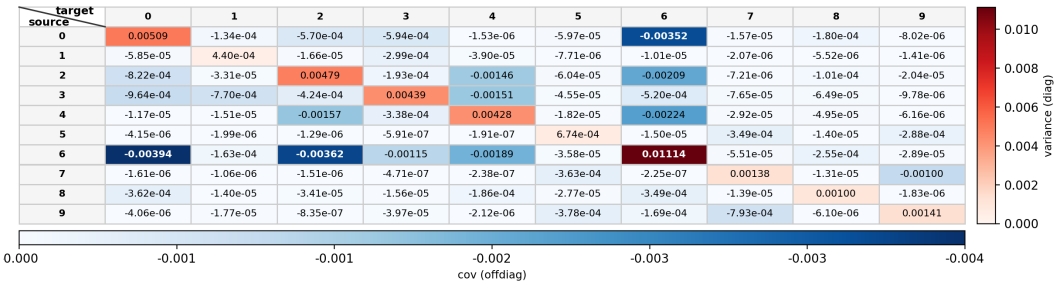


Figure 5: Unified SOIV matrix for **FashionMNIST + ResNet-18** computed from clean validation data. The class indices for FashionMNIST correspond to the following labels: 0: T-shirt/top, 1: Trouser, 2: Pullover, 3: Dress, 4: Coat, 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, and 9: Ankle boot.

C.4 CIFAR-100 + Adversarially Trained PreActResNet-18

For CIFAR-100 + Adversarially Trained PreActResNet-18, we present ranking comparisons between the proposed SOIV metrics and the corresponding ASR under both untargeted and targeted settings. The model is obtained from RobustBench (Croce et al., 2020; Addepalli et al., 2022a). The corresponding results are provided in Table 19 and Table 20. Since FGSM yields negligible attack success in the targeted setting, we replace it with Iterative FGSM (I-FGSM) for the targeted experiments. Even in the CIFAR-100 setting with a substantially larger number of classes, the rank correlations remain high under both untargeted and targeted settings. This indicates that SOIV-U and SOIV-T remain effective in capturing class-wise and directional adversarial vulnerability, even for an adversarially trained model.

We omit the unified SOIV matrix because it would be considerably large for 100 classes. Nevertheless, since both SOIV-U and SOIV-T values are computed, the unified matrix can be readily constructed following the same procedure.

Table 19: Comparison between SOIV-U ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) for **CIFAR-100 + Adversarially Trained PreActResNet-18**. Due to page width constraints, class indices are used instead of class names.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	MIXED
1	55 (8.85e-01)	80 (94.12)	77 (95.00)	80 (94.12)	55 (73.33)	80 (88.24)
2	27 (8.63e-01)	55 (93.33)	80 (94.12)	77 (90.00)	80 (70.59)	55 (85.00)
3	65 (8.54e-01)	77 (90.00)	55 (93.33)	38 (88.46)	4 (68.00)	77 (83.75)

4	80 (8.43e-01)	4 (80.00)	33 (86.21)	44 (85.00)	42 (64.00)	4 (79.00)
5	72 (8.36e-01)	44 (80.00)	44 (85.00)	4 (84.00)	74 (64.00)	38 (76.92)
6	98 (8.34e-01)	38 (76.92)	38 (84.62)	74 (84.00)	35 (62.50)	44 (76.25)
7	38 (8.24e-01)	3 (76.47)	4 (84.00)	72 (81.82)	96 (62.50)	74 (75.00)
8	3 (8.21e-01)	42 (76.00)	96 (81.25)	55 (80.00)	2 (61.54)	42 (74.00)
9	44 (8.20e-01)	74 (76.00)	42 (80.00)	79 (78.57)	77 (60.00)	35 (71.88)
10	88 (8.19e-01)	35 (75.00)	35 (79.17)	42 (76.00)	34 (58.62)	34 (68.97)
11	4 (8.14e-01)	2 (73.08)	79 (78.57)	66 (73.53)	38 (57.69)	96 (68.75)
12	77 (8.12e-01)	27 (72.73)	98 (77.78)	33 (72.41)	44 (55.00)	72 (68.18)
13	35 (8.08e-01)	72 (72.73)	25 (77.27)	34 (72.41)	50 (52.94)	79 (67.86)
14	18 (8.04e-01)	34 (72.41)	27 (77.27)	35 (70.83)	25 (50.00)	2 (67.31)
15	74 (8.03e-01)	98 (72.22)	2 (76.92)	50 (70.59)	27 (50.00)	50 (66.18)
16	78 (7.97e-01)	50 (70.59)	3 (76.47)	88 (70.37)	79 (50.00)	27 (65.91)
17	31 (7.96e-01)	33 (68.97)	18 (76.32)	3 (64.71)	11 (47.62)	33 (65.52)
18	42 (7.87e-01)	96 (68.75)	74 (76.00)	25 (63.64)	72 (45.45)	25 (63.64)
19	33 (7.83e-01)	79 (64.29)	78 (74.07)	27 (63.64)	98 (44.44)	3 (63.24)
20	45 (7.80e-01)	25 (63.64)	72 (72.73)	96 (62.50)	63 (40.74)	98 (62.50)
21	37 (7.77e-01)	18 (63.16)	34 (72.41)	14 (61.29)	18 (39.47)	88 (57.41)
22	64 (7.77e-01)	14 (61.29)	67 (70.83)	64 (60.87)	14 (38.71)	18 (56.58)
23	50 (7.76e-01)	5 (59.38)	50 (70.59)	43 (60.00)	90 (38.71)	14 (56.45)
24	2 (7.74e-01)	63 (59.26)	6 (70.00)	5 (59.38)	10 (38.10)	63 (55.56)
25	84 (7.73e-01)	88 (59.26)	19 (69.23)	65 (59.09)	3 (35.29)	67 (55.21)
26	19 (7.67e-01)	67 (58.33)	63 (66.67)	15 (58.82)	59 (34.78)	11 (53.57)
27	96 (7.66e-01)	19 (57.69)	88 (66.67)	67 (58.33)	33 (34.48)	19 (52.88)
28	79 (7.65e-01)	11 (57.14)	14 (64.52)	97 (58.06)	43 (34.29)	43 (52.86)
29	10 (7.63e-01)	43 (57.14)	46 (64.29)	2 (57.69)	30 (33.33)	78 (51.85)
30	93 (7.58e-01)	46 (57.14)	10 (61.90)	63 (55.56)	45 (33.33)	5 (51.56)
31	43 (7.57e-01)	70 (56.25)	49 (61.11)	98 (55.56)	67 (33.33)	10 (51.19)
32	46 (7.43e-01)	78 (55.56)	59 (60.87)	19 (53.85)	78 (33.33)	65 (50.00)
33	34 (7.37e-01)	84 (55.17)	43 (60.00)	10 (52.38)	88 (33.33)	15 (49.26)
34	15 (7.17e-01)	15 (52.94)	5 (59.38)	11 (52.38)	85 (32.43)	46 (48.21)
35	11 (7.17e-01)	10 (52.38)	70 (59.38)	45 (52.38)	65 (31.82)	64 (47.83)
36	6 (7.03e-01)	45 (52.38)	65 (59.09)	36 (51.52)	92 (31.82)	59 (47.83)
37	81 (6.98e-01)	59 (52.17)	15 (58.82)	23 (50.00)	7 (31.25)	66 (47.79)
38	36 (6.88e-01)	6 (50.00)	11 (57.14)	93 (50.00)	70 (31.25)	45 (47.62)
39	73 (6.86e-01)	49 (50.00)	31 (57.14)	18 (47.37)	19 (30.77)	6 (46.67)
40	25 (6.82e-01)	65 (50.00)	37 (56.00)	8 (47.06)	64 (30.43)	70 (46.09)
41	92 (6.81e-01)	90 (48.39)	66 (55.88)	7 (46.88)	36 (30.30)	90 (45.97)
42	66 (6.78e-01)	37 (48.00)	84 (55.17)	89 (46.88)	97 (29.03)	49 (45.83)
43	89 (6.77e-01)	64 (47.83)	30 (54.55)	17 (46.15)	71 (28.95)	30 (43.94)
44	7 (6.75e-01)	7 (46.88)	81 (54.55)	30 (45.45)	46 (28.57)	36 (43.94)
45	67 (6.75e-01)	93 (44.44)	45 (52.38)	90 (45.16)	5 (28.12)	7 (43.75)
46	63 (6.73e-01)	8 (44.12)	64 (52.17)	32 (44.83)	40 (28.00)	97 (43.55)
47	13 (6.70e-01)	89 (43.75)	90 (51.61)	49 (44.44)	49 (27.78)	84 (43.10)
48	30 (6.63e-01)	31 (42.86)	36 (51.52)	78 (44.44)	32 (27.59)	89 (41.41)
49	83 (6.63e-01)	30 (42.42)	7 (50.00)	59 (43.48)	95 (27.27)	85 (41.22)
50	59 (6.63e-01)	36 (42.42)	8 (50.00)	85 (43.24)	52 (26.83)	8 (40.44)
51	51 (6.62e-01)	97 (41.94)	13 (50.00)	31 (42.86)	6 (26.67)	93 (40.28)
52	22 (6.54e-01)	66 (41.18)	89 (50.00)	46 (42.86)	83 (26.67)	31 (40.18)
53	97 (6.54e-01)	13 (40.91)	92 (50.00)	81 (42.42)	15 (26.47)	81 (40.15)
54	26 (6.52e-01)	92 (40.91)	51 (48.65)	69 (41.86)	91 (25.71)	37 (40.00)
55	14 (6.43e-01)	51 (40.54)	85 (48.65)	84 (41.38)	73 (25.00)	92 (39.77)
56	32 (6.38e-01)	85 (40.54)	87 (47.22)	6 (40.00)	87 (25.00)	87 (37.50)
57	47 (6.34e-01)	83 (40.00)	97 (45.16)	40 (40.00)	89 (25.00)	32 (37.07)
58	90 (6.32e-01)	81 (39.39)	61 (44.74)	26 (39.13)	81 (24.24)	13 (36.36)
59	5 (6.23e-01)	26 (39.13)	93 (44.44)	87 (38.89)	13 (22.73)	26 (35.87)
60	70 (6.18e-01)	87 (38.89)	26 (43.48)	28 (37.50)	57 (22.58)	51 (35.81)
61	8 (6.14e-01)	56 (38.10)	83 (43.33)	70 (37.50)	93 (22.22)	17 (35.26)
62	29 (6.09e-01)	71 (36.84)	32 (41.38)	92 (36.36)	26 (21.74)	40 (35.00)
63	16 (6.01e-01)	52 (36.59)	17 (41.03)	37 (36.00)	1 (21.62)	83 (35.00)
64	85 (5.96e-01)	40 (36.00)	21 (40.00)	57 (35.48)	51 (21.62)	71 (32.89)
65	21 (5.94e-01)	21 (35.56)	28 (40.00)	12 (35.29)	56 (21.43)	56 (32.74)
66	49 (5.77e-01)	57 (35.48)	71 (39.47)	22 (34.38)	84 (20.69)	57 (32.26)
67	40 (5.74e-01)	1 (35.14)	95 (39.39)	91 (34.29)	8 (20.59)	91 (32.14)
68	54 (5.72e-01)	32 (34.48)	52 (39.02)	61 (34.21)	12 (20.59)	21 (31.67)

69	87 (5.50e-01)	17 (33.33)	54 (38.10)	21 (33.33)	66 (20.59)	28 (31.25)
70	12 (5.44e-01)	28 (32.50)	56 (38.10)	56 (33.33)	17 (20.51)	95 (31.06)
71	17 (5.38e-01)	91 (31.43)	91 (37.14)	51 (32.43)	37 (20.00)	1 (30.41)
72	57 (5.36e-01)	99 (30.56)	40 (36.00)	73 (32.14)	99 (19.44)	52 (29.88)
73	61 (5.32e-01)	95 (30.30)	57 (35.48)	13 (31.82)	22 (18.75)	61 (29.61)
74	99 (5.30e-01)	12 (29.41)	1 (35.14)	9 (30.56)	31 (17.86)	73 (29.46)
75	62 (5.27e-01)	47 (29.17)	22 (34.38)	99 (30.56)	21 (17.78)	22 (28.91)
76	23 (5.27e-01)	16 (28.57)	47 (33.33)	83 (30.00)	47 (16.67)	12 (28.68)
77	56 (5.16e-01)	54 (28.57)	99 (33.33)	1 (29.73)	54 (16.67)	99 (28.47)
78	1 (5.13e-01)	73 (28.57)	58 (32.43)	95 (27.27)	61 (15.79)	69 (27.91)
79	28 (5.10e-01)	22 (28.12)	16 (32.14)	58 (27.03)	48 (15.56)	54 (27.38)
80	9 (4.84e-01)	69 (27.91)	73 (32.14)	71 (26.32)	28 (15.00)	23 (25.69)
81	48 (4.83e-01)	39 (26.83)	39 (31.71)	54 (26.19)	39 (14.63)	47 (25.00)
82	52 (4.62e-01)	48 (26.67)	69 (30.23)	29 (25.81)	53 (13.95)	58 (24.32)
83	39 (4.61e-01)	58 (24.32)	12 (29.41)	75 (22.92)	68 (13.95)	39 (23.17)
84	69 (4.59e-01)	61 (23.68)	48 (28.89)	86 (21.88)	9 (13.89)	48 (22.78)
85	71 (4.50e-01)	41 (23.08)	41 (28.21)	47 (20.83)	94 (13.64)	16 (22.32)
86	86 (4.38e-01)	23 (22.22)	94 (25.00)	76 (20.51)	58 (13.51)	9 (21.53)
87	58 (4.33e-01)	68 (20.93)	68 (23.26)	48 (20.00)	41 (12.82)	41 (20.51)
88	68 (4.32e-01)	75 (20.83)	29 (22.58)	39 (19.51)	75 (12.50)	75 (19.27)
89	95 (4.31e-01)	60 (19.57)	9 (22.22)	68 (18.60)	86 (12.50)	68 (19.19)
90	75 (4.13e-01)	9 (19.44)	23 (22.22)	94 (18.18)	69 (11.63)	94 (18.75)
91	91 (4.08e-01)	86 (18.75)	60 (21.74)	41 (17.95)	16 (10.71)	86 (17.97)
92	76 (3.82e-01)	94 (18.18)	75 (20.83)	16 (17.86)	76 (10.26)	29 (17.74)
93	94 (3.66e-01)	62 (17.95)	62 (20.51)	52 (17.07)	82 (9.09)	53 (15.70)
94	41 (3.61e-01)	53 (16.28)	86 (18.75)	53 (16.28)	23 (8.33)	60 (15.22)
95	60 (3.44e-01)	29 (16.13)	0 (17.50)	82 (15.91)	60 (6.52)	62 (14.74)
96	0 (3.37e-01)	82 (15.91)	53 (16.28)	62 (15.38)	29 (6.45)	76 (14.74)
97	24 (3.12e-01)	0 (15.00)	82 (15.91)	20 (13.33)	62 (5.13)	82 (14.20)
98	53 (2.72e-01)	76 (12.82)	76 (15.38)	60 (13.04)	24 (4.65)	20 (10.56)
99	82 (2.56e-01)	20 (11.11)	24 (13.95)	24 (9.30)	20 (4.44)	0 (10.00)
100	20 (2.42e-01)	24 (6.98)	20 (13.33)	0 (5.00)	0 (2.50)	24 (8.72)
Pearson		0.8558 (8.3e-30)	0.8692 (9.8e-32)	0.8167 (3.9e-25)	0.7420 (1.0e-18)	0.8552 (1.0e-29)
Spearman		0.8825 (7.1e-34)	0.8832 (5.4e-34)	0.8360 (2.7e-27)	0.7917 (1.1e-22)	0.8810 (1.3e-33)
Kendall		0.7048 (2.8e-25)	0.7068 (2.0e-25)	0.6485 (1.2e-21)	0.6020 (7.0e-19)	0.6971 (9.2e-25)

Table 20: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (I-FGSM, PGD, C&W, and MIXED) when the source class is 0 (Apple) for **CIFAR-100 + Adversarially Trained PreActResNet-18**.

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
1	53 (-0.035224)	83 (87.50)	83 (95.00)	83 (100.00)	83 (94.17)
2	83 (-0.028484)	10 (72.50)	10 (77.50)	26 (97.50)	70 (80.83)
3	57 (-0.023557)	70 (70.00)	70 (77.50)	57 (95.00)	10 (80.00)
4	70 (-0.018124)	57 (67.50)	57 (72.50)	70 (95.00)	57 (78.33)
5	92 (-0.015560)	62 (65.00)	62 (67.50)	36 (92.50)	62 (72.50)
6	10 (-0.011853)	53 (62.50)	82 (67.50)	10 (90.00)	36 (69.17)
7	62 (-0.009505)	82 (57.50)	53 (65.00)	62 (85.00)	26 (68.33)
8	45 (-0.007237)	25 (55.00)	16 (62.50)	32 (82.50)	53 (68.33)
9	61 (-0.006246)	32 (55.00)	22 (62.50)	11 (80.00)	32 (65.00)
10	1 (-0.005304)	61 (55.00)	36 (62.50)	22 (80.00)	25 (63.33)
11	54 (-0.005114)	26 (52.50)	47 (60.00)	45 (80.00)	82 (63.33)
12	26 (-0.004982)	36 (52.50)	61 (60.00)	25 (77.50)	22 (63.33)
13	2 (-0.004570)	92 (52.50)	94 (60.00)	53 (77.50)	61 (62.50)
14	36 (-0.004524)	11 (50.00)	25 (57.50)	94 (77.50)	94 (62.50)
15	84 (-0.004211)	16 (50.00)	32 (57.50)	43 (75.00)	16 (60.83)
16	77 (-0.003864)	47 (50.00)	86 (57.50)	7 (72.50)	11 (60.00)
17	51 (-0.003598)	94 (50.00)	26 (55.00)	61 (72.50)	47 (60.00)
18	32 (-0.003572)	22 (47.50)	13 (52.50)	16 (70.00)	45 (58.33)
19	7 (-0.003549)	77 (47.50)	92 (52.50)	47 (70.00)	92 (57.50)
20	25 (-0.003412)	28 (45.00)	98 (52.50)	84 (67.50)	43 (54.17)

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
21	11 (-0.002962)	45 (45.00)	11 (50.00)	92 (67.50)	98 (53.33)
22	22 (-0.002908)	40 (42.50)	28 (50.00)	4 (65.00)	86 (51.67)
23	16 (-0.002835)	46 (42.50)	45 (50.00)	82 (65.00)	77 (51.67)
24	94 (-0.002699)	87 (42.50)	77 (50.00)	98 (65.00)	46 (50.00)
25	50 (-0.002662)	88 (42.50)	14 (47.50)	46 (60.00)	7 (48.33)
26	35 (-0.002642)	98 (42.50)	40 (47.50)	50 (60.00)	4 (46.67)
27	28 (-0.002565)	5 (40.00)	43 (47.50)	86 (60.00)	54 (45.83)
28	98 (-0.002528)	43 (40.00)	46 (47.50)	2 (57.50)	84 (45.83)
29	24 (-0.002445)	54 (40.00)	54 (45.00)	77 (57.50)	28 (45.00)
30	47 (-0.002307)	1 (37.50)	87 (45.00)	6 (55.00)	40 (45.00)
31	63 (-0.002258)	6 (37.50)	88 (45.00)	51 (55.00)	88 (45.00)
32	65 (-0.002257)	13 (37.50)	1 (42.50)	35 (52.50)	2 (45.00)
33	4 (-0.002241)	63 (37.50)	2 (42.50)	54 (52.50)	13 (44.17)
34	43 (-0.002237)	86 (37.50)	5 (42.50)	63 (52.50)	6 (44.17)
35	78 (-0.002196)	2 (35.00)	9 (42.50)	93 (52.50)	51 (43.33)
36	14 (-0.002170)	4 (35.00)	58 (42.50)	15 (47.50)	63 (42.50)
37	86 (-0.002167)	7 (35.00)	81 (42.50)	78 (47.50)	35 (41.67)
38	67 (-0.002159)	9 (35.00)	4 (40.00)	88 (47.50)	50 (41.67)
39	41 (-0.002025)	14 (35.00)	6 (40.00)	40 (45.00)	5 (40.83)
40	80 (-0.002015)	51 (35.00)	35 (40.00)	13 (42.50)	1 (40.00)
41	5 (-0.002004)	58 (35.00)	51 (40.00)	19 (42.50)	14 (40.00)
42	46 (-0.001978)	15 (32.50)	84 (40.00)	1 (40.00)	93 (39.17)
43	93 (-0.001947)	19 (32.50)	7 (37.50)	5 (40.00)	81 (38.33)
44	6 (-0.001937)	29 (32.50)	19 (37.50)	21 (40.00)	15 (38.33)
45	82 (-0.001936)	31 (32.50)	29 (37.50)	28 (40.00)	78 (38.33)
46	34 (-0.001915)	35 (32.50)	42 (37.50)	29 (40.00)	19 (37.50)
47	3 (-0.001889)	41 (32.50)	63 (37.50)	65 (40.00)	87 (37.50)
48	40 (-0.001887)	50 (32.50)	89 (37.50)	72 (40.00)	29 (36.67)
49	15 (-0.001863)	78 (32.50)	3 (35.00)	81 (40.00)	9 (35.00)
50	19 (-0.001849)	81 (32.50)	15 (35.00)	14 (37.50)	21 (34.17)
51	9 (-0.001847)	90 (32.50)	21 (35.00)	85 (37.50)	58 (34.17)
52	99 (-0.001823)	3 (30.00)	31 (35.00)	3 (35.00)	3 (33.33)
53	88 (-0.001784)	20 (30.00)	39 (35.00)	24 (35.00)	42 (32.50)
54	89 (-0.001737)	33 (30.00)	41 (35.00)	34 (35.00)	31 (31.67)
55	64 (-0.001711)	42 (30.00)	48 (35.00)	64 (35.00)	64 (31.67)
56	72 (-0.001686)	71 (30.00)	78 (35.00)	67 (32.50)	85 (31.67)
57	55 (-0.001684)	84 (30.00)	90 (35.00)	80 (32.50)	90 (31.67)
58	90 (-0.001671)	89 (30.00)	93 (35.00)	96 (32.50)	80 (30.83)
59	81 (-0.001671)	93 (30.00)	20 (32.50)	42 (30.00)	96 (30.83)
60	48 (-0.001635)	96 (30.00)	37 (32.50)	55 (30.00)	34 (30.83)
61	74 (-0.001631)	21 (27.50)	50 (32.50)	99 (30.00)	41 (30.83)
62	39 (-0.001575)	27 (27.50)	64 (32.50)	9 (27.50)	65 (30.83)
63	18 (-0.001571)	34 (27.50)	80 (32.50)	31 (27.50)	89 (30.83)
64	42 (-0.001542)	37 (27.50)	91 (32.50)	37 (27.50)	99 (30.00)
65	85 (-0.001529)	39 (27.50)	99 (32.50)	74 (27.50)	37 (29.17)
66	13 (-0.001524)	44 (27.50)	8 (30.00)	90 (27.50)	67 (29.17)
67	87 (-0.001510)	48 (27.50)	18 (30.00)	91 (27.50)	91 (29.17)
68	33 (-0.001494)	64 (27.50)	34 (30.00)	18 (25.00)	48 (29.17)
69	21 (-0.001464)	66 (27.50)	66 (30.00)	41 (25.00)	72 (28.33)
70	58 (-0.001456)	80 (27.50)	67 (30.00)	48 (25.00)	24 (28.33)
71	29 (-0.001395)	85 (27.50)	71 (30.00)	58 (25.00)	20 (27.50)
72	96 (-0.001381)	91 (27.50)	85 (30.00)	79 (25.00)	55 (26.67)
73	37 (-0.001378)	99 (27.50)	96 (30.00)	87 (25.00)	66 (26.67)
74	69 (-0.001366)	8 (25.00)	24 (27.50)	89 (25.00)	71 (26.67)
75	31 (-0.001345)	23 (25.00)	27 (27.50)	44 (22.50)	18 (25.83)
76	79 (-0.001339)	38 (25.00)	33 (27.50)	52 (22.50)	44 (25.83)
77	91 (-0.001335)	65 (25.00)	44 (27.50)	66 (22.50)	39 (25.83)
78	66 (-0.001318)	67 (25.00)	55 (27.50)	69 (22.50)	33 (25.00)
79	23 (-0.001314)	12 (22.50)	56 (27.50)	20 (20.00)	27 (24.17)
80	38 (-0.001308)	18 (22.50)	65 (27.50)	71 (20.00)	69 (23.33)
81	71 (-0.001306)	24 (22.50)	12 (25.00)	97 (20.00)	23 (22.50)
82	27 (-0.001300)	55 (22.50)	17 (25.00)	23 (17.50)	38 (22.50)

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
83	52 (-0.001287)	69 (22.50)	23 (25.00)	27 (17.50)	8 (20.00)
84	73 (-0.001284)	72 (22.50)	38 (25.00)	33 (17.50)	17 (20.00)
85	56 (-0.001284)	17 (20.00)	69 (25.00)	38 (17.50)	52 (20.00)
86	44 (-0.001278)	52 (20.00)	30 (22.50)	17 (15.00)	56 (20.00)
87	20 (-0.001263)	56 (17.50)	72 (22.50)	39 (15.00)	74 (20.00)
88	97 (-0.001229)	59 (17.50)	97 (22.50)	56 (15.00)	79 (20.00)
89	68 (-0.001186)	75 (17.50)	49 (20.00)	76 (15.00)	97 (20.00)
90	8 (-0.001151)	97 (17.50)	59 (20.00)	73 (12.50)	12 (17.50)
91	17 (-0.001117)	30 (15.00)	60 (20.00)	30 (10.00)	30 (15.83)
92	60 (-0.001048)	73 (15.00)	73 (20.00)	49 (7.50)	73 (15.83)
93	12 (-0.001005)	74 (15.00)	79 (20.00)	59 (7.50)	76 (15.83)
94	30 (-0.000989)	76 (15.00)	52 (17.50)	60 (7.50)	59 (15.00)
95	95 (-0.000980)	79 (15.00)	68 (17.50)	68 (7.50)	60 (13.33)
96	59 (-0.000977)	60 (12.50)	74 (17.50)	8 (5.00)	49 (12.50)
97	76 (-0.000957)	68 (12.50)	75 (17.50)	12 (5.00)	68 (12.50)
98	49 (-0.000889)	49 (10.00)	76 (17.50)	75 (2.50)	75 (12.50)
99	75 (-0.000862)	95 (7.50)	95 (12.50)	95 (0.00)	95 (6.67)
	Pearson	0.6845 (5.7e-15)	0.6411 (8.8e-13)	0.5716 (6.4e-10)	0.643 (7.2e-13)
	Spearman	0.814 (1.3e-24)	0.8117 (2.2e-24)	0.9115 (3.3e-39)	0.8898 (8.2e-35)
	Kendall	0.6448 (3.2e-21)	0.6361 (1.1e-20)	0.7567 (1.3e-28)	0.7207 (6.9e-26)

C.5 CIFAR-100 + Adversarially Trained ResNet-18

For CIFAR-100 + Adversarially Trained ResNet-18, we report the same set of ranking comparisons using another adversarially trained model obtained from RobustBench (Croce et al., 2020; Addepalli et al., 2022b). The results are summarized in Table 21 and Table 22. As in the previous setting, I-FGSM is used in place of FGSM for the targeted experiments. Strong rank correlations are observed under both untargeted and targeted settings. These results suggest that the proposed metrics effectively capture class-wise and directional adversarial vulnerability, even for an adversarially trained model.

We omit the unified SOIV matrix because it would be considerably large for 100 classes. Nevertheless, since both SOIV-U and SOIV-T values are computed, the unified matrix can be readily constructed following the same procedure.

Table 21: Comparison between SOIV-U ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) for **CIFAR-100 + Adversarially Trained ResNet-18**. Due to page width constraints, class indices are used instead of class names.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	MIXED
1	55 (8.90e-01)	80 (90.00)	80 (95.00)	80 (90.00)	96 (78.26)	80 (83.75)
2	27 (8.64e-01)	77 (85.19)	55 (93.33)	77 (85.19)	4 (74.07)	77 (79.63)
3	72 (8.62e-01)	96 (82.61)	77 (92.59)	4 (81.48)	50 (61.90)	96 (79.35)
4	65 (8.61e-01)	55 (80.00)	42 (85.71)	44 (76.19)	38 (61.76)	4 (78.70)
5	98 (8.45e-01)	4 (77.78)	2 (82.76)	42 (75.00)	42 (60.71)	55 (76.67)
6	80 (8.44e-01)	98 (76.47)	35 (82.61)	96 (73.91)	55 (60.00)	42 (74.11)
7	38 (8.40e-01)	79 (75.76)	96 (82.61)	55 (73.33)	80 (60.00)	38 (69.85)
8	3 (8.38e-01)	10 (75.00)	4 (81.48)	66 (71.43)	72 (56.25)	44 (67.86)
9	44 (8.34e-01)	42 (75.00)	72 (81.25)	34 (67.74)	77 (55.56)	34 (67.74)
10	4 (8.27e-01)	34 (74.19)	78 (80.65)	38 (67.65)	35 (52.17)	72 (67.19)
11	77 (8.24e-01)	35 (73.91)	38 (79.41)	27 (66.67)	27 (51.85)	50 (66.67)
12	35 (8.24e-01)	3 (73.68)	3 (78.95)	74 (66.67)	34 (51.61)	35 (66.30)
13	74 (8.19e-01)	44 (71.43)	79 (78.79)	3 (63.16)	44 (47.62)	3 (65.79)
14	18 (8.16e-01)	50 (71.43)	34 (77.42)	72 (62.50)	3 (47.37)	27 (64.81)
15	64 (8.14e-01)	74 (71.43)	98 (76.47)	63 (62.07)	2 (44.83)	74 (64.29)
16	31 (8.09e-01)	38 (70.59)	44 (76.19)	64 (60.87)	74 (42.86)	79 (62.88)
17	33 (8.07e-01)	46 (70.59)	50 (76.19)	65 (60.00)	10 (41.67)	10 (61.46)
18	88 (8.05e-01)	72 (68.75)	74 (76.19)	15 (58.82)	63 (41.38)	2 (60.34)

19	78 (7.97e-01)	27 (66.67)	33 (75.00)	10 (58.33)	46 (41.18)	63 (60.34)
20	96 (7.95e-01)	5 (65.71)	27 (74.07)	88 (58.06)	70 (40.00)	98 (58.82)
21	42 (7.93e-01)	2 (65.52)	63 (72.41)	79 (57.58)	79 (39.39)	46 (57.35)
22	45 (7.91e-01)	63 (65.52)	65 (72.00)	50 (57.14)	90 (39.39)	43 (56.08)
23	50 (7.87e-01)	78 (64.52)	10 (70.83)	43 (56.76)	18 (38.89)	78 (55.65)
24	79 (7.87e-01)	25 (62.50)	46 (70.59)	35 (56.52)	19 (37.93)	65 (54.00)
25	93 (7.86e-01)	37 (62.50)	18 (69.44)	89 (55.26)	43 (37.84)	5 (53.57)
26	2 (7.85e-01)	43 (62.16)	5 (68.57)	8 (54.05)	6 (35.90)	19 (53.45)
27	37 (7.82e-01)	6 (61.54)	43 (67.57)	37 (53.12)	98 (35.29)	64 (53.26)
28	84 (7.81e-01)	33 (60.71)	6 (66.67)	19 (51.72)	64 (34.78)	6 (53.21)
29	19 (7.79e-01)	11 (60.00)	25 (66.67)	25 (50.00)	93 (34.78)	25 (53.12)
30	43 (7.73e-01)	31 (59.38)	19 (65.52)	6 (48.72)	37 (34.38)	37 (53.12)
31	10 (7.70e-01)	92 (59.09)	64 (65.22)	5 (48.57)	89 (34.21)	33 (52.68)
32	11 (7.57e-01)	19 (58.62)	81 (64.86)	36 (48.39)	14 (33.33)	88 (51.61)
33	34 (7.46e-01)	18 (58.33)	83 (64.52)	78 (48.39)	25 (33.33)	89 (50.00)
34	46 (7.44e-01)	88 (58.06)	92 (63.64)	2 (48.28)	84 (33.33)	66 (50.00)
35	81 (7.34e-01)	65 (56.00)	31 (62.50)	46 (47.06)	15 (32.35)	18 (50.00)
36	6 (7.27e-01)	89 (55.26)	37 (62.50)	98 (47.06)	83 (32.26)	15 (49.26)
37	15 (7.26e-01)	70 (54.29)	88 (61.29)	31 (46.88)	92 (31.82)	70 (47.86)
38	73 (7.25e-01)	64 (52.17)	11 (60.00)	97 (46.88)	5 (31.43)	31 (47.66)
39	36 (7.24e-01)	14 (51.85)	7 (58.82)	33 (46.43)	32 (31.03)	92 (46.59)
40	63 (7.21e-01)	59 (51.72)	84 (57.58)	32 (44.83)	59 (31.03)	84 (46.21)
41	67 (7.19e-01)	83 (51.61)	15 (55.88)	7 (44.12)	73 (31.03)	90 (45.45)
42	51 (7.15e-01)	84 (51.52)	14 (55.56)	93 (43.48)	11 (30.00)	14 (45.37)
43	92 (7.09e-01)	66 (51.43)	89 (55.26)	69 (43.18)	7 (29.41)	11 (45.00)
44	66 (7.03e-01)	81 (51.35)	67 (54.55)	70 (42.86)	36 (29.03)	7 (44.85)
45	59 (7.03e-01)	15 (50.00)	66 (54.29)	84 (42.42)	78 (29.03)	83 (44.35)
46	83 (7.02e-01)	90 (48.48)	70 (54.29)	90 (42.42)	88 (29.03)	59 (43.97)
47	7 (7.00e-01)	93 (47.83)	49 (52.50)	45 (42.31)	33 (28.57)	93 (43.48)
48	13 (6.96e-01)	85 (47.37)	32 (51.72)	87 (41.67)	95 (28.57)	81 (43.24)
49	30 (6.92e-01)	7 (47.06)	59 (51.72)	59 (41.38)	65 (28.00)	32 (43.10)
50	89 (6.84e-01)	8 (45.95)	90 (51.52)	14 (40.74)	45 (26.92)	8 (42.57)
51	22 (6.81e-01)	67 (45.45)	8 (51.35)	85 (39.47)	85 (26.32)	36 (41.13)
52	47 (6.77e-01)	32 (44.83)	45 (50.00)	51 (38.24)	16 (25.71)	85 (40.13)
53	14 (6.77e-01)	87 (44.44)	93 (47.83)	23 (37.84)	30 (25.00)	49 (38.75)
54	97 (6.77e-01)	52 (42.86)	85 (47.37)	81 (37.84)	49 (25.00)	67 (38.64)
55	32 (6.74e-01)	49 (42.50)	87 (47.22)	1 (36.59)	57 (25.00)	45 (38.46)
56	90 (6.68e-01)	36 (41.94)	51 (47.06)	67 (36.36)	1 (24.39)	97 (38.28)
57	26 (6.67e-01)	13 (40.91)	30 (46.88)	21 (35.71)	66 (22.86)	87 (38.19)
58	70 (6.63e-01)	30 (40.62)	97 (46.88)	49 (35.00)	91 (22.86)	30 (36.72)
59	25 (6.61e-01)	95 (40.00)	52 (45.24)	40 (34.48)	69 (22.73)	51 (36.03)
60	5 (6.45e-01)	61 (39.47)	36 (45.16)	30 (34.38)	47 (22.22)	69 (34.09)
61	21 (6.40e-01)	51 (38.24)	40 (44.83)	18 (33.33)	31 (21.88)	57 (33.59)
62	29 (6.38e-01)	40 (37.93)	54 (44.44)	26 (33.33)	97 (21.88)	95 (33.57)
63	49 (6.33e-01)	54 (37.78)	61 (42.11)	92 (31.82)	52 (21.43)	1 (33.54)
64	8 (6.30e-01)	57 (37.50)	13 (40.91)	57 (31.25)	26 (20.83)	40 (32.76)
65	85 (6.29e-01)	97 (37.50)	47 (40.74)	11 (30.00)	51 (20.59)	13 (31.82)
66	23 (6.13e-01)	21 (35.71)	57 (40.62)	17 (29.73)	71 (20.51)	54 (31.67)
67	16 (6.10e-01)	45 (34.62)	17 (40.54)	83 (29.03)	53 (20.45)	21 (31.55)
68	54 (6.08e-01)	73 (34.48)	95 (40.00)	16 (28.57)	87 (19.44)	26 (31.25)
69	12 (6.01e-01)	56 (34.21)	1 (39.02)	39 (28.57)	8 (18.92)	73 (31.03)
70	99 (5.98e-01)	1 (34.15)	99 (38.89)	9 (27.50)	81 (18.92)	52 (30.95)
71	17 (5.90e-01)	69 (34.09)	21 (38.10)	13 (27.27)	56 (18.42)	61 (30.26)
72	40 (5.84e-01)	26 (33.33)	26 (37.50)	54 (26.67)	13 (18.18)	16 (30.00)
73	61 (5.83e-01)	47 (33.33)	16 (37.14)	61 (26.32)	67 (18.18)	71 (28.85)
74	56 (5.82e-01)	71 (33.33)	56 (36.84)	95 (25.71)	86 (18.18)	47 (28.70)
75	62 (5.77e-01)	17 (32.43)	69 (36.36)	58 (25.64)	54 (17.78)	17 (28.38)
76	87 (5.73e-01)	58 (30.77)	58 (35.90)	71 (25.64)	9 (17.50)	56 (28.29)
77	57 (5.68e-01)	16 (28.57)	71 (35.90)	28 (25.00)	76 (17.07)	23 (27.70)
78	1 (5.44e-01)	91 (28.57)	73 (34.48)	29 (24.24)	21 (16.67)	58 (26.92)
79	28 (5.43e-01)	12 (28.12)	39 (33.33)	86 (24.24)	94 (16.28)	91 (26.43)
80	52 (5.38e-01)	99 (27.78)	91 (31.43)	73 (24.14)	23 (16.22)	39 (25.60)
81	71 (5.32e-01)	9 (27.50)	12 (31.25)	56 (23.68)	58 (15.38)	9 (25.00)
82	69 (5.31e-01)	29 (27.27)	28 (30.00)	75 (23.40)	29 (15.15)	99 (24.31)
83	68 (5.16e-01)	23 (27.03)	75 (29.79)	91 (22.86)	39 (14.29)	29 (23.48)

84	95 (5.03e-01)	62 (26.83)	23 (29.73)	12 (21.88)	40 (13.79)	86 (22.73)
85	48 (5.02e-01)	39 (26.19)	62 (29.27)	22 (21.21)	68 (13.64)	12 (22.66)
86	39 (4.96e-01)	94 (25.58)	9 (27.50)	94 (20.93)	61 (13.16)	28 (22.50)
87	9 (4.95e-01)	28 (25.00)	41 (27.50)	41 (20.00)	60 (13.04)	94 (22.09)
88	75 (4.89e-01)	86 (24.24)	22 (27.27)	76 (19.51)	41 (12.50)	75 (21.81)
89	58 (4.85e-01)	75 (23.40)	29 (27.27)	99 (19.44)	22 (12.12)	41 (20.62)
90	76 (4.65e-01)	68 (22.73)	94 (25.58)	47 (18.52)	99 (11.11)	22 (20.45)
91	86 (4.62e-01)	41 (22.50)	86 (24.24)	48 (15.91)	17 (10.81)	76 (20.12)
92	91 (4.45e-01)	76 (21.95)	53 (22.73)	53 (15.91)	75 (10.64)	62 (20.12)
93	94 (4.29e-01)	22 (21.21)	68 (22.73)	68 (15.91)	28 (10.00)	53 (19.89)
94	60 (4.20e-01)	48 (20.45)	76 (21.95)	62 (14.63)	62 (9.76)	68 (18.75)
95	0 (3.76e-01)	53 (20.45)	60 (21.74)	52 (14.29)	12 (9.38)	48 (15.91)
96	24 (3.73e-01)	60 (19.57)	48 (20.45)	24 (13.95)	82 (9.09)	60 (14.67)
97	41 (3.72e-01)	82 (15.91)	24 (18.60)	82 (13.64)	24 (6.98)	82 (13.64)
98	53 (3.55e-01)	24 (11.63)	82 (15.91)	20 (11.36)	48 (6.82)	24 (12.79)
99	82 (3.21e-01)	20 (11.36)	20 (11.36)	0 (5.00)	0 (5.00)	20 (9.66)
100	20 (2.77e-01)	0 (10.00)	0 (10.00)	60 (4.35)	20 (4.55)	0 (7.50)
Pearson		0.8547 (1.2e-29)	0.8822 (8.0e-34)	0.8004 (1.7e-23)	0.7451 (6.2e-19)	0.8608 (1.7e-30)
Spearman		0.8678 (1.6e-31)	0.8930 (9.4e-36)	0.8263 (3.6e-26)	0.8087 (2.6e-24)	0.8867 (1.4e-34)
Kendall		0.6795 (1.3e-23)	0.7151 (5.5e-26)	0.6330 (1.0e-20)	0.6058 (4.2e-19)	0.6981 (7.9e-25)

Table 22: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (I-FGSM, PGD, C&W, and MIXED) when the source class is 0 (Apple) for **CIFAR-100 + Adversarially Trained ResNet-18**.

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
1	83 (-0.031213)	83 (97.50)	83 (97.50)	26 (100.00)	83 (98.33)
2	53 (-0.028814)	10 (90.00)	10 (92.50)	70 (100.00)	10 (91.67)
3	57 (-0.021192)	70 (77.50)	22 (80.00)	83 (100.00)	70 (85.83)
4	70 (-0.015888)	22 (70.00)	70 (80.00)	10 (92.50)	22 (79.17)
5	10 (-0.014880)	53 (70.00)	82 (75.00)	36 (92.50)	26 (78.33)
6	92 (-0.014000)	57 (70.00)	26 (72.50)	57 (92.50)	57 (77.50)
7	62 (-0.008992)	25 (62.50)	36 (72.50)	22 (87.50)	36 (75.00)
8	61 (-0.007863)	26 (62.50)	53 (72.50)	25 (87.50)	53 (75.00)
9	26 (-0.007461)	47 (62.50)	62 (72.50)	32 (87.50)	25 (72.50)
10	45 (-0.006376)	61 (62.50)	57 (70.00)	45 (87.50)	62 (70.83)
11	1 (-0.006021)	62 (62.50)	86 (70.00)	11 (82.50)	32 (69.17)
12	54 (-0.005958)	32 (60.00)	25 (67.50)	53 (82.50)	61 (69.17)
13	77 (-0.005619)	36 (60.00)	47 (67.50)	16 (80.00)	45 (67.50)
14	32 (-0.005599)	82 (60.00)	61 (65.00)	61 (80.00)	47 (66.67)
15	36 (-0.005446)	16 (57.50)	94 (62.50)	62 (77.50)	16 (65.83)
16	2 (-0.005248)	94 (57.50)	11 (60.00)	94 (75.00)	82 (65.83)
17	7 (-0.004702)	45 (55.00)	16 (60.00)	43 (72.50)	11 (65.00)
18	84 (-0.004668)	11 (52.50)	32 (60.00)	46 (72.50)	94 (65.00)
19	51 (-0.004639)	13 (52.50)	45 (60.00)	84 (72.50)	86 (61.67)
20	25 (-0.004364)	92 (52.50)	13 (57.50)	92 (72.50)	92 (60.00)
21	22 (-0.004362)	4 (50.00)	6 (55.00)	98 (72.50)	43 (59.17)
22	50 (-0.004057)	43 (50.00)	43 (55.00)	7 (70.00)	46 (56.67)
23	11 (-0.003646)	77 (50.00)	92 (55.00)	47 (70.00)	77 (55.83)
24	80 (-0.003441)	86 (50.00)	77 (52.50)	2 (67.50)	84 (55.83)
25	35 (-0.003438)	2 (47.50)	81 (52.50)	4 (65.00)	4 (55.00)
26	65 (-0.003396)	5 (47.50)	4 (50.00)	77 (65.00)	98 (54.17)
27	94 (-0.003244)	40 (47.50)	5 (50.00)	86 (65.00)	2 (54.17)
28	16 (-0.003226)	46 (47.50)	46 (50.00)	54 (62.50)	6 (53.33)
29	86 (-0.003166)	54 (47.50)	54 (50.00)	82 (62.50)	54 (53.33)
30	4 (-0.003069)	6 (45.00)	84 (50.00)	6 (60.00)	7 (52.50)
31	28 (-0.002999)	63 (45.00)	88 (50.00)	35 (60.00)	63 (50.83)
32	43 (-0.002950)	84 (45.00)	1 (47.50)	63 (60.00)	13 (50.00)
33	98 (-0.002946)	87 (45.00)	2 (47.50)	50 (57.50)	35 (49.17)
34	72 (-0.002918)	1 (42.50)	14 (47.50)	40 (52.50)	40 (49.17)
35	93 (-0.002777)	7 (42.50)	28 (47.50)	65 (52.50)	50 (48.33)

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
36	47 (-0.002774)	50 (42.50)	35 (47.50)	28 (50.00)	88 (47.50)
37	78 (-0.002740)	81 (42.50)	40 (47.50)	88 (50.00)	5 (46.67)
38	67 (-0.002734)	88 (42.50)	63 (47.50)	93 (50.00)	28 (45.83)
39	5 (-0.002685)	98 (42.50)	87 (47.50)	3 (47.50)	3 (44.17)
40	63 (-0.002639)	3 (40.00)	89 (47.50)	72 (45.00)	87 (43.33)
41	64 (-0.002638)	28 (40.00)	98 (47.50)	78 (45.00)	93 (43.33)
42	46 (-0.002635)	35 (40.00)	3 (45.00)	80 (45.00)	78 (42.50)
43	88 (-0.002630)	14 (37.50)	7 (45.00)	5 (42.50)	81 (41.67)
44	6 (-0.002575)	29 (37.50)	9 (45.00)	15 (42.50)	1 (40.83)
45	3 (-0.002564)	64 (37.50)	20 (45.00)	51 (42.50)	65 (40.83)
46	19 (-0.002521)	78 (37.50)	29 (45.00)	13 (40.00)	29 (39.17)
47	34 (-0.002495)	93 (37.50)	50 (45.00)	64 (40.00)	64 (39.17)
48	15 (-0.002489)	19 (35.00)	78 (45.00)	34 (37.50)	80 (39.17)
49	55 (-0.002486)	37 (35.00)	37 (42.50)	42 (37.50)	15 (38.33)
50	74 (-0.002445)	41 (35.00)	39 (42.50)	55 (37.50)	51 (37.50)
51	24 (-0.002423)	58 (35.00)	90 (42.50)	87 (37.50)	89 (37.50)
52	40 (-0.002323)	65 (35.00)	93 (42.50)	96 (37.50)	34 (36.67)
53	42 (-0.002305)	67 (35.00)	15 (40.00)	21 (35.00)	21 (35.00)
54	33 (-0.002251)	80 (35.00)	19 (40.00)	29 (35.00)	31 (35.00)
55	18 (-0.002249)	89 (35.00)	21 (40.00)	31 (35.00)	67 (35.00)
56	99 (-0.002185)	9 (32.50)	31 (40.00)	1 (32.50)	96 (35.00)
57	14 (-0.002170)	15 (32.50)	34 (40.00)	67 (32.50)	72 (34.17)
58	87 (-0.002097)	18 (32.50)	64 (40.00)	18 (30.00)	37 (34.17)
59	82 (-0.002076)	20 (32.50)	41 (37.50)	81 (30.00)	42 (34.17)
60	44 (-0.002070)	34 (32.50)	42 (37.50)	89 (30.00)	14 (34.17)
61	79 (-0.002052)	39 (32.50)	48 (37.50)	74 (27.50)	19 (33.33)
62	38 (-0.002010)	51 (32.50)	51 (37.50)	85 (27.50)	18 (31.67)
63	96 (-0.001977)	71 (32.50)	58 (37.50)	99 (27.50)	9 (30.83)
64	27 (-0.001912)	90 (32.50)	67 (37.50)	19 (25.00)	90 (30.83)
65	81 (-0.001878)	91 (32.50)	80 (37.50)	33 (25.00)	85 (30.83)
66	66 (-0.001842)	96 (32.50)	56 (35.00)	37 (25.00)	20 (30.00)
67	89 (-0.001824)	21 (30.00)	65 (35.00)	44 (25.00)	39 (30.00)
68	31 (-0.001816)	27 (30.00)	66 (35.00)	79 (25.00)	55 (30.00)
69	37 (-0.001788)	31 (30.00)	85 (35.00)	27 (22.50)	99 (30.00)
70	9 (-0.001779)	33 (30.00)	91 (35.00)	38 (22.50)	91 (29.17)
71	21 (-0.001757)	48 (30.00)	96 (35.00)	66 (22.50)	58 (29.17)
72	39 (-0.001747)	85 (30.00)	99 (35.00)	23 (20.00)	27 (28.33)
73	85 (-0.001664)	42 (27.50)	18 (32.50)	24 (20.00)	33 (28.33)
74	23 (-0.001657)	44 (27.50)	27 (32.50)	91 (20.00)	44 (28.33)
75	97 (-0.001653)	56 (27.50)	44 (32.50)	14 (17.50)	66 (28.33)
76	29 (-0.001642)	66 (27.50)	71 (32.50)	71 (17.50)	71 (27.50)
77	90 (-0.001619)	72 (27.50)	17 (30.00)	90 (17.50)	41 (26.67)
78	13 (-0.001573)	99 (27.50)	23 (30.00)	9 (15.00)	56 (25.83)
79	91 (-0.001564)	8 (25.00)	33 (30.00)	39 (15.00)	23 (25.00)
80	58 (-0.001549)	23 (25.00)	72 (30.00)	56 (15.00)	38 (25.00)
81	48 (-0.001514)	38 (25.00)	8 (27.50)	58 (15.00)	79 (25.00)
82	41 (-0.001501)	55 (25.00)	12 (27.50)	20 (12.50)	48 (24.17)
83	56 (-0.001385)	12 (22.50)	24 (27.50)	52 (12.50)	74 (23.33)
84	8 (-0.001291)	17 (22.50)	38 (27.50)	97 (12.50)	24 (22.50)
85	52 (-0.001230)	79 (22.50)	55 (27.50)	69 (10.00)	8 (20.00)
86	20 (-0.001184)	24 (20.00)	69 (27.50)	8 (7.50)	69 (18.33)
87	71 (-0.001165)	74 (20.00)	79 (27.50)	30 (7.50)	97 (18.33)
88	59 (-0.001130)	75 (20.00)	97 (25.00)	41 (7.50)	17 (17.50)
89	73 (-0.001120)	30 (17.50)	49 (22.50)	59 (7.50)	12 (17.50)
90	30 (-0.001084)	59 (17.50)	59 (22.50)	48 (5.00)	59 (15.83)
91	17 (-0.001060)	69 (17.50)	73 (22.50)	73 (5.00)	30 (15.00)
92	69 (-0.001029)	97 (17.50)	74 (22.50)	76 (5.00)	52 (14.17)
93	12 (-0.000957)	49 (15.00)	30 (20.00)	12 (2.50)	49 (13.33)
94	75 (-0.000888)	52 (15.00)	75 (20.00)	49 (2.50)	73 (13.33)
95	68 (-0.000858)	68 (15.00)	76 (20.00)	68 (2.50)	75 (13.33)
96	49 (-0.000801)	73 (12.50)	68 (17.50)	17 (0.00)	76 (12.50)
97	60 (-0.000797)	76 (12.50)	52 (15.00)	60 (0.00)	68 (11.67)

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
98	76 (-0.000786)	60 (7.50)	60 (15.00)	75 (0.00)	60 (7.50)
99	95 (-0.000780)	95 (7.50)	95 (7.50)	95 (0.00)	95 (5.00)
	Pearson	0.7098 (2.0e-16)	0.6494 (3.6e-13)	0.6150 (1.3e-11)	0.6716 (2.8e-14)
	Spearman	0.8111 (2.5e-24)	0.7606 (6.7e-20)	0.9187 (6.4e-41)	0.8751 (2.4e-32)
	Kendall	0.6402 (6.1e-21)	0.5884 (6.2e-18)	0.7669 (2.4e-29)	0.7041 (6.8e-25)

C.6 IMAGENET-100 + VIT-L/16

For ImageNet-100 + ViT-L/16, we use a model pretrained on ImageNet-1K. Since the model outputs 1000-class logits, evaluation is performed using the 100 logits corresponding to the ImageNet-100 classes. We report ranking comparisons between the proposed SOIV metrics and the corresponding ASR under both untargeted and targeted settings. The corresponding results are provided in Table 23 and Table 24.

Since FGSM yields negligible attack success in the targeted setting, we replace it with Iterative FGSM (I-FGSM) for the targeted experiments. For each attack we evaluate several attack strengths and average the resulting class-wise ASRs, since a single strength did not yield class-wise ASRs that are sufficiently spread out for this dataset-model combination.

While SOIV exhibits strong correlations with ASR in many settings, comparatively weaker correlations are observed in more challenging large-scale settings. The rank coefficients are reduced because most target classes differ only slightly in both SOIV-T and ASR, so inversions among these near-ties lower the coefficients. In particular, the two target classes with the largest-magnitude SOIV-T values are also the two highest-ASR targets under every attack. Nevertheless, the consistently positive alignment suggests that second-order statistics still capture a substantial component of adversarial vulnerability across diverse architectures and datasets.

We omit the unified SOIV matrix because it would be considerably large for 100 classes. Nevertheless, since both SOIV-U and SOIV-T values are computed, the unified matrix can be readily constructed following the same procedure.

Table 23: Comparison between SOIV-U ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, and MIXED) for **ImageNet-100 + ViT-L/16**. Classes are identified by their WordNet synset IDs. Each attack is evaluated at multiple attack strengths and the resulting ASRs are averaged. In this experiment, Square Attack is omitted due to its high computational cost for large-scale evaluation.

Rank	SOIV-U	FGSM	PGD	C&W	MIXED
1	n02089973 (3.79e-1)	n02089973 (76.92)	n02085620 (65.91)	n02093428 (72.92)	n02085620 (69.19)
2	n04589890 (2.62e-1)	n02085620 (75.76)	n02109047 (63.64)	n02109047 (68.18)	n02093428 (67.36)
3	n04127249 (2.27e-1)	n02093428 (70.83)	n02089973 (61.54)	n02085620 (65.91)	n02109047 (66.16)
4	n03764736 (2.21e-1)	n02087046 (68.12)	n02093428 (58.33)	n02104029 (58.00)	n02089973 (64.10)
5	n04429376 (1.94e-1)	n02089867 (68.06)	n02087046 (56.52)	n02106550 (56.52)	n02087046 (59.66)
6	n01978455 (1.93e-1)	n02109047 (66.67)	n03642806 (54.76)	n02099849 (56.00)	n02106550 (55.07)
7	n02093428 (1.84e-1)	n02106550 (65.22)	n04493381 (50.00)	n02087046 (54.35)	n03642806 (52.91)
8	n04099969 (1.83e-1)	n03642806 (63.49)	n02099849 (48.00)	n02089973 (53.85)	n02099849 (52.44)
9	n02085620 (1.78e-1)	n01978455 (55.07)	n02104029 (48.00)	n04493381 (45.65)	n02089867 (51.16)
10	n03594734 (1.77e-1)	n02488291 (54.17)	n02089867 (47.92)	n02107142 (42.50)	n02104029 (49.11)
11	n03891251 (1.76e-1)	n02099849 (53.33)	n03777754 (47.62)	n03062245 (42.00)	n01978455 (45.89)
12	n03584829 (1.73e-1)	n04429376 (52.17)	n02114855 (45.45)	n04429376 (41.30)	n03777754 (45.77)
13	n03017168 (1.72e-1)	n03777754 (49.21)	n03764736 (45.24)	n03764736 (40.48)	n02107142 (44.44)
14	n02087046 (1.62e-1)	n02107142 (48.33)	n02106550 (43.48)	n03642806 (40.48)	n02488291 (44.44)
15	n04229816 (1.60e-1)	n04336792 (48.33)	n01978455 (43.48)	n03777754 (40.48)	n04493381 (43.48)
16	n02974003 (1.56e-1)	n03062245 (46.67)	n02107142 (42.50)	n03594734 (39.58)	n02114855 (43.18)
17	n01980166 (1.54e-1)	n02113978 (45.83)	n02488291 (39.58)	n02488291 (39.58)	n04429376 (42.75)
18	n02877765 (1.54e-1)	n02114855 (45.45)	n03594734 (39.58)	n04099969 (39.58)	n03764736 (42.33)
19	n04336792 (1.51e-1)	n04099969 (44.44)	n02113978 (37.50)	n02804414 (39.13)	n02113978 (40.28)
20	n03642806 (1.49e-1)	n03930630 (44.44)	n02804414 (36.96)	n01978455 (39.13)	n04336792 (40.28)
21	n02114855 (1.48e-1)	n02090622 (44.00)	n02091831 (36.00)	n02869837 (38.64)	n03062245 (40.22)
22	n03903868 (1.46e-1)	n01980166 (41.67)	n03494278 (36.00)	n02114855 (38.64)	n03594734 (39.81)
23	n01729322 (1.44e-1)	n04026417 (41.67)	n03017168 (35.42)	n04238763 (38.64)	n04099969 (39.81)
24	n04067472 (1.43e-1)	n02091831 (41.33)	n04099969 (35.42)	n02113978 (37.50)	n02869837 (37.88)
25	n02488291 (1.34e-1)	n02104029 (41.33)	n01980166 (35.00)	n02089867 (37.50)	n04026417 (37.50)

Rank	SOIV-U	FGSM	PGD	C&W	MIXED
26	n03032252 (1.25e-1)	n03764736 (41.27)	n04127249 (35.00)	n04026417 (37.50)	n02804414 (37.44)
27	n04493381 (1.18e-1)	n02869837 (40.91)	n04336792 (35.00)	n04336792 (37.50)	n01980166 (37.22)
28	n04238763 (1.16e-1)	n03594734 (40.28)	n04429376 (34.78)	n04517823 (36.96)	n02091831 (37.11)
29	n03637318 (1.16e-1)	n03903868 (39.39)	n02869837 (34.09)	n02113799 (36.00)	n04517823 (35.75)
30	n02104029 (1.16e-1)	n04517823 (37.68)	n07836838 (34.00)	n01980166 (35.00)	n03494278 (35.11)
31	n03085013 (1.16e-1)	n03494278 (37.33)	n04026417 (33.33)	n04127249 (35.00)	n04127249 (34.44)
32	n02869837 (1.13e-1)	n02804414 (36.23)	n04517823 (32.61)	n02091831 (34.00)	n07836838 (33.56)
33	n04136333 (1.13e-1)	n01729322 (34.92)	n03062245 (32.00)	n02859443 (33.33)	n02090622 (33.33)
34	n03777754 (1.12e-1)	n04493381 (34.78)	n02788148 (31.25)	n07836838 (32.00)	n03930630 (32.87)
35	n02109047 (9.76e-2)	n07836838 (34.67)	n03032252 (30.95)	n03494278 (32.00)	n03903868 (31.31)
36	n03259280 (9.74e-2)	n02108089 (33.33)	n02090622 (28.00)	n03891251 (31.82)	n04238763 (31.31)
37	n03930630 (9.25e-2)	n03837869 (33.33)	n03794056 (28.00)	n02788148 (31.25)	n03017168 (30.56)
38	n02086240 (8.78e-2)	n04127249 (33.33)	n04229816 (28.00)	n03017168 (31.25)	n03032252 (30.42)
39	n02788148 (8.73e-2)	n03032252 (31.75)	n03903868 (27.27)	n01729322 (30.95)	n02788148 (30.09)
40	n04026417 (8.71e-2)	n04136333 (31.75)	n02086240 (27.08)	n04435653 (30.43)	n01729322 (29.89)
41	n01749939 (8.60e-2)	n03259280 (30.56)	n03787032 (27.08)	n03930630 (29.17)	n02859443 (28.70)
42	n03775546 (8.42e-2)	n07714571 (30.56)	n03637318 (26.09)	n03032252 (28.57)	n04229816 (28.44)
43	n02108089 (8.26e-2)	n04238763 (30.30)	n07714571 (25.00)	n04136333 (28.57)	n03794056 (27.78)
44	n03530642 (7.92e-2)	n03794056 (29.33)	n03530642 (25.00)	n02090622 (28.00)	n02086240 (27.31)
45	n07753275 (7.78e-2)	n04229816 (29.33)	n04238763 (25.00)	n04229816 (28.00)	n03891251 (27.27)
46	n02099849 (7.33e-2)	n02086240 (27.78)	n04485082 (25.00)	n03903868 (27.27)	n04136333 (27.25)
47	n03062245 (7.28e-2)	n02483362 (27.78)	n04589890 (25.00)	n02086240 (27.08)	n04435653 (26.57)
48	n02086910 (7.16e-2)	n02788148 (27.78)	n02859443 (25.00)	n03085013 (26.09)	n02113799 (26.44)
49	n02119022 (6.88e-2)	n02859443 (27.78)	n03930630 (25.00)	n01692333 (26.00)	n02108089 (26.26)
50	n04435653 (6.79e-2)	n03085013 (27.54)	n04111531 (25.00)	n03794056 (26.00)	n03085013 (25.12)
51	n04517823 (6.67e-2)	n04435653 (27.54)	n01692333 (24.00)	n02138441 (25.00)	n03837869 (25.00)
52	n07836838 (5.97e-2)	n03891251 (27.27)	n01729322 (23.81)	n03424325 (22.92)	n07714571 (24.77)
53	n03794056 (5.95e-2)	n04485082 (26.39)	n03424325 (22.92)	n02105505 (22.92)	n02483362 (24.54)
54	n07831146 (5.78e-2)	n02974003 (25.76)	n02483362 (22.92)	n02483362 (22.92)	n04589890 (24.49)
55	n02804414 (5.67e-2)	n04589890 (25.76)	n03584829 (22.92)	n02100583 (22.92)	n01692333 (24.22)
56	n02089867 (5.47e-2)	n01735189 (25.33)	n02100583 (22.92)	n02108089 (22.73)	n04485082 (24.07)
57	n03492542 (5.03e-2)	n03017168 (25.00)	n02108089 (22.73)	n04589890 (22.73)	n03637318 (23.67)
58	n01692333 (4.94e-2)	n03492542 (24.24)	n03891251 (22.73)	n03530642 (22.50)	n03530642 (23.61)
59	n03424325 (4.81e-2)	n01855672 (24.00)	n02113799 (22.00)	n01855672 (22.00)	n03259280 (23.38)
60	n02106550 (4.47e-2)	n02086910 (24.00)	n03085013 (21.74)	n03637318 (21.74)	n02100583 (23.15)
61	n03785016 (4.45e-2)	n04418357 (24.00)	n04435653 (21.74)	n02123045 (20.83)	n03787032 (22.92)
62	n02483362 (4.32e-2)	n02100583 (23.61)	n04136333 (21.43)	n03259280 (20.83)	n03424325 (22.69)
63	n04485082 (4.22e-2)	n03530642 (23.33)	n13040303 (20.83)	n03584829 (20.83)	n02974003 (22.22)
64	n01983481 (4.12e-2)	n03637318 (23.19)	n02123045 (20.83)	n03787032 (20.83)	n01855672 (21.33)
65	n02107142 (4.06e-2)	n01692333 (22.67)	n03837869 (20.83)	n03837869 (20.83)	n03584829 (20.60)
66	n03787032 (3.57e-2)	n03424325 (22.22)	n02974003 (20.45)	n04485082 (20.83)	n02105505 (20.37)
67	n03837869 (3.18e-2)	n02877765 (21.74)	n04418357 (20.00)	n02974003 (20.45)	n02138441 (20.14)
68	n02396427 (3.11e-2)	n04067472 (21.67)	n03259280 (18.75)	n07714571 (18.75)	n01735189 (19.78)
69	n02123045 (2.95e-2)	n02113799 (21.33)	n02105505 (18.75)	n03492542 (18.18)	n03492542 (19.44)
70	n02113978 (2.86e-2)	n03787032 (20.83)	n02138441 (18.75)	n01735189 (18.00)	n02123045 (18.98)
71	n02105505 (2.76e-2)	n02105505 (19.44)	n01855672 (18.00)	n02259212 (18.00)	n02086910 (18.67)
72	n04592741 (2.73e-2)	n03584829 (18.06)	n02259212 (18.00)	n04067472 (17.50)	n04418357 (18.67)
73	n13040303 (2.46e-2)	n04111531 (18.06)	n02119022 (16.67)	n02877765 (17.39)	n04111531 (18.52)
74	n04111531 (2.39e-2)	n02231487 (18.06)	n01735189 (16.00)	n02119022 (16.67)	n04067472 (18.06)
75	n03494278 (2.07e-2)	n07831146 (16.67)	n02086910 (16.00)	n02086910 (16.00)	n02259212 (16.89)
76	n04418357 (2.05e-2)	n02138441 (16.67)	n03492542 (15.91)	n03947888 (14.58)	n02119022 (16.67)
77	n03947888 (1.90e-2)	n02119022 (16.67)	n04067472 (15.00)	n01749939 (14.00)	n02877765 (15.94)
78	n02091831 (1.88e-2)	n01749939 (16.00)	n02172182 (14.58)	n07831146 (12.50)	n01749939 (14.67)
79	n02231487 (1.77e-2)	n02123045 (15.28)	n01749939 (14.00)	n04111531 (12.50)	n13040303 (14.58)
80	n02701002 (1.61e-2)	n02259212 (14.67)	n02231487 (12.50)	n04418357 (12.00)	n02231487 (13.66)
81	n01773797 (1.08e-2)	n02172182 (13.89)	n13037406 (12.00)	n04592741 (12.00)	n02172182 (12.96)
82	n02172182 (1.02e-2)	n03947888 (13.89)	n04592741 (12.00)	n02172182 (10.42)	n03947888 (12.96)
83	n02100583 (8.29e-3)	n03785016 (13.89)	n03947888 (10.42)	n13040303 (10.42)	n07831146 (12.50)
84	n07714571 (7.61e-3)	n04592741 (13.33)	n01983481 (8.70)	n02231487 (10.42)	n04592741 (12.44)
85	n01855672 (7.00e-3)	n13040303 (12.50)	n02877765 (8.70)	n02396427 (10.00)	n13037406 (9.78)
86	n02326432 (4.30e-3)	n03775546 (11.59)	n07831146 (8.33)	n02018207 (9.09)	n02396427 (9.11)
87	n02090622 (3.84e-3)	n07715103 (11.11)	n07715103 (8.33)	n01983481 (8.70)	n03775546 (8.94)
88	n02859443 (3.84e-3)	n02018207 (10.61)	n02009229 (8.00)	n03379051 (8.70)	n03785016 (8.80)
89	n02113799 (1.52e-3)	n13037406 (9.33)	n02396427 (8.00)	n03775546 (8.70)	n01983481 (8.70)
90	n02018207 (9.69e-4)	n02009229 (9.33)	n03379051 (6.52)	n03785016 (8.33)	n07715103 (8.56)
91	n07715103 (9.43e-4)	n02396427 (9.33)	n03775546 (6.52)	n13037406 (8.00)	n02018207 (8.08)
92	n01558993 (8.01e-4)	n01983481 (8.70)	n07753275 (6.25)	n07715103 (6.25)	n02009229 (7.78)
93	n02138441 (7.25e-4)	n07753275 (5.56)	n02116738 (6.00)	n02009229 (6.00)	n03379051 (5.56)
94	n01735189 (5.24e-4)	n02116738 (4.00)	n02018207 (4.55)	n07753275 (4.17)	n07753275 (5.32)
95	n02009229 (1.66e-4)	n02326432 (2.67)	n01773797 (4.17)	n02116738 (4.00)	n02116738 (4.67)
96	n13037406 (1.59e-4)	n03379051 (1.45)	n03785016 (4.17)	n02326432 (4.00)	n02326432 (2.89)
97	n02259212 (1.40e-4)	n01820546 (1.33)	n01558993 (4.00)	n01773797 (2.08)	n01773797 (2.08)
98	n03379051 (1.37e-4)	n02701002 (0.00)	n02701002 (2.00)	n01820546 (2.00)	n01558993 (2.00)
99	n02116738 (1.33e-4)	n01773797 (0.00)	n02326432 (2.00)	n01558993 (2.00)	n01820546 (1.78)
100	n01820546 (1.00e-4)	n01558993 (0.00)	n01820546 (2.00)	n02701002 (0.00)	n02701002 (0.67)
	Pearson	0.5687 (6.7e-10)	0.5708 (5.6e-10)	0.5481 (3.6e-09)	0.5773 (3.2e-10)
	Spearman	0.6017 (3.6e-11)	0.5919 (8.8e-11)	0.6016 (3.6e-11)	0.6145 (1.0e-11)
	Kendall	0.4290 (2.5e-10)	0.4195 (6.2e-10)	0.4294 (2.5e-10)	0.4354 (1.4e-10)

Table 24: Comparison between SOIV-T ranking and ASR rankings under multiple targeted attacks (I-FGSM, PGD, C&W, and MIXED) when the source class is n03642806 (laptop) for **ImageNet-100 + ViT-L/16**. Classes are identified by their WordNet synset IDs. Each attack is evaluated at multiple attack strengths and the resulting ASRs are averaged.

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
1	n03777754 (-0.038455)	n03085013 (93.65)	n03085013 (88.57)	n03085013 (89.29)	n03085013 (90.50)
2	n03085013 (-0.028934)	n03777754 (93.65)	n03777754 (88.57)	n03777754 (85.71)	n03777754 (89.31)
3	n03775546 (-0.012650)	n03492542 (75.40)	n04238763 (69.52)	n03492542 (63.10)	n03492542 (68.39)
4	n02974003 (-0.005525)	n04238763 (71.43)	n03492542 (66.67)	n02974003 (58.33)	n04238763 (63.25)
5	n03492542 (-0.002092)	n03062245 (68.25)	n03775546 (66.67)	n03794056 (55.95)	n03794056 (61.14)
6	n03032252 (-0.002051)	n03494278 (63.49)	n03062245 (65.71)	n04589890 (55.95)	n04589890 (59.66)
7	n03494278 (-0.001933)	n04485082 (63.49)	n03794056 (64.76)	n03584829 (51.19)	n04485082 (57.91)
8	n03584829 (-0.001402)	n03794056 (62.70)	n04485082 (63.81)	n04517823 (50.00)	n03584829 (57.75)
9	n04111531 (-0.001288)	n03584829 (61.11)	n04517823 (62.86)	n03775546 (48.81)	n04517823 (57.72)
10	n04589890 (-0.000898)	n04127249 (61.11)	n07836838 (61.90)	n04238763 (48.81)	n03062245 (56.96)
11	n03637318 (-0.000873)	n04589890 (61.11)	n04589890 (61.90)	n03787032 (47.62)	n03775546 (56.75)
12	n03787032 (-0.000848)	n04517823 (60.32)	n04592741 (61.90)	n04485082 (46.43)	n03787032 (55.50)
13	n02018207 (-0.000840)	n03787032 (57.94)	n03494278 (60.95)	n01983481 (46.43)	n04336792 (53.20)
14	n04517823 (-0.000835)	n02804414 (56.35)	n03584829 (60.95)	n01773797 (45.24)	n03494278 (52.99)
15	n03794056 (-0.000834)	n04336792 (55.56)	n03787032 (60.95)	n04136333 (45.24)	n01773797 (52.01)
16	n04485082 (-0.000813)	n02085620 (55.56)	n01773797 (60.00)	n04493381 (45.24)	n02804414 (51.88)
17	n07831146 (-0.000749)	n03775546 (54.76)	n04336792 (60.00)	n02804414 (44.05)	n02869837 (50.98)
18	n04136333 (-0.000729)	n04111531 (54.76)	n04418357 (59.05)	n07715103 (44.05)	n02974003 (50.87)
19	n03379051 (-0.000723)	n02869837 (53.17)	n02869837 (58.10)	n01749939 (44.05)	n04592741 (50.53)
20	n02105505 (-0.000722)	n07836838 (53.17)	n04111531 (57.14)	n04336792 (44.05)	n07836838 (49.87)
21	n04238763 (-0.000713)	n03637318 (51.59)	n04493381 (56.19)	n02104029 (42.86)	n04493381 (49.42)
22	n03017168 (-0.000709)	n03032252 (50.79)	n03424325 (55.24)	n02869837 (41.67)	n02085620 (48.44)
23	n01749939 (-0.000705)	n01773797 (50.79)	n02804414 (55.24)	n04592741 (40.48)	n01983481 (47.91)
24	n02488291 (-0.000703)	n04229816 (50.00)	n03032252 (54.29)	n02085620 (39.29)	n04136333 (47.14)
25	n01558993 (-0.000703)	n02086910 (49.21)	n03947888 (54.29)	n02109047 (36.90)	n04229816 (47.06)
26	n04493381 (-0.000694)	n04592741 (49.21)	n04127249 (54.29)	n02877765 (36.90)	n03032252 (46.93)
27	n02085620 (-0.000692)	n03424325 (48.41)	n04229816 (54.29)	n03062245 (36.90)	n03424325 (46.85)
28	n03930630 (-0.000684)	n03764736 (48.41)	n02086910 (53.33)	n03424325 (36.90)	n03637318 (46.48)
29	n02086910 (-0.000681)	n04026417 (48.41)	n03637318 (53.33)	n04229816 (36.90)	n04111531 (44.84)
30	n04336792 (-0.000681)	n02974003 (47.62)	n03594734 (52.38)	n02113978 (35.71)	n04127249 (44.81)
31	n04592741 (-0.000678)	n03594734 (47.62)	n03017168 (51.43)	n07831146 (35.71)	n04418357 (44.55)
32	n02109047 (-0.000678)	n04136333 (47.62)	n01983481 (50.48)	n03032252 (35.71)	n02086910 (44.10)
33	n02869837 (-0.000678)	n01983481 (46.83)	n02085620 (50.48)	n07714571 (35.71)	n01749939 (43.62)
34	n07836838 (-0.000674)	n04493381 (46.83)	n02788148 (50.48)	n07753275 (35.71)	n03764736 (43.28)
35	n02804414 (-0.000670)	n02231487 (46.83)	n02009229 (50.48)	n02106550 (34.52)	n03379051 (43.04)
36	n04026417 (-0.000663)	n03017168 (46.83)	n03764736 (50.48)	n07836838 (34.52)	n02104029 (42.91)
37	n03903868 (-0.000653)	n03379051 (46.03)	n03837869 (49.52)	n03494278 (34.52)	n03594734 (42.86)
38	n02119022 (-0.000652)	n04099969 (46.03)	n02018207 (48.57)	n03637318 (34.52)	n02109047 (42.78)
39	n04435653 (-0.000646)	n04418357 (46.03)	n02109047 (48.57)	n03379051 (34.52)	n07831146 (42.65)
40	n07715103 (-0.000643)	n02009229 (45.24)	n07831146 (48.57)	n03930630 (33.33)	n02877765 (41.88)
41	n04229816 (-0.000636)	n04067472 (45.24)	n03379051 (48.57)	n02018207 (32.14)	n04026417 (41.85)
42	n03424325 (-0.000632)	n02100583 (45.24)	n04026417 (48.57)	n01729322 (30.95)	n07714571 (41.75)
43	n01983481 (-0.000630)	n02018207 (44.44)	n04136333 (48.57)	n03764736 (30.95)	n02018207 (41.72)
44	n01773797 (-0.000630)	n03259280 (43.65)	n04067472 (47.62)	n02086910 (29.76)	n07715103 (41.61)
45	n02089973 (-0.000628)	n07831146 (43.65)	n02090622 (46.67)	n02089867 (28.57)	n03017168 (40.29)
46	n02396427 (-0.000622)	n03947888 (43.65)	n02877765 (46.67)	n03594734 (28.57)	n02009229 (40.24)
47	n04418357 (-0.000605)	n02107142 (43.65)	n02974003 (46.67)	n04026417 (28.57)	n02113978 (40.16)
48	n02483362 (-0.000597)	n07714571 (42.86)	n07714571 (46.67)	n04418357 (28.57)	n02231487 (39.97)
49	n02009229 (-0.000595)	n02109047 (42.86)	n04099969 (46.67)	n13037406 (27.38)	n04067472 (39.68)
50	n02113978 (-0.000594)	n02113978 (42.86)	n02231487 (45.71)	n02105505 (27.38)	n02106550 (39.60)
51	n03764736 (-0.000589)	n02326432 (42.86)	n02107142 (44.76)	n02231487 (27.38)	n07753275 (39.15)
52	n03891251 (-0.000589)	n01749939 (42.06)	n03259280 (44.76)	n04067472 (26.19)	n03930630 (39.10)
53	n01855672 (-0.000587)	n02104029 (42.06)	n01749939 (44.76)	n03530642 (26.19)	n02788148 (38.92)
54	n02106550 (-0.000585)	n02877765 (42.06)	n02100583 (44.76)	n02119022 (26.19)	n02090622 (36.72)
55	n03530642 (-0.000583)	n03930630 (42.06)	n02106550 (43.81)	n02483362 (25.00)	n02107142 (36.61)
56	n13037406 (-0.000579)	n02788148 (41.27)	n02091831 (43.81)	n02089973 (25.00)	n03530642 (35.98)
57	n02259212 (-0.000575)	n02089973 (40.48)	n02104029 (43.81)	n02091831 (25.00)	n03947888 (35.82)
58	n02090622 (-0.000570)	n02106550 (40.48)	n02086240 (42.86)	n02788148 (25.00)	n03259280 (35.82)
59	n02788148 (-0.000568)	n02701002 (40.48)	n01980166 (42.86)	n01980166 (25.00)	n03837869 (35.69)
60	n07753275 (-0.000566)	n02087046 (40.48)	n03530642 (42.86)	n02009229 (25.00)	n04099969 (35.66)
61	n02100583 (-0.000566)	n02090622 (39.68)	n02138441 (42.86)	n02090622 (23.81)	n02100583 (35.16)
62	n02877765 (-0.000563)	n03837869 (39.68)	n02172182 (42.86)	n03903868 (23.81)	n02701002 (35.00)
63	n01820546 (-0.000561)	n07715103 (38.89)	n03903868 (42.86)	n04435653 (23.81)	n02089867 (34.66)
64	n04099969 (-0.000559)	n02086240 (38.89)	n07753275 (42.86)	n02396427 (23.81)	n03903868 (34.66)
65	n01692333 (-0.000557)	n02108089 (38.89)	n02113978 (41.90)	n02701002 (22.62)	n01980166 (34.52)
66	n02114855 (-0.000549)	n02172182 (38.89)	n02701002 (41.90)	n02108089 (22.62)	n02108089 (34.47)
67	n01978455 (-0.000547)	n03530642 (38.89)	n02108089 (41.90)	n03017168 (22.62)	n02172182 (34.39)
68	n01729322 (-0.000546)	n07753275 (38.89)	n07715103 (41.90)	n02488291 (22.62)	n02119022 (34.02)
69	n04067472 (-0.000545)	n01735189 (38.10)	n03930630 (41.90)	n02259212 (22.62)	n02086240 (33.99)
70	n03259280 (-0.000544)	n02089867 (37.30)	n02119022 (40.95)	n04111531 (22.62)	n01729322 (33.54)
71	n02086240 (-0.000539)	n03903868 (37.30)	n02123045 (40.95)	n02093428 (21.43)	n04435653 (33.17)

Rank	SOIV-T	I-FGSM	PGD	C&W	MIXED
72	n02107142 (-0.000538)	n02093428 (36.51)	n02093428 (40.00)	n02107142 (21.43)	n13037406 (33.15)
73	n13040303 (-0.000536)	n03785016 (36.51)	n02326432 (40.00)	n02859443 (21.43)	n02091831 (32.99)
74	n02231487 (-0.000535)	n02259212 (35.71)	n02488291 (40.00)	n02172182 (21.43)	n02087046 (32.78)
75	n02093428 (-0.000529)	n02123045 (35.71)	n01558993 (40.00)	n02086240 (20.24)	n02093428 (32.65)
76	n02701002 (-0.000528)	n04435653 (35.71)	n04435653 (40.00)	n01558993 (20.24)	n02089973 (32.62)
77	n02104029 (-0.000527)	n02099849 (35.71)	n02087046 (40.00)	n03259280 (19.05)	n02326432 (32.38)
78	n02138441 (-0.000523)	n01980166 (35.71)	n02099849 (38.10)	n04127249 (19.05)	n02105505 (31.61)
79	n01980166 (-0.000523)	n01855672 (34.92)	n04429376 (38.10)	n03837869 (17.86)	n02259212 (31.19)
80	n02113799 (-0.000522)	n13037406 (34.92)	n02089867 (38.10)	n03891251 (17.86)	n02483362 (30.56)
81	n02859443 (-0.000521)	n02138441 (34.92)	n02105505 (38.10)	n02087046 (17.86)	n02488291 (30.40)
82	n02172182 (-0.000519)	n02119022 (34.92)	n01820546 (38.10)	n02113799 (17.86)	n01558993 (30.40)
83	n02091831 (-0.000518)	n03891251 (34.13)	n13037406 (37.14)	n02116738 (16.67)	n03891251 (29.71)
84	n04429376 (-0.000517)	n01820546 (33.33)	n01729322 (37.14)	n02114855 (16.67)	n01820546 (28.57)
85	n03062245 (-0.000507)	n01978455 (33.33)	n03891251 (37.14)	n01978455 (15.48)	n02396427 (28.52)
86	n02108089 (-0.000499)	n02483362 (33.33)	n01735189 (35.24)	n13040303 (15.48)	n01735189 (28.41)
87	n02087046 (-0.000495)	n01729322 (32.54)	n01855672 (35.24)	n02100583 (15.48)	n02859443 (28.36)
88	n01735189 (-0.000487)	n02113799 (31.75)	n02259212 (35.24)	n01820546 (14.29)	n02123045 (28.33)
89	n07714571 (-0.000486)	n01558993 (30.95)	n03785016 (35.24)	n02326432 (14.29)	n02138441 (27.91)
90	n03594734 (-0.000483)	n02091831 (30.16)	n02859443 (34.29)	n04099969 (14.29)	n01855672 (27.75)
91	n02099849 (-0.000476)	n02114855 (30.16)	n01978455 (34.29)	n01855672 (13.10)	n01978455 (27.70)
92	n02326432 (-0.000470)	n02859443 (29.37)	n02483362 (33.33)	n04429376 (13.10)	n03785016 (27.09)
93	n02116738 (-0.000469)	n02105505 (29.37)	n02089973 (32.38)	n01735189 (11.90)	n02099849 (26.98)
94	n03837869 (-0.000452)	n02396427 (29.37)	n02396427 (32.38)	n01692333 (9.52)	n02113799 (26.69)
95	n03785016 (-0.000446)	n01692333 (28.57)	n01692333 (32.38)	n03785016 (9.52)	n04429376 (26.59)
96	n03947888 (-0.000438)	n02488291 (28.57)	n13040303 (31.43)	n03947888 (9.52)	n02114855 (25.13)
97	n04127249 (-0.000430)	n04429376 (28.57)	n02113799 (30.48)	n02123045 (8.33)	n13040303 (24.63)
98	n02123045 (-0.000413)	n02116738 (28.57)	n02116738 (28.57)	n02099849 (7.14)	n02116738 (24.60)
99	n02089867 (-0.000365)	n13040303 (26.98)	n02114855 (28.57)	n02138441 (5.95)	n01692333 (23.49)
	Pearson	0.5907 (1.2e-10)	0.5594 (1.8e-09)	0.5923 (1.1e-10)	0.6184 (9.1e-12)
	Spearman	0.5482 (4.3e-09)	0.5520 (3.2e-09)	0.6992 (8.4e-16)	0.6519 (2.7e-13)
	Kendall	0.3968 (5.9e-09)	0.3921 (8.9e-09)	0.5221 (1.9e-14)	0.4732 (4.0e-12)

D Further Analysis of Second-Order Statistics

D.1 Additional PGD Evaluation of SOIV-guided Adversarial Training

To complement the AutoAttack results reported in the main paper Section 5.2, we additionally evaluate the same Baseline, AT-Uniform, and AT-SOIV models under PGD. The PGD evaluation is conducted under the L_∞ norm with $\epsilon = 8/255$, a step size of $2/255$, and 7 steps. Thus, this experiment evaluates the same trained models under a different attack setting. AT-Uniform and AT-SOIV are each trained with a single seed.

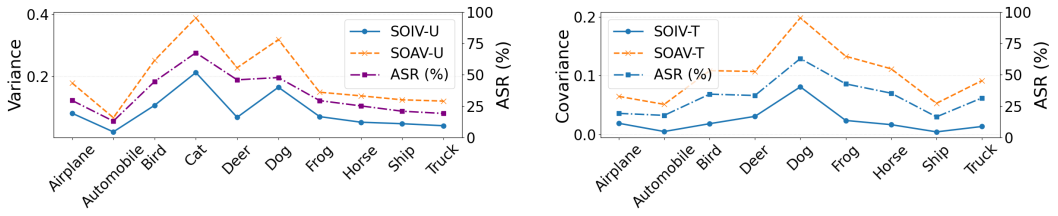
As shown in Table 25, AT-SOIV achieves lower ASR than AT-Uniform in 8 of the 10 classes. In particular, for the high-SOIV-U classes *cat*, *deer*, *dog*, and *bird*, ASR decreases by 7.98, 12.41, 4.79, and 0.78 percentage points, respectively. The overall ASR also decreases from 52.29% for AT-Uniform to 49.94% for AT-SOIV. These results show a similar overall trend to the AutoAttack evaluation reported in the main paper Section 5.2.

D.2 Clean Distribution as a Proxy for Adversarial Vulnerability

As discussed in the main paper Section 3.4, the proposed framework is grounded in the assumption that the structural pattern of adversarial vulnerability, typically measured via ASR, is already reflected in the clean data distribution and can therefore be estimated through SOIV without generating adversarial examples. To examine this assumption, we conduct a supplementary experiment introducing SOAV (Second-Order Adversarial Vulnerability), defined as the class-wise variance (analogous to the main paper Eq. 2) and covariance (analogous to the main paper Eq. 5) of softmax outputs computed on adversarial data $\mathcal{D}_{\text{adv}}$ (entirely disjoint from the clean validation data). SOAV serves as an intermediate reference that directly captures second-order statistics under adversarial conditions, bridging SOIV (clean-based) and ASR (attack-based). If the assumption holds, we expect SOIV, SOAV, and ASR to exhibit consistent class-wise trends. As shown in Fig. 6, the relative class-wise rankings are largely consistent across all three metrics in both untargeted and targeted settings, supporting the core assumption and indicating that clean distribution-based analysis can serve as a practical pre-screening tool for adversarial vulnerability diagnosis.

Table 25: Class-wise ASR under PGD evaluation for the Baseline, AT-Uniform, and AT-SOIV models. PGD is evaluated under the L_∞ norm with $\epsilon = 8/255$, a step size of $2/255$, and 7 steps.

Class	SOIV-U	Baseline	AT-Uniform	AT-SOIV
cat	0.019349	100.00%	71.63%	63.65%
deer	0.012349	100.00%	76.94%	64.53%
dog	0.011844	99.62%	64.27%	59.48%
bird	0.011219	97.90%	54.53%	53.75%
horse	0.008957	98.13%	45.10%	31.15%
airplane	0.007544	99.78%	53.77%	53.02%
frog	0.006985	99.89%	53.47%	52.65%
truck	0.005784	97.94%	44.52%	40.84%
automobile	0.005586	93.21%	49.74%	55.97%
ship	0.003910	99.25%	36.37%	42.24%
Overall ASR	–	98.52%	52.29%	49.94%
Clean Accuracy	–	88.24%	82.43%	85.17%



Untargeted (SQUARE) **Targeted (C&W, Source class: Cat)**

Figure 6: Comparison of SOIV, SOAV, and ASR in untargeted and targeted attack settings for VGG-16 on CIFAR-10. Class-wise trends are largely consistent across all metrics. For visualization in the targeted case, the absolute values of SOIV-T and SOAV-T are used to facilitate direct comparison with ASR.

D.3 Comparison with Existing Attack-free Metrics

To further analyze the characteristics of SOIV-U, we compare it with existing attack-free metrics. These include CLEVER (Weng et al., 2018), an attack-agnostic robustness score based on local Lipschitz constant estimation, as well as uncertainty-related measures computed from clean validation data, namely margin, confidence, entropy, NLL, and the Brier score. Here, margin denotes the difference between the true-class probability and the largest competing probability, confidence denotes the true-class softmax probability, entropy measures the uncertainty of the predictive distribution, NLL denotes the negative log-probability of the true class, and the Brier score measures the squared error between the predictive distribution and the one-hot label, which subsumes the true-class deviation term.

We compute CLEVER with the Adversarial Robustness Toolbox (ART) (Nicolae et al., 2018) using the sampling parameters $N_b = 500$ and $N_s = 1024$. Due to its high computational cost, CLEVER-U is evaluated on a subset of 500 validation images (50 per class). This follows the original work, which likewise evaluates CLEVER on a subset of 500 CIFAR-10 test images (Weng et al., 2018). Since CLEVER produces a robustness score per input, we aggregate the sample-wise scores by true class for CLEVER-U and by source–target pair for CLEVER-T to obtain quantities comparable to ours.

For a fair comparison, all untargeted metrics are computed on the same 500-image subset, and SOIV-T and CLEVER-T are computed on all 500 validation images of the source class. Table 26 summarizes the resulting rankings and correlations in the untargeted setting for CIFAR-10 with ResNet-18. SOIV-U achieves Pearson, Spearman, and Kendall correlations of 0.865, 0.879, and 0.733, respectively, outperforming CLEVER-U (0.842, 0.855, and 0.689) across all coefficients. Compared with

Table 26: Comparison between SOIV-U and existing attack-free metrics under MIXED untargeted ASR for **CIFAR-10 + ResNet-18**. The MIXED ASR is computed by aggregating FGSM, PGD, C&W, SQUARE, and AutoAttack results. Absolute correlation values are used for CLEVER-U, Confidence, and Margin because larger values indicate lower vulnerability.

Rank	MIXED ASR	SOIV-U	CLEVER-U	Margin	Confidence	Entropy	NLL	Brier
1	cat (79.34)	cat (1.96e-02)	cat (2.81e-03)	cat (0.8948)	cat (0.9389)	cat (0.1778)	cat (0.0748)	cat (0.0322)
2	deer (70.41)	frog (1.75e-02)	bird (3.68e-03)	frog (0.9074)	frog (0.9502)	airplane (0.1298)	frog (0.0625)	frog (0.0309)
3	dog (64.23)	deer (1.54e-02)	frog (3.69e-03)	airplane (0.9198)	airplane (0.9552)	frog (0.1256)	bird (0.0540)	deer (0.0254)
4	bird (58.28)	bird (1.44e-02)	dog (3.84e-03)	bird (0.9224)	bird (0.9559)	bird (0.1231)	deer (0.0536)	bird (0.0249)
5	frog (52.62)	dog (1.36e-02)	deer (3.91e-03)	deer (0.9266)	deer (0.9580)	deer (0.1164)	airplane (0.0536)	dog (0.0244)
6	airplane (45.58)	airplane (1.31e-02)	airplane (5.15e-03)	dog (0.9332)	dog (0.9641)	dog (0.0945)	dog (0.0461)	airplane (0.0225)
7	horse (44.14)	truck (6.65e-03)	ship (5.63e-03)	truck (0.9531)	truck (0.9757)	automobile (0.0716)	truck (0.0287)	truck (0.0124)
8	ship (36.58)	horse (4.94e-03)	truck (6.42e-03)	horse (0.9603)	horse (0.9784)	horse (0.0708)	horse (0.0248)	horse (0.0090)
9	truck (35.89)	ship (4.40e-03)	horse (7.65e-03)	ship (0.9664)	ship (0.9812)	truck (0.0678)	ship (0.0216)	ship (0.0075)
10	automobile (31.78)	automobile (3.84e-03)	automobile (8.33e-03)	automobile (0.9673)	automobile (0.9813)	ship (0.0622)	automobile (0.0212)	automobile (0.0060)
Pearson		0.8646 (1.2e-03)	0.8417 (2.3e-03)	0.7954 (5.9e-03)	0.8152 (4.1e-03)	0.7969 (5.8e-03)	0.8361 (2.6e-03)	0.8427 (2.2e-03)
Spearman		0.8788 (8.1e-04)	0.8545 (1.6e-03)	0.7455 (1.3e-02)	0.7455 (1.3e-02)	0.6848 (2.9e-02)	0.8182 (3.8e-03)	0.8788 (8.1e-04)
Kendall		0.7333 (2.2e-03)	0.6889 (4.7e-03)	0.5556 (2.9e-02)	0.5556 (2.9e-02)	0.4222 (1.1e-01)	0.6444 (9.1e-03)	0.7333 (2.2e-03)

the remaining uncertainty-related metrics, SOIV-U achieves the highest or tied-highest correlation. These results indicate that SOIV-U shows consistently strong overall alignment with the observed adversarial vulnerability structure. The efficiency gap is far larger. Computing SOIV-U required 0.117 seconds including clean inference, whereas CLEVER-U required 411.7 minutes (an approximately 2.1×10^5-fold difference), since CLEVER performs repeated gradient evaluations for each input.

Table 27 reports the corresponding comparison in the targeted setting. Since the uncertainty-related metrics are not defined with respect to a target class, CLEVER-T is the only applicable baseline here. SOIV-T achieves Pearson, Spearman, and Kendall correlations of 0.975, 0.733, and 0.611, substantially outperforming CLEVER-T (0.840, 0.400, and 0.278), indicating that the proposed metric effectively captures directional vulnerability in the targeted setting.

Tables 28 and 29 extend the untargeted comparison to the two adversarially trained CIFAR-100 models. CLEVER is omitted here because of its computational cost, which grows with both the number of images and the number of classes. SOIV-U ranks first or second in every coefficient. It attains the highest Pearson correlation, and the largest gap in the rank correlations is 0.004. SOIV-U is therefore among the strongest of the compared metrics in this 100-class, adversarially trained setting.

These results suggest that second-order statistics derived from clean predictive distributions can effectively capture adversarial vulnerability patterns. Furthermore, unlike conventional uncertainty-related metrics that provide only a scalar uncertainty estimate per prediction, the SOIV framework can also be naturally extended to directional source-to-target analysis through SOIV-T. CLEVER-T captures such directionality but requires a separate estimate for each target class. In contrast, SOIV-T explicitly models the interaction between source class and target class probabilities through conditional covariance, recovering all source–target scores from a single set of clean softmax outputs. As a result, the proposed framework can characterize directional vulnerability patterns in targeted attack settings, enabling a more detailed analysis of adversarial behavior beyond class-wise vulnerability estimation.

Table 27: Comparison between SOIV-T and CLEVER-T under MIXED targeted ASR for **CIFAR-10 + ResNet-18** when the source class is *automobile*. The MIXED ASR is computed by aggregating FGSM, PGD, and C&W results. Absolute correlation values are used for both metrics to facilitate direct comparison.

Rank	MIXED ASR	SOIV-T	CLEVER-T
1	truck (62.83)	truck (-0.002972)	truck (0.013524)
2	ship (31.31)	ship (-0.001217)	frog (0.018627)
3	airplane (22.00)	horse (-0.000430)	dog (0.019155)
4	frog (19.94)	airplane (-0.000418)	cat (0.019244)
5	bird (13.58)	frog (-0.000178)	ship (0.019650)
6	cat (12.65)	dog (-0.000129)	horse (0.019694)
7	horse (10.98)	cat (-0.000117)	bird (0.020133)
8	dog (10.84)	bird (-0.000094)	airplane (0.021091)
9	deer (10.13)	deer (-0.000030)	deer (0.022283)
Pearson		0.9751 (7.8e-06)	0.8398 (4.6e-03)
Spearman		0.7333 (2.5e-02)	0.4000 (2.9e-01)
Kendall		0.6111 (2.5e-02)	0.2778 (3.6e-01)

Table 28: Comparison between SOIV-U and uncertainty-related metrics under MIXED untargeted ASR for **CIFAR-100 + Adversarially Trained PreActResNet-18**. The MIXED ASR is computed by aggregating FGSM, PGD, C&W, and SQUARE results. Due to page width constraints, class indices are used instead of class names. Absolute correlation values are used for Margin and Confidence because larger values indicate lower vulnerability.

Rank	MIXED ASR	SOIV-U	Margin	Confidence	Entropy	NLL	Brier
1	80 (88.24)	55 (8.85e-01)	96 (-0.047340)	55 (0.063439)	27 (4.049265)	55 (3.268114)	55 (0.913669)
2	55 (85.00)	27 (8.63e-01)	55 (-0.027877)	27 (0.073503)	80 (4.008991)	27 (3.019690)	27 (0.887361)
3	77 (83.75)	65 (8.54e-01)	10 (-0.022810)	65 (0.079253)	88 (4.008682)	65 (3.007586)	65 (0.883095)
4	4 (79.00)	80 (8.43e-01)	65 (-0.019931)	80 (0.087450)	77 (4.006274)	72 (2.981222)	98 (0.869596)
5	38 (76.92)	72 (8.36e-01)	35 (-0.018710)	98 (0.095274)	38 (4.003911)	45 (2.896978)	80 (0.867885)
6	44 (76.25)	98 (8.34e-01)	98 (-0.016926)	72 (0.098039)	55 (3.998926)	80 (2.887271)	72 (0.864791)
7	74 (75.00)	38 (8.24e-01)	27 (-0.006837)	44 (0.098052)	43 (3.968169)	98 (2.883999)	44 (0.848826)
8	42 (74.00)	3 (8.21e-01)	37 (-0.003762)	88 (0.098170)	42 (3.967852)	3 (2.850220)	3 (0.848823)
9	35 (71.88)	44 (8.20e-01)	72 (0.004377)	38 (0.099419)	65 (3.961259)	4 (2.789923)	35 (0.847471)
10	34 (68.97)	88 (8.19e-01)	44 (0.004883)	3 (0.102914)	44 (3.955306)	38 (2.776946)	88 (0.845178)
11	96 (68.75)	4 (8.14e-01)	80 (0.005718)	35 (0.105028)	79 (3.936160)	74 (2.748516)	38 (0.844581)
12	72 (68.18)	77 (8.12e-01)	30 (0.005978)	77 (0.106584)	64 (3.929165)	64 (2.730769)	4 (0.835449)
13	79 (67.86)	35 (8.08e-01)	3 (0.008180)	18 (0.109125)	74 (3.914433)	18 (2.728605)	96 (0.833130)
14	2 (67.31)	18 (8.04e-01)	92 (0.008457)	78 (0.111943)	78 (3.908839)	84 (2.725014)	77 (0.831226)
15	50 (66.18)	74 (8.03e-01)	78 (0.011515)	74 (0.112998)	18 (3.903355)	88 (2.707985)	18 (0.830641)
16	27 (65.91)	78 (7.97e-01)	88 (0.013832)	31 (0.117792)	31 (3.902697)	44 (2.683464)	78 (0.828357)
17	33 (65.52)	31 (7.96e-01)	46 (0.016935)	33 (0.119406)	33 (3.901645)	77 (2.678576)	74 (0.827317)
18	25 (63.64)	42 (7.87e-01)	2 (0.017005)	42 (0.120518)	34 (3.890946)	50 (2.669202)	10 (0.821595)
19	3 (63.24)	33 (7.83e-01)	18 (0.020887)	37 (0.122174)	3 (3.883579)	31 (2.631642)	31 (0.819948)
20	98 (62.50)	45 (7.80e-01)	74 (0.023831)	4 (0.125924)	72 (3.865530)	42 (2.619467)	45 (0.819231)
21	88 (57.41)	37 (7.77e-01)	33 (0.026610)	96 (0.128129)	4 (3.842789)	93 (2.618598)	37 (0.817428)
22	18 (56.58)	64 (7.77e-01)	38 (0.029858)	2 (0.129036)	19 (3.812962)	78 (2.609753)	33 (0.811010)
23	14 (56.45)	50 (7.76e-01)	7 (0.029967)	79 (0.131537)	66 (3.812229)	35 (2.569087)	2 (0.809468)
24	63 (55.56)	2 (7.74e-01)	31 (0.035391)	64 (0.131636)	98 (3.789781)	19 (2.553222)	42 (0.806630)
25	67 (55.21)	84 (7.73e-01)	83 (0.040300)	10 (0.134120)	36 (3.778949)	2 (2.513631)	50 (0.799947)
26	11 (53.57)	19 (7.67e-01)	77 (0.041188)	43 (0.134661)	50 (3.771835)	22 (2.487041)	64 (0.796418)
27	19 (52.88)	96 (7.66e-01)	45 (0.045942)	19 (0.139380)	35 (3.756805)	37 (2.468114)	84 (0.795840)
28	43 (52.86)	79 (7.65e-01)	93 (0.046378)	50 (0.139873)	2 (3.749850)	46 (2.442053)	93 (0.792683)
29	78 (51.85)	10 (7.63e-01)	73 (0.046962)	93 (0.152553)	89 (3.742357)	10 (2.419359)	19 (0.790990)
30	5 (51.56)	93 (7.58e-01)	59 (0.047864)	34 (0.152645)	97 (3.727168)	11 (2.418108)	79 (0.786229)
31	10 (51.19)	43 (7.57e-01)	67 (0.050454)	46 (0.152864)	37 (3.726434)	33 (2.404770)	46 (0.784916)
32	65 (50.00)	46 (7.43e-01)	11 (0.051579)	45 (0.153070)	15 (3.710654)	79 (2.393688)	43 (0.775527)
33	15 (49.26)	34 (7.37e-01)	47 (0.052515)	84 (0.163791)	81 (3.699076)	43 (2.339956)	11 (0.760638)
34	46 (48.21)	15 (7.17e-01)	4 (0.053020)	15 (0.172521)	6 (3.676655)	15 (2.296345)	34 (0.755275)
35	59 (47.83)	11 (7.17e-01)	19 (0.053922)	81 (0.173934)	84 (3.673809)	73 (2.289592)	7 (0.753539)
36	64 (47.83)	6 (7.03e-01)	42 (0.054838)	6 (0.176248)	51 (3.659895)	25 (2.263602)	92 (0.751466)
37	66 (47.79)	81 (6.98e-01)	50 (0.055099)	11 (0.179908)	93 (3.605604)	96 (2.260328)	15 (0.741676)
38	45 (47.62)	36 (6.88e-01)	79 (0.055359)	36 (0.185304)	45 (3.602932)	32 (2.222597)	73 (0.737388)
39	6 (46.67)	73 (6.86e-01)	64 (0.063380)	89 (0.186783)	10 (3.565498)	36 (2.212205)	30 (0.736585)
40	70 (46.09)	25 (6.82e-01)	43 (0.069381)	92 (0.188259)	25 (3.555836)	34 (2.208009)	6 (0.734089)
41	90 (45.97)	92 (6.81e-01)	81 (0.072574)	73 (0.191258)	21 (3.553747)	63 (2.204978)	83 (0.727565)
42	49 (45.83)	66 (6.78e-01)	70 (0.075262)	67 (0.192175)	46 (3.542164)	29 (2.203711)	81 (0.726245)
43	30 (43.94)	89 (6.77e-01)	6 (0.078670)	66 (0.195746)	63 (3.536972)	6 (2.199840)	67 (0.725084)
44	36 (43.94)	7 (6.75e-01)	84 (0.084433)	25 (0.196000)	90 (3.531525)	66 (2.145420)	59 (0.722420)
45	7 (43.75)	67 (6.75e-01)	15 (0.084490)	7 (0.196903)	13 (3.492639)	26 (2.140950)	25 (0.710524)
46	97 (43.55)	63 (6.73e-01)	13 (0.086008)	30 (0.198075)	5 (3.490953)	14 (2.122410)	22 (0.708979)
47	84 (43.10)	13 (6.70e-01)	34 (0.086129)	13 (0.205145)	11 (3.477466)	83 (2.110535)	36 (0.706631)
48	89 (41.41)	30 (6.63e-01)	25 (0.095247)	51 (0.207528)	14 (3.448603)	59 (2.081988)	13 (0.701700)
49	85 (41.22)	83 (6.63e-01)	22 (0.100865)	59 (0.211521)	96 (3.443939)	51 (2.081426)	89 (0.698146)
50	8 (40.44)	59 (6.63e-01)	89 (0.107022)	83 (0.214370)	26 (3.417144)	92 (2.076894)	47 (0.694368)
51	93 (40.28)	51 (6.62e-01)	36 (0.116948)	97 (0.216696)	85 (3.408585)	67 (2.062191)	63 (0.692090)
52	31 (40.18)	22 (6.54e-01)	90 (0.122696)	90 (0.221469)	8 (3.378717)	40 (2.052231)	66 (0.691959)
53	81 (40.15)	97 (6.54e-01)	49 (0.130884)	47 (0.224406)	87 (3.336462)	81 (2.035138)	51 (0.681129)
54	37 (40.00)	26 (6.52e-01)	5 (0.131244)	63 (0.225155)	54 (3.326638)	13 (2.014033)	26 (0.676329)
55	92 (39.77)	14 (6.43e-01)	51 (0.132915)	5 (0.232386)	7 (3.301136)	97 (2.012096)	70 (0.672624)
56	87 (37.50)	32 (6.38e-01)	66 (0.144214)	70 (0.233381)	67 (3.295818)	7 (1.998871)	97 (0.668387)
57	32 (37.07)	47 (6.34e-01)	63 (0.157231)	26 (0.238444)	16 (3.290009)	8 (1.948515)	14 (0.665347)
58	13 (36.36)	90 (6.32e-01)	21 (0.157384)	22 (0.241989)	17 (3.286870)	89 (1.940955)	32 (0.661881)
59	26 (35.87)	5 (6.23e-01)	26 (0.157976)	21 (0.242279)	1 (3.281491)	30 (1.889003)	90 (0.657629)
60	51 (35.81)	70 (6.18e-01)	62 (0.159208)	14 (0.242559)	32 (3.276536)	16 (1.862400)	5 (0.650178)
61	17 (35.26)	8 (6.14e-01)	61 (0.162046)	49 (0.260590)	73 (3.246379)	5 (1.841542)	8 (0.633199)
62	40 (35.00)	29 (6.09e-01)	97 (0.162744)	85 (0.260661)	92 (3.238973)	70 (1.838855)	16 (0.630528)
63	83 (35.00)	16 (6.01e-01)	14 (0.163062)	32 (0.264234)	56 (3.229499)	47 (1.834451)	29 (0.628834)
64	71 (32.89)	85 (5.96e-01)	16 (0.164351)	16 (0.266588)	29 (3.219214)	90 (1.828033)	21 (0.618240)
65	56 (32.74)	21 (5.94e-01)	54 (0.173052)	54 (0.272680)	23 (3.212216)	28 (1.827721)	85 (0.617764)
66	57 (32.26)	49 (5.77e-01)	17 (0.175096)	8 (0.275361)	59 (3.187339)	85 (1.816259)	49 (0.616280)
67	91 (32.14)	40 (5.74e-01)	85 (0.176070)	17 (0.280070)	47 (3.186884)	12 (1.747147)	40 (0.613082)
68	21 (31.67)	54 (5.72e-01)	32 (0.181995)	29 (0.287497)	49 (3.174138)	57 (1.706726)	54 (0.598615)
69	28 (31.25)	87 (5.50e-01)	52 (0.187346)	87 (0.295719)	30 (3.154219)	54 (1.698834)	62 (0.578243)

Rank	MIXED ASR	SOIV-U	Margin	Confidence	Entropy	NLL	Brier
70	95 (31.06)	12 (5.44e-01)	40 (0.188980)	23 (0.297027)	12 (3.139542)	49 (1.687334)	12 (0.573688)
71	1 (30.41)	17 (5.38e-01)	23 (0.196730)	61 (0.298742)	99 (3.137752)	99 (1.673305)	61 (0.570780)
72	52 (29.88)	57 (5.36e-01)	12 (0.203903)	12 (0.300628)	22 (3.135215)	21 (1.649974)	87 (0.569537)
73	61 (29.61)	61 (5.32e-01)	8 (0.208990)	40 (0.311680)	70 (3.126609)	87 (1.636785)	17 (0.566548)
74	73 (29.46)	99 (5.30e-01)	29 (0.216933)	62 (0.313082)	83 (3.101433)	61 (1.623403)	57 (0.559995)
75	22 (28.91)	62 (5.27e-01)	87 (0.217122)	1 (0.314037)	61 (3.053190)	9 (1.612183)	99 (0.553795)
76	12 (28.68)	23 (5.27e-01)	99 (0.224214)	56 (0.315619)	57 (3.027529)	62 (1.565659)	23 (0.553379)
77	99 (28.47)	56 (5.16e-01)	56 (0.234572)	99 (0.316769)	48 (3.013283)	23 (1.560039)	56 (0.537881)
78	69 (27.91)	1 (5.13e-01)	71 (0.236322)	57 (0.333740)	28 (2.994811)	56 (1.524480)	28 (0.535596)
79	54 (27.38)	28 (5.10e-01)	1 (0.237578)	28 (0.341485)	75 (2.978152)	86 (1.519956)	1 (0.533355)
80	23 (25.69)	9 (4.84e-01)	28 (0.246760)	52 (0.342646)	40 (2.966252)	17 (1.483199)	52 (0.507821)
81	47 (25.00)	48 (4.83e-01)	57 (0.248772)	69 (0.360123)	58 (2.959387)	39 (1.442628)	9 (0.506743)
82	58 (24.32)	52 (4.62e-01)	68 (0.252549)	71 (0.360856)	69 (2.954665)	1 (1.437899)	48 (0.497041)
83	39 (23.17)	39 (4.61e-01)	69 (0.253657)	68 (0.366237)	39 (2.903665)	69 (1.425565)	69 (0.491965)
84	48 (22.78)	69 (4.59e-01)	95 (0.268737)	48 (0.377121)	94 (2.902243)	48 (1.416328)	71 (0.484389)
85	16 (22.32)	71 (4.50e-01)	58 (0.309910)	9 (0.387369)	68 (2.902109)	91 (1.291844)	39 (0.478770)
86	9 (21.53)	86 (4.38e-01)	9 (0.310433)	58 (0.391301)	71 (2.847773)	52 (1.289225)	68 (0.468604)
87	41 (20.51)	58 (4.33e-01)	48 (0.316618)	95 (0.394392)	9 (2.837624)	95 (1.244793)	95 (0.468354)
88	75 (19.27)	68 (4.32e-01)	60 (0.331271)	39 (0.395152)	62 (2.825598)	71 (1.233262)	86 (0.460909)
89	68 (19.19)	95 (4.31e-01)	39 (0.335848)	75 (0.406927)	52 (2.763316)	58 (1.221274)	58 (0.450781)
90	94 (18.75)	75 (4.13e-01)	75 (0.351194)	86 (0.433912)	95 (2.660295)	75 (1.217762)	75 (0.425049)
91	86 (17.97)	91 (4.08e-01)	86 (0.353218)	94 (0.440176)	91 (2.613581)	41 (1.216563)	91 (0.423953)
92	29 (17.74)	76 (3.82e-01)	76 (0.362261)	76 (0.447756)	86 (2.605620)	76 (1.160463)	76 (0.406627)
93	53 (15.70)	94 (3.66e-01)	0 (0.378413)	60 (0.448239)	76 (2.602492)	68 (1.141840)	94 (0.376305)
94	60 (15.22)	41 (3.61e-01)	94 (0.389280)	91 (0.454455)	60 (2.568199)	94 (1.016795)	60 (0.373999)
95	62 (14.74)	60 (3.44e-01)	91 (0.394356)	0 (0.492007)	0 (2.302903)	0 (1.008835)	41 (0.372181)
96	76 (14.74)	0 (3.37e-01)	24 (0.449135)	24 (0.529762)	24 (2.272628)	24 (0.963854)	0 (0.371272)
97	82 (14.20)	24 (3.12e-01)	53 (0.459535)	41 (0.543010)	41 (2.216967)	60 (0.940529)	24 (0.327346)
98	20 (10.56)	53 (2.72e-01)	41 (0.499536)	53 (0.553685)	82 (2.209916)	20 (0.828497)	53 (0.296548)
99	0 (10.00)	82 (2.56e-01)	82 (0.517210)	82 (0.572552)	53 (2.091506)	53 (0.823325)	82 (0.267173)
100	24 (8.72)	20 (2.42e-01)	20 (0.565012)	20 (0.627366)	20 (1.829131)	82 (0.748533)	20 (0.256676)
	Pearson	0.8552 (1.0e-29)	0.7742 (3.5e-21)	0.8373 (1.9e-27)	0.8259 (4.0e-26)	0.8253 (4.6e-26)	0.8417 (5.6e-28)
	Spearman	0.8810 (1.3e-33)	0.8024 (1.1e-23)	0.8828 (6.3e-34)	0.8631 (7.8e-31)	0.8342 (4.5e-27)	0.8684 (1.3e-31)
	Kendall	0.6971 (9.2e-25)	0.6122 (1.8e-19)	0.7012 (4.8e-25)	0.6721 (3.8e-23)	0.6413 (3.3e-21)	0.6813 (9.8e-24)

Table 29: Comparison between SOIV-U and uncertainty-related metrics under MIXED untargeted ASR for **CIFAR-100 + Adversarially Trained ResNet-18**. The MIXED ASR is computed by aggregating FGSM, PGD, C&W, and SQUARE results. Due to page width constraints, class indices are used instead of class names. Absolute correlation values are used for Margin and Confidence because larger values indicate lower vulnerability.

Rank	MIXED ASR	SOIV-U	Margin	Confidence	Entropy	NLL	Brier
1	80 (83.75)	55 (8.90e-01)	96 (-0.031512)	55 (0.059514)	27 (4.152514)	55 (3.192542)	55 (0.915131)
2	77 (79.63)	27 (8.64e-01)	55 (-0.019698)	27 (0.072375)	38 (4.123389)	65 (3.020832)	72 (0.886665)
3	96 (79.35)	72 (8.62e-01)	98 (-0.015828)	65 (0.075167)	77 (4.112690)	72 (2.981505)	27 (0.884272)
4	4 (78.70)	65 (8.61e-01)	35 (-0.007908)	72 (0.076467)	80 (4.109611)	27 (2.965802)	65 (0.884215)
5	55 (76.67)	98 (8.45e-01)	65 (-0.006353)	98 (0.085426)	55 (4.108843)	98 (2.896311)	98 (0.874126)
6	42 (74.11)	80 (8.44e-01)	72 (-0.003643)	80 (0.086087)	64 (4.087991)	45 (2.869786)	80 (0.862605)
7	38 (69.85)	38 (8.40e-01)	10 (0.003518)	38 (0.087139)	44 (4.083058)	3 (2.850515)	3 (0.860269)
8	44 (67.86)	3 (8.38e-01)	27 (0.008213)	3 (0.088888)	33 (4.079519)	4 (2.844616)	38 (0.857861)
9	34 (67.74)	44 (8.34e-01)	3 (0.010136)	44 (0.089542)	79 (4.078345)	80 (2.828350)	44 (0.855525)
10	72 (67.19)	4 (8.27e-01)	44 (0.012961)	35 (0.094939)	65 (4.074974)	64 (2.819495)	35 (0.853748)
11	50 (66.67)	77 (8.24e-01)	30 (0.014578)	77 (0.096449)	88 (4.050375)	38 (2.800854)	4 (0.845880)
12	35 (66.30)	35 (8.24e-01)	37 (0.017331)	18 (0.101355)	72 (4.048343)	74 (2.765593)	96 (0.842196)
13	3 (65.79)	74 (8.19e-01)	80 (0.021705)	74 (0.103114)	43 (4.048121)	44 (2.742513)	77 (0.840748)
14	27 (64.81)	18 (8.16e-01)	92 (0.023178)	64 (0.104492)	42 (4.046430)	18 (2.728067)	74 (0.839329)
15	74 (64.29)	64 (8.14e-01)	38 (0.029361)	33 (0.104669)	18 (4.041493)	84 (2.725154)	18 (0.835561)
16	79 (62.88)	31 (8.09e-01)	78 (0.030325)	88 (0.106908)	3 (4.034004)	93 (2.697921)	64 (0.829968)
17	10 (61.46)	33 (8.07e-01)	74 (0.031020)	31 (0.107413)	31 (4.027771)	77 (2.677707)	31 (0.828244)
18	2 (60.34)	88 (8.05e-01)	18 (0.031340)	96 (0.110368)	74 (4.026056)	50 (2.664383)	33 (0.826694)
19	63 (60.34)	78 (7.97e-01)	93 (0.033232)	78 (0.111528)	78 (4.020716)	31 (2.630854)	88 (0.825109)
20	98 (58.82)	96 (7.95e-01)	2 (0.033438)	42 (0.115098)	34 (3.981411)	35 (2.607979)	45 (0.820899)
21	46 (57.35)	42 (7.93e-01)	11 (0.035837)	4 (0.116398)	19 (3.961722)	42 (2.574540)	78 (0.820850)
22	43 (56.08)	45 (7.91e-01)	33 (0.037550)	79 (0.117625)	36 (3.945743)	78 (2.563723)	93 (0.814789)
23	78 (55.65)	50 (7.87e-01)	88 (0.037842)	37 (0.118618)	4 (3.936081)	88 (2.561911)	2 (0.810905)
24	65 (54.00)	79 (7.87e-01)	83 (0.038436)	2 (0.120986)	98 (3.935023)	19 (2.542447)	42 (0.810134)
25	5 (53.57)	93 (7.86e-01)	31 (0.039434)	43 (0.125146)	50 (3.934924)	33 (2.503316)	37 (0.809944)
26	19 (53.45)	2 (7.85e-01)	77 (0.040577)	93 (0.126502)	35 (3.926672)	2 (2.496205)	10 (0.808919)
27	64 (53.26)	37 (7.82e-01)	46 (0.041861)	50 (0.127309)	66 (3.919607)	11 (2.485958)	50 (0.806140)
28	6 (53.21)	84 (7.81e-01)	73 (0.043197)	19 (0.128760)	2 (3.913764)	22 (2.460112)	79 (0.803185)
29	25 (53.12)	19 (7.79e-01)	7 (0.043799)	10 (0.129478)	37 (3.910584)	79 (2.449902)	84 (0.799691)
30	37 (53.12)	43 (7.73e-01)	47 (0.044702)	45 (0.140780)	51 (3.900959)	37 (2.435016)	19 (0.797388)

Rank	MIXED ASR	SOIV-U	Margin	Confidence	Entropy	NLL	Brier
31	33 (52.68)	10 (7.70e-01)	59 (0.045077)	11 (0.142287)	81 (3.880084)	46 (2.421156)	43 (0.788618)
32	88 (51.61)	11 (7.57e-01)	67 (0.047440)	84 (0.145358)	84 (3.854676)	10 (2.404027)	11 (0.788137)
33	18 (50.00)	34 (7.46e-01)	64 (0.050627)	34 (0.145809)	63 (3.854326)	96 (2.376130)	46 (0.774376)
34	66 (50.00)	46 (7.44e-01)	45 (0.052807)	46 (0.148850)	6 (3.853596)	73 (2.360649)	73 (0.762095)
35	89 (50.00)	81 (7.34e-01)	4 (0.053026)	81 (0.149265)	89 (3.850551)	43 (2.356239)	34 (0.761574)
36	15 (49.26)	6 (7.27e-01)	79 (0.058149)	6 (0.158424)	97 (3.842514)	15 (2.350389)	92 (0.759300)
37	70 (47.86)	15 (7.26e-01)	42 (0.058539)	36 (0.159601)	15 (3.824472)	63 (2.343850)	81 (0.757378)
38	31 (47.66)	73 (7.25e-01)	81 (0.060090)	67 (0.160840)	93 (3.815487)	36 (2.322562)	67 (0.756405)
39	92 (46.59)	36 (7.24e-01)	50 (0.060696)	73 (0.161219)	10 (3.775731)	32 (2.310108)	7 (0.752038)
40	84 (46.21)	63 (7.21e-01)	19 (0.065744)	15 (0.164488)	45 (3.771871)	6 (2.249894)	6 (0.749130)
41	90 (45.45)	67 (7.19e-01)	43 (0.069423)	51 (0.165581)	11 (3.768718)	29 (2.245866)	30 (0.746836)
42	14 (45.37)	51 (7.15e-01)	70 (0.069525)	92 (0.167408)	90 (3.722341)	51 (2.242494)	15 (0.746288)
43	11 (45.00)	92 (7.09e-01)	84 (0.073161)	63 (0.170584)	13 (3.722270)	34 (2.239941)	83 (0.746088)
44	7 (44.85)	66 (7.03e-01)	13 (0.081885)	66 (0.174947)	46 (3.721493)	14 (2.236593)	59 (0.745636)
45	83 (44.35)	59 (7.03e-01)	6 (0.083444)	30 (0.176090)	96 (3.713125)	67 (2.223745)	36 (0.739105)
46	59 (43.97)	83 (7.02e-01)	22 (0.087179)	7 (0.177606)	21 (3.711386)	83 (2.185010)	63 (0.737536)
47	93 (43.48)	7 (7.00e-01)	34 (0.091830)	59 (0.178817)	5 (3.711231)	66 (2.173644)	51 (0.730403)
48	81 (43.24)	13 (6.96e-01)	15 (0.092796)	83 (0.180197)	25 (3.689656)	59 (2.173354)	47 (0.721905)
49	32 (43.10)	30 (6.92e-01)	36 (0.104565)	13 (0.180606)	14 (3.679920)	26 (2.172216)	13 (0.720679)
50	8 (42.57)	89 (6.84e-01)	51 (0.108228)	89 (0.182157)	23 (3.650258)	81 (2.155943)	22 (0.720507)
51	36 (41.13)	22 (6.81e-01)	90 (0.108501)	47 (0.190542)	26 (3.637000)	92 (2.155579)	66 (0.716435)
52	85 (40.13)	47 (6.77e-01)	49 (0.111995)	90 (0.193208)	85 (3.629399)	25 (2.128081)	89 (0.700637)
53	49 (38.75)	14 (6.77e-01)	63 (0.114551)	97 (0.194231)	8 (3.619923)	97 (2.088816)	70 (0.700516)
54	67 (38.64)	97 (6.77e-01)	89 (0.118161)	70 (0.200283)	54 (3.607331)	7 (2.083077)	14 (0.694558)
55	45 (38.46)	32 (6.74e-01)	66 (0.126592)	14 (0.207004)	67 (3.600891)	13 (2.078760)	97 (0.690888)
56	97 (38.28)	90 (6.68e-01)	5 (0.130675)	25 (0.207894)	56 (3.571416)	40 (2.025219)	32 (0.690590)
57	87 (38.19)	26 (6.67e-01)	21 (0.131296)	22 (0.210348)	73 (3.566610)	70 (2.004057)	90 (0.689410)
58	30 (36.72)	70 (6.63e-01)	25 (0.134107)	21 (0.210641)	32 (3.564819)	47 (1.995199)	26 (0.685388)
59	51 (36.03)	25 (6.61e-01)	61 (0.135916)	5 (0.211862)	7 (3.563076)	30 (1.964956)	25 (0.680298)
60	69 (34.09)	5 (6.45e-01)	52 (0.140412)	26 (0.215230)	17 (3.561069)	89 (1.962710)	5 (0.666519)
61	57 (33.59)	21 (6.40e-01)	14 (0.142127)	49 (0.215406)	87 (3.539196)	8 (1.952530)	21 (0.663784)
62	95 (33.57)	29 (6.38e-01)	62 (0.142285)	32 (0.224445)	92 (3.533288)	90 (1.949890)	49 (0.661562)
63	1 (33.54)	49 (6.33e-01)	97 (0.146071)	23 (0.228047)	99 (3.525908)	85 (1.933040)	29 (0.653851)
64	40 (32.76)	8 (6.30e-01)	26 (0.148095)	85 (0.229285)	49 (3.517611)	16 (1.928248)	85 (0.646900)
65	13 (31.82)	85 (6.29e-01)	23 (0.151080)	54 (0.238429)	59 (3.502421)	5 (1.875663)	8 (0.643755)
66	54 (31.67)	23 (6.13e-01)	17 (0.158359)	17 (0.241459)	47 (3.489519)	28 (1.875459)	23 (0.633044)
67	21 (31.55)	16 (6.10e-01)	85 (0.160054)	8 (0.244878)	12 (3.488294)	12 (1.853447)	16 (0.632536)
68	26 (31.25)	54 (6.08e-01)	32 (0.161289)	12 (0.248808)	1 (3.477514)	99 (1.845998)	54 (0.626424)
69	73 (31.03)	12 (6.01e-01)	54 (0.166624)	61 (0.255541)	16 (3.446947)	57 (1.825631)	12 (0.622225)
70	52 (30.95)	99 (5.98e-01)	12 (0.169009)	99 (0.255870)	30 (3.445722)	49 (1.810244)	99 (0.614766)
71	61 (30.26)	17 (5.90e-01)	16 (0.171947)	56 (0.256341)	70 (3.444258)	21 (1.801024)	61 (0.613983)
72	16 (30.00)	40 (5.84e-01)	71 (0.183206)	16 (0.256884)	83 (3.440892)	54 (1.766024)	62 (0.613887)
73	71 (28.85)	61 (5.83e-01)	99 (0.184062)	29 (0.258903)	29 (3.417927)	23 (1.742178)	17 (0.611068)
74	47 (28.70)	56 (5.82e-01)	56 (0.189975)	62 (0.265309)	22 (3.417049)	61 (1.700016)	40 (0.605212)
75	17 (28.38)	62 (5.77e-01)	8 (0.192257)	87 (0.270133)	69 (3.379997)	56 (1.699762)	56 (0.598935)
76	56 (28.29)	87 (5.73e-01)	68 (0.193364)	52 (0.279607)	61 (3.374662)	62 (1.698298)	87 (0.588012)
77	23 (27.70)	57 (5.68e-01)	29 (0.201165)	1 (0.286441)	75 (3.349147)	87 (1.672990)	57 (0.585472)
78	58 (26.92)	1 (5.44e-01)	87 (0.204257)	71 (0.287360)	58 (3.302054)	9 (1.633457)	52 (0.576901)
79	91 (26.43)	28 (5.43e-01)	40 (0.206851)	40 (0.289813)	28 (3.298571)	86 (1.619522)	28 (0.561227)
80	39 (25.60)	52 (5.38e-01)	69 (0.211622)	69 (0.293276)	71 (3.295477)	17 (1.615133)	1 (0.561061)
81	9 (25.00)	71 (5.32e-01)	95 (0.215420)	68 (0.294150)	68 (3.293872)	69 (1.608744)	71 (0.558271)
82	99 (24.31)	69 (5.31e-01)	1 (0.220038)	57 (0.303970)	40 (3.283739)	1 (1.548873)	69 (0.552292)
83	29 (23.48)	68 (5.16e-01)	28 (0.229249)	28 (0.304898)	39 (3.268130)	39 (1.520000)	68 (0.545307)
84	86 (22.73)	95 (5.03e-01)	57 (0.236920)	95 (0.324351)	57 (3.263642)	52 (1.479511)	95 (0.534054)
85	12 (22.66)	48 (5.02e-01)	58 (0.258603)	58 (0.330602)	62 (3.229440)	48 (1.466380)	48 (0.512968)
86	28 (22.50)	39 (4.96e-01)	60 (0.271189)	75 (0.333007)	48 (3.224728)	95 (1.463141)	39 (0.511300)
87	94 (22.09)	9 (4.95e-01)	75 (0.284263)	39 (0.339922)	94 (3.204693)	71 (1.454781)	9 (0.511271)
88	75 (21.81)	75 (4.89e-01)	76 (0.285195)	48 (0.348724)	52 (3.142531)	75 (1.431868)	58 (0.500483)
89	41 (20.62)	58 (4.85e-01)	39 (0.286461)	76 (0.360140)	76 (3.101561)	76 (1.393401)	75 (0.500304)
90	22 (20.45)	76 (4.65e-01)	48 (0.301114)	9 (0.365756)	95 (3.069411)	91 (1.385461)	76 (0.483411)
91	62 (20.12)	86 (4.62e-01)	9 (0.302937)	60 (0.374024)	9 (3.038831)	58 (1.365945)	86 (0.476715)
92	76 (20.12)	91 (4.45e-01)	94 (0.332770)	94 (0.378476)	60 (2.981044)	68 (1.351330)	91 (0.457432)
93	53 (19.89)	94 (4.29e-01)	86 (0.343583)	86 (0.405809)	91 (2.842398)	41 (1.286253)	60 (0.443605)
94	68 (18.75)	60 (4.20e-01)	0 (0.347172)	91 (0.417283)	86 (2.838117)	94 (1.178855)	94 (0.438701)
95	48 (15.91)	0 (3.76e-01)	91 (0.365745)	0 (0.450079)	24 (2.661874)	24 (1.174160)	0 (0.402689)
96	60 (14.67)	24 (3.73e-01)	53 (0.371971)	24 (0.455070)	82 (2.636044)	0 (1.120435)	24 (0.387815)
97	82 (13.64)	41 (3.72e-01)	24 (0.373416)	53 (0.461258)	0 (2.568712)	60 (1.116308)	41 (0.380413)
98	24 (12.79)	53 (3.55e-01)	82 (0.440265)	82 (0.491840)	53 (2.563948)	53 (1.028482)	53 (0.377585)
99	20 (9.66)	82 (3.21e-01)	41 (0.510815)	41 (0.546197)	41 (2.196816)	82 (0.931082)	82 (0.330600)
100	0 (7.50)	20 (2.77e-01)	20 (0.514114)	20 (0.570000)	20 (2.191192)	20 (0.928843)	20 (0.288749)
	Pearson	0.8608 (1.7e-30)	0.7727 (4.7e-21)	0.8322 (7.6e-27)	0.8112 (1.4e-24)	0.8286 (2.0e-26)	0.8514 (3.2e-29)
	Spearman	0.8867 (1.4e-34)	0.8121 (1.2e-24)	0.8899 (3.5e-35)	0.8725 (3.0e-32)	0.8359 (2.8e-27)	0.8788 (3.0e-33)
	Kendall	0.6981 (7.9e-25)	0.6197 (6.5e-20)	0.7001 (5.7e-25)	0.6815 (9.5e-24)	0.6483 (1.2e-21)	0.6900 (2.7e-24)

D.4 Results Based on Conditional Sample Second-Order Statistics

In the main paper, Sections 3.2 and 3.3 introduce two types of conditional second-order statistics derived from clean validation data: the conditional sample variance (Eq. 1 of the main paper)/covariance (Eq. 4 of the main paper) and the ideal-centered conditional variance (Eq. 2 of the main paper)/covariance (Eq. 5 of the main paper). In the experiments reported in the main paper, as well as in Section C of this Appendix, we focus on the ideal-centered methods and use them to define the SOIV metrics. In this section, we additionally examine whether a similar analysis can be conducted using statistics centered at the empirical sample mean (as defined in Eq. 1 and Eq. 4 of the main paper), on CIFAR-10 with ResNet-18 and on the two adversarially trained CIFAR-100 models of Appendix Sections C.4 and C.5. Specifically, we compute the conditional sample variance and covariance and analyze their relationships with ASR. The CIFAR-10 comparisons are summarized in Table 30, Table 31, and Fig. 7.

On CIFAR-10 with ResNet-18, the rankings derived from conditional sample variance-based statistics also exhibit strong correlations with ASR. In the untargeted case, the sample-variance ranking achieves high correlations with ASR across multiple attacks (e.g., Spearman correlations ranging from 0.91 to 0.96). Similarly, for the targeted case, the sample-covariance ranking shows consistently strong correlations with targeted ASR (e.g., Spearman correlations around 0.73). In this setting, sample mean-centered statistics can also reflect class-wise and directional adversarial vulnerability patterns.

On the adversarially trained CIFAR-100 models, the sample-centered statistic is negatively correlated with class-wise ASR under every untargeted attack, whereas the ideal-centered statistic reported in Tables 19 and 21 remains positively correlated. Tables 32 and 34 report the untargeted comparison, and Tables 33 and 35 the targeted one. Both models assign low confidence to the true class, with a mean true-class probability of 0.244 and 0.214 on the validation set. Since the predicted probabilities lie in $[0, 1]$, a class whose true-class probabilities are all close to zero shows little variability, and the sample-centered statistic therefore assigns the smallest values to the most vulnerable classes. Ideal centering measures deviation from the ideal confidence of 1 rather than from the class mean, and is not subject to this effect. In the targeted setting both formulations remain positively correlated.

Table 30: Comparison between sample-variance ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED) for **CIFAR-10 + ResNet-18**.

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	AUTOATTACK	MIXED
1	cat (1.49e-2)	cat (75.1)	cat (78.2)	cat (75.6)	cat (86.6)	cat (81.2)	cat (79.3)
2	deer (1.11e-2)	deer (62.0)	deer (67.5)	deer (73.3)	deer (78.1)	deer (71.1)	deer (70.4)
3	dog (1.02e-2)	dog (60.9)	dog (62.6)	dog (62.1)	dog (71.8)	dog (63.8)	dog (64.2)
4	bird (9.61e-3)	bird (51.3)	bird (55.8)	bird (58.9)	bird (66.7)	bird (58.8)	bird (58.3)
5	horse (8.24e-3)	frog (45.5)	frog (50.8)	frog (53.4)	frog (60.8)	frog (52.6)	frog (52.6)
6	airplane (6.82e-3)	horse (41.0)	airplane (43.0)	airplane (44.4)	airplane (55.3)	airplane (45.2)	airplane (45.6)
7	frog (6.38e-3)	airplane (40.0)	horse (42.4)	horse (40.8)	horse (52.3)	horse (44.2)	horse (44.1)
8	truck (5.39e-3)	truck (33.7)	truck (33.9)	ship (35.7)	ship (46.6)	ship (36.5)	ship (36.6)
9	automobile (5.10e-3)	automobile (31.2)	ship (33.8)	truck (32.1)	truck (44.0)	truck (35.7)	truck (35.9)
10	ship (3.66e-3)	ship (30.3)	automobile (30.4)	automobile (28.1)	automobile (37.3)	automobile (31.8)	automobile (31.8)
Pearson		0.966 (5.6e-06)	0.946 (3.5e-05)	0.899 (4.0e-04)	0.925 (1.3e-04)	0.944 (4.0e-05)	0.940 (5.3e-05)
Spearman		0.964 (7.0e-06)	0.939 (5.6e-05)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)
Kendall		0.911 (3.0e-05)	0.822 (3.6e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)

Table 32: Comparison between sample-variance ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) for **CIFAR-100 + Adversarially Trained PreActResNet-18**. Due to page width constraints, class indices are used instead of class names.

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	MIXED
1	41 (1.55e-01)	80 (94.12)	77 (95.00)	80 (94.12)	55 (73.33)	80 (88.24)
2	86 (1.20e-01)	55 (93.33)	80 (94.12)	77 (90.00)	80 (70.59)	55 (85.00)
3	91 (1.12e-01)	77 (90.00)	55 (93.33)	38 (88.46)	4 (68.00)	77 (83.75)
4	9 (1.11e-01)	4 (80.00)	33 (86.21)	44 (85.00)	42 (64.00)	4 (79.00)
5	20 (1.05e-01)	44 (80.00)	44 (85.00)	4 (84.00)	74 (64.00)	38 (76.92)

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	MIXED
6	29 (1.04e-01)	38 (76.92)	38 (84.62)	74 (84.00)	35 (62.50)	44 (76.25)
7	40 (1.03e-01)	3 (76.47)	4 (84.00)	72 (81.82)	96 (62.50)	74 (75.00)
8	32 (9.90e-02)	42 (76.00)	96 (81.25)	55 (80.00)	2 (61.54)	42 (74.00)
9	48 (9.73e-02)	74 (76.00)	42 (80.00)	79 (78.57)	77 (60.00)	35 (71.88)
10	39 (9.67e-02)	35 (75.00)	35 (79.17)	42 (76.00)	34 (58.62)	34 (68.97)
11	57 (9.45e-02)	2 (73.08)	79 (78.57)	66 (73.53)	38 (57.69)	96 (68.75)
12	24 (9.23e-02)	27 (72.73)	98 (77.78)	33 (72.41)	44 (55.00)	72 (68.18)
13	8 (9.11e-02)	72 (72.73)	25 (77.27)	34 (72.41)	50 (52.94)	79 (67.86)
14	22 (8.14e-02)	34 (72.41)	27 (77.27)	35 (70.83)	25 (50.00)	2 (67.31)
15	0 (8.09e-02)	98 (72.22)	2 (76.92)	50 (70.59)	27 (50.00)	50 (66.18)
16	76 (7.87e-02)	50 (70.59)	3 (76.47)	88 (70.37)	79 (50.00)	27 (65.91)
17	28 (7.82e-02)	33 (68.97)	18 (76.32)	3 (64.71)	11 (47.62)	33 (65.52)
18	84 (7.55e-02)	96 (68.75)	74 (76.00)	25 (63.64)	72 (45.45)	25 (63.64)
19	82 (7.51e-02)	79 (64.29)	78 (74.07)	27 (63.64)	98 (44.44)	3 (63.24)
20	53 (7.47e-02)	25 (63.64)	72 (72.73)	96 (62.50)	63 (40.74)	98 (62.50)
21	63 (7.39e-02)	18 (63.16)	34 (72.41)	14 (61.29)	18 (39.47)	88 (57.41)
22	26 (7.38e-02)	14 (61.29)	67 (70.83)	64 (60.87)	14 (38.71)	18 (56.58)
23	14 (7.11e-02)	5 (59.38)	50 (70.59)	43 (60.00)	90 (38.71)	14 (56.45)
24	95 (6.60e-02)	63 (59.26)	6 (70.00)	5 (59.38)	10 (38.10)	63 (55.56)
25	16 (6.48e-02)	88 (59.26)	19 (69.23)	65 (59.09)	3 (35.29)	67 (55.21)
26	99 (6.43e-02)	67 (58.33)	63 (66.67)	15 (58.82)	59 (34.78)	11 (53.57)
27	45 (6.39e-02)	19 (57.69)	88 (66.67)	67 (58.33)	33 (34.48)	19 (52.88)
28	58 (6.33e-02)	11 (57.14)	14 (64.52)	97 (58.06)	43 (34.29)	43 (52.86)
29	75 (6.20e-02)	43 (57.14)	46 (64.29)	2 (57.69)	30 (33.33)	78 (51.85)
30	62 (5.62e-02)	46 (57.14)	10 (61.90)	63 (55.56)	45 (33.33)	5 (51.56)
31	12 (5.56e-02)	70 (56.25)	49 (61.11)	98 (55.56)	67 (33.33)	10 (51.19)
32	87 (5.52e-02)	78 (55.56)	59 (60.87)	19 (53.85)	78 (33.33)	65 (50.00)
33	94 (5.41e-02)	84 (55.17)	43 (60.00)	10 (52.38)	88 (33.33)	15 (49.26)
34	4 (5.08e-02)	15 (52.94)	5 (59.38)	11 (52.38)	85 (32.43)	46 (48.21)
35	69 (5.05e-02)	10 (52.38)	70 (59.38)	45 (52.38)	65 (31.82)	64 (47.83)
36	85 (4.99e-02)	45 (52.38)	65 (59.09)	36 (51.52)	92 (31.82)	59 (47.83)
37	56 (4.90e-02)	59 (52.17)	15 (58.82)	23 (50.00)	7 (31.25)	66 (47.79)
38	83 (4.68e-02)	6 (50.00)	11 (57.14)	93 (50.00)	70 (31.25)	45 (47.62)
39	11 (4.50e-02)	49 (50.00)	31 (57.14)	18 (47.37)	19 (30.77)	6 (46.67)
40	54 (4.38e-02)	65 (50.00)	37 (56.00)	8 (47.06)	64 (30.43)	70 (46.09)
41	1 (4.34e-02)	90 (48.39)	66 (55.88)	7 (46.88)	36 (30.30)	90 (45.97)
42	71 (4.27e-02)	37 (48.00)	84 (55.17)	89 (46.88)	97 (29.03)	49 (45.83)
43	59 (4.20e-02)	64 (47.83)	30 (54.55)	17 (46.15)	71 (28.95)	30 (43.94)
44	97 (4.12e-02)	7 (46.88)	81 (54.55)	30 (45.45)	46 (28.57)	36 (43.94)
45	61 (4.11e-02)	93 (44.44)	45 (52.38)	90 (45.16)	5 (28.12)	7 (43.75)
46	60 (4.05e-02)	8 (44.12)	64 (52.17)	32 (44.83)	40 (28.00)	97 (43.55)
47	93 (4.02e-02)	89 (43.75)	90 (51.61)	49 (44.44)	49 (27.78)	84 (43.10)
48	13 (3.87e-02)	31 (42.86)	36 (51.52)	78 (44.44)	32 (27.59)	89 (41.41)
49	50 (3.70e-02)	30 (42.42)	7 (50.00)	59 (43.48)	95 (27.27)	85 (41.22)
50	25 (3.61e-02)	36 (42.42)	8 (50.00)	85 (43.24)	52 (26.83)	8 (40.44)
51	5 (3.44e-02)	97 (41.94)	13 (50.00)	31 (42.86)	6 (26.67)	93 (40.28)
52	51 (3.44e-02)	66 (41.18)	89 (50.00)	46 (42.86)	83 (26.67)	31 (40.18)
53	23 (3.34e-02)	13 (40.91)	92 (50.00)	81 (42.42)	15 (26.47)	81 (40.15)
54	15 (3.31e-02)	92 (40.91)	51 (48.65)	69 (41.86)	91 (25.71)	37 (40.00)
55	73 (3.28e-02)	51 (40.54)	85 (48.65)	84 (41.38)	73 (25.00)	92 (39.77)
56	47 (3.28e-02)	85 (40.54)	87 (47.22)	6 (40.00)	87 (25.00)	87 (37.50)
57	66 (3.14e-02)	83 (40.00)	97 (45.16)	40 (40.00)	89 (25.00)	32 (37.07)
58	70 (3.10e-02)	81 (39.39)	61 (44.74)	26 (39.13)	81 (24.24)	13 (36.36)
59	52 (3.09e-02)	26 (39.13)	93 (44.44)	87 (38.89)	13 (22.73)	26 (35.87)
60	68 (3.09e-02)	87 (38.89)	26 (43.48)	28 (37.50)	57 (22.58)	51 (35.81)
61	7 (3.08e-02)	56 (38.10)	83 (43.33)	70 (37.50)	93 (22.22)	17 (35.26)
62	49 (3.05e-02)	71 (36.84)	32 (41.38)	92 (36.36)	26 (21.74)	40 (35.00)
63	19 (2.67e-02)	52 (36.59)	17 (41.03)	37 (36.00)	1 (21.62)	83 (35.00)
64	90 (2.65e-02)	40 (36.00)	21 (40.00)	57 (35.48)	51 (21.62)	71 (32.89)
65	46 (2.58e-02)	21 (35.56)	28 (40.00)	12 (35.29)	56 (21.43)	56 (32.74)
66	6 (2.51e-02)	57 (35.48)	71 (39.47)	22 (34.38)	84 (20.69)	57 (32.26)
67	36 (2.49e-02)	1 (35.14)	95 (39.39)	91 (34.29)	8 (20.59)	91 (32.14)
68	64 (2.36e-02)	32 (34.48)	52 (39.02)	61 (34.21)	12 (20.59)	21 (31.67)
69	92 (2.30e-02)	17 (33.33)	54 (38.10)	21 (33.33)	66 (20.59)	28 (31.25)
70	67 (2.30e-02)	28 (32.50)	56 (38.10)	56 (33.33)	17 (20.51)	95 (31.06)
71	72 (2.28e-02)	91 (31.43)	91 (37.14)	51 (32.43)	37 (20.00)	1 (30.41)
72	17 (2.06e-02)	99 (30.56)	40 (36.00)	73 (32.14)	99 (19.44)	52 (29.88)
73	30 (2.04e-02)	95 (30.30)	57 (35.48)	13 (31.82)	22 (18.75)	61 (29.61)
74	21 (1.99e-02)	12 (29.41)	1 (35.14)	9 (30.56)	31 (17.86)	73 (29.46)
75	34 (1.91e-02)	47 (29.17)	22 (34.38)	99 (30.56)	21 (17.78)	22 (28.91)
76	31 (1.85e-02)	16 (28.57)	47 (33.33)	83 (30.00)	47 (16.67)	12 (28.68)
77	3 (1.67e-02)	54 (28.57)	99 (33.33)	1 (29.73)	54 (16.67)	99 (28.47)
78	74 (1.63e-02)	73 (28.57)	58 (32.43)	95 (27.27)	61 (15.79)	69 (27.91)
79	89 (1.60e-02)	22 (28.12)	16 (32.14)	58 (27.03)	48 (15.56)	54 (27.38)
80	2 (1.57e-02)	69 (27.91)	73 (32.14)	71 (26.32)	28 (15.00)	23 (25.69)
81	98 (1.55e-02)	39 (26.83)	39 (31.71)	54 (26.19)	39 (14.63)	47 (25.00)
82	81 (1.55e-02)	48 (26.67)	69 (30.23)	29 (25.81)	53 (13.95)	58 (24.32)
83	10 (1.40e-02)	58 (24.32)	12 (29.41)	75 (22.92)	68 (13.95)	39 (23.17)

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	MIXED
84	42 (1.37e-02)	61 (23.68)	48 (28.89)	86 (21.88)	9 (13.89)	48 (22.78)
85	77 (1.36e-02)	41 (23.08)	41 (28.21)	47 (20.83)	94 (13.64)	16 (22.32)
86	38 (1.32e-02)	23 (22.22)	94 (25.00)	76 (20.51)	58 (13.51)	9 (21.53)
87	79 (1.12e-02)	68 (20.93)	68 (23.26)	48 (20.00)	41 (12.82)	41 (20.51)
88	80 (1.09e-02)	75 (20.83)	29 (22.58)	39 (19.51)	75 (12.50)	75 (19.27)
89	18 (1.08e-02)	60 (19.57)	9 (22.22)	68 (18.60)	86 (12.50)	68 (19.19)
90	43 (8.77e-03)	9 (19.44)	23 (22.22)	94 (18.18)	69 (11.63)	94 (18.75)
91	78 (8.34e-03)	86 (18.75)	60 (21.74)	41 (17.95)	16 (10.71)	86 (17.97)
92	55 (8.04e-03)	94 (18.18)	75 (20.83)	16 (17.86)	76 (10.26)	29 (17.74)
93	33 (7.49e-03)	62 (17.95)	62 (20.51)	52 (17.07)	82 (9.09)	53 (15.70)
94	37 (7.06e-03)	53 (16.28)	86 (18.75)	53 (16.28)	23 (8.33)	60 (15.22)
95	35 (6.69e-03)	29 (16.13)	0 (17.50)	82 (15.91)	60 (6.52)	62 (14.74)
96	44 (6.55e-03)	82 (15.91)	53 (16.28)	62 (15.38)	29 (6.45)	76 (14.74)
97	88 (6.18e-03)	0 (15.00)	82 (15.91)	20 (13.33)	62 (5.13)	82 (14.20)
98	65 (6.09e-03)	76 (12.82)	76 (15.38)	60 (13.04)	24 (4.65)	20 (10.56)
99	96 (5.70e-03)	20 (11.11)	24 (13.95)	24 (9.30)	20 (4.44)	0 (10.00)
100	27 (4.43e-03)	24 (6.98)	20 (13.33)	0 (5.00)	0 (2.50)	24 (8.72)
	Pearson	-0.6462 (3.9e-13)	-0.6871 (2.9e-15)	-0.5960 (6.1e-11)	-0.5661 (8.4e-10)	-0.6499 (2.6e-13)
	Spearman	-0.7048 (2.8e-16)	-0.7469 (4.6e-19)	-0.6509 (2.3e-13)	-0.6427 (5.7e-13)	-0.7096 (1.4e-16)
	Kendall	-0.5098 (5.7e-14)	-0.5569 (2.2e-16)	-0.4567 (1.7e-11)	-0.4630 (8.8e-12)	-0.5092 (6.1e-14)

Table 33: Comparison between sample-covariance ranking and ASR rankings under multiple targeted attacks (I-FGSM, PGD, C&W, and MIXED) when the source class is 0 (Apple) for **CIFAR-100 + Adversarially Trained PreActResNet-18**. For correlation analysis, the absolute value of the sample covariance is used since it is non-positive while ASR is non-negative.

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
1	53 (-0.008754)	83 (87.50)	83 (95.00)	83 (100.00)	83 (94.17)
2	57 (-0.006967)	10 (72.50)	10 (77.50)	26 (97.50)	70 (80.83)
3	70 (-0.004472)	70 (70.00)	70 (77.50)	57 (95.00)	10 (80.00)
4	92 (-0.003971)	57 (67.50)	57 (72.50)	70 (95.00)	57 (78.33)
5	45 (-0.002616)	62 (65.00)	62 (67.50)	36 (92.50)	62 (72.50)
6	62 (-0.002399)	53 (62.50)	82 (67.50)	10 (90.00)	36 (69.17)
7	10 (-0.001770)	82 (57.50)	53 (65.00)	62 (85.00)	26 (68.33)
8	61 (-0.001641)	25 (55.00)	16 (62.50)	32 (82.50)	53 (68.33)
9	84 (-0.001561)	32 (55.00)	22 (62.50)	11 (80.00)	32 (65.00)
10	54 (-0.001516)	61 (55.00)	36 (62.50)	22 (80.00)	25 (63.33)
11	2 (-0.001460)	26 (52.50)	47 (60.00)	45 (80.00)	82 (63.33)
12	26 (-0.001456)	36 (52.50)	61 (60.00)	25 (77.50)	22 (63.33)
13	83 (-0.001430)	92 (52.50)	94 (60.00)	53 (77.50)	61 (62.50)
14	1 (-0.001414)	11 (50.00)	25 (57.50)	94 (77.50)	94 (62.50)
15	36 (-0.001156)	16 (50.00)	32 (57.50)	43 (75.00)	16 (60.83)
16	51 (-0.000974)	47 (50.00)	86 (57.50)	7 (72.50)	11 (60.00)
17	77 (-0.000973)	94 (50.00)	26 (55.00)	61 (72.50)	47 (60.00)
18	7 (-0.000886)	22 (47.50)	13 (52.50)	16 (70.00)	45 (58.33)
19	32 (-0.000876)	77 (47.50)	92 (52.50)	47 (70.00)	92 (57.50)
20	25 (-0.000839)	28 (45.00)	98 (52.50)	84 (67.50)	43 (54.17)
21	11 (-0.000838)	45 (45.00)	11 (50.00)	92 (67.50)	98 (53.33)
22	16 (-0.000782)	40 (42.50)	28 (50.00)	4 (65.00)	86 (51.67)
23	22 (-0.000748)	46 (42.50)	45 (50.00)	82 (65.00)	77 (51.67)
24	35 (-0.000742)	87 (42.50)	77 (50.00)	98 (65.00)	46 (50.00)
25	98 (-0.000679)	88 (42.50)	14 (47.50)	46 (60.00)	7 (48.33)
26	14 (-0.000679)	98 (42.50)	40 (47.50)	50 (60.00)	4 (46.67)
27	78 (-0.000679)	5 (40.00)	43 (47.50)	86 (60.00)	54 (45.83)
28	28 (-0.000662)	43 (40.00)	46 (47.50)	2 (57.50)	84 (45.83)
29	5 (-0.000657)	54 (40.00)	54 (45.00)	77 (57.50)	28 (45.00)
30	9 (-0.000629)	1 (37.50)	87 (45.00)	6 (55.00)	40 (45.00)
31	50 (-0.000628)	6 (37.50)	88 (45.00)	51 (55.00)	88 (45.00)
32	43 (-0.000619)	13 (37.50)	1 (42.50)	35 (52.50)	2 (45.00)
33	41 (-0.000610)	63 (37.50)	2 (42.50)	54 (52.50)	13 (44.17)
34	65 (-0.000593)	86 (37.50)	5 (42.50)	63 (52.50)	6 (44.17)
35	4 (-0.000586)	2 (35.00)	9 (42.50)	93 (52.50)	51 (43.33)

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
36	80 (-0.000572)	4 (35.00)	58 (42.50)	15 (47.50)	63 (42.50)
37	63 (-0.000550)	7 (35.00)	81 (42.50)	78 (47.50)	35 (41.67)
38	86 (-0.000550)	9 (35.00)	4 (40.00)	88 (47.50)	50 (41.67)
39	24 (-0.000547)	14 (35.00)	6 (40.00)	40 (45.00)	5 (40.83)
40	3 (-0.000545)	51 (35.00)	35 (40.00)	13 (42.50)	1 (40.00)
41	88 (-0.000534)	58 (35.00)	51 (40.00)	19 (42.50)	14 (40.00)
42	46 (-0.000516)	15 (32.50)	84 (40.00)	1 (40.00)	93 (39.17)
43	6 (-0.000512)	19 (32.50)	7 (37.50)	5 (40.00)	81 (38.33)
44	89 (-0.000496)	29 (32.50)	19 (37.50)	21 (40.00)	15 (38.33)
45	82 (-0.000493)	31 (32.50)	29 (37.50)	28 (40.00)	78 (38.33)
46	47 (-0.000479)	35 (32.50)	42 (37.50)	29 (40.00)	19 (37.50)
47	39 (-0.000474)	41 (32.50)	63 (37.50)	65 (40.00)	87 (37.50)
48	48 (-0.000465)	50 (32.50)	89 (37.50)	72 (40.00)	29 (36.67)
49	67 (-0.000457)	78 (32.50)	3 (35.00)	81 (40.00)	9 (35.00)
50	34 (-0.000456)	81 (32.50)	15 (35.00)	14 (37.50)	21 (34.17)
51	94 (-0.000455)	90 (32.50)	21 (35.00)	85 (37.50)	58 (34.17)
52	99 (-0.000445)	3 (30.00)	31 (35.00)	3 (35.00)	3 (33.33)
53	64 (-0.000444)	20 (30.00)	39 (35.00)	24 (35.00)	42 (32.50)
54	81 (-0.000441)	33 (30.00)	41 (35.00)	34 (35.00)	31 (31.67)
55	19 (-0.000438)	42 (30.00)	48 (35.00)	64 (35.00)	64 (31.67)
56	90 (-0.000431)	71 (30.00)	78 (35.00)	67 (32.50)	85 (31.67)
57	18 (-0.000417)	84 (30.00)	90 (35.00)	80 (32.50)	90 (31.67)
58	40 (-0.000414)	89 (30.00)	93 (35.00)	96 (32.50)	80 (30.83)
59	55 (-0.000413)	93 (30.00)	20 (32.50)	42 (30.00)	96 (30.83)
60	27 (-0.000403)	96 (30.00)	37 (32.50)	55 (30.00)	34 (30.83)
61	33 (-0.000400)	21 (27.50)	50 (32.50)	99 (30.00)	41 (30.83)
62	15 (-0.000398)	27 (27.50)	64 (32.50)	9 (27.50)	65 (30.83)
63	93 (-0.000397)	34 (27.50)	80 (32.50)	31 (27.50)	89 (30.83)
64	13 (-0.000397)	37 (27.50)	91 (32.50)	37 (27.50)	99 (30.00)
65	42 (-0.000394)	39 (27.50)	99 (32.50)	74 (27.50)	37 (29.17)
66	72 (-0.000386)	44 (27.50)	8 (30.00)	90 (27.50)	67 (29.17)
67	87 (-0.000376)	48 (27.50)	18 (30.00)	91 (27.50)	91 (29.17)
68	74 (-0.000372)	64 (27.50)	34 (30.00)	18 (25.00)	48 (29.17)
69	44 (-0.000352)	66 (27.50)	66 (30.00)	41 (25.00)	72 (28.33)
70	56 (-0.000352)	80 (27.50)	67 (30.00)	48 (25.00)	24 (28.33)
71	79 (-0.000351)	85 (27.50)	71 (30.00)	58 (25.00)	20 (27.50)
72	38 (-0.000351)	91 (27.50)	85 (30.00)	79 (25.00)	55 (26.67)
73	69 (-0.000351)	99 (27.50)	96 (30.00)	87 (25.00)	66 (26.67)
74	58 (-0.000346)	8 (25.00)	24 (27.50)	89 (25.00)	71 (26.67)
75	91 (-0.000345)	23 (25.00)	27 (27.50)	44 (22.50)	18 (25.83)
76	20 (-0.000342)	38 (25.00)	33 (27.50)	52 (22.50)	44 (25.83)
77	66 (-0.000325)	65 (25.00)	44 (27.50)	66 (22.50)	39 (25.83)
78	29 (-0.000324)	67 (25.00)	55 (27.50)	69 (22.50)	33 (25.00)
79	97 (-0.000312)	12 (22.50)	56 (27.50)	20 (20.00)	27 (24.17)
80	37 (-0.000310)	18 (22.50)	65 (27.50)	71 (20.00)	69 (23.33)
81	8 (-0.000306)	24 (22.50)	12 (25.00)	97 (20.00)	23 (22.50)
82	21 (-0.000305)	55 (22.50)	17 (25.00)	23 (17.50)	38 (22.50)
83	23 (-0.000303)	69 (22.50)	23 (25.00)	27 (17.50)	8 (20.00)
84	96 (-0.000296)	72 (22.50)	38 (25.00)	33 (17.50)	17 (20.00)
85	31 (-0.000285)	17 (20.00)	69 (25.00)	38 (17.50)	52 (20.00)
86	73 (-0.000275)	52 (20.00)	30 (22.50)	17 (15.00)	56 (20.00)
87	17 (-0.000259)	56 (17.50)	72 (22.50)	39 (15.00)	74 (20.00)
88	68 (-0.000258)	59 (17.50)	97 (22.50)	56 (15.00)	79 (20.00)
89	59 (-0.000251)	75 (17.50)	49 (20.00)	76 (15.00)	97 (20.00)
90	85 (-0.000237)	97 (17.50)	59 (20.00)	73 (12.50)	12 (17.50)
91	12 (-0.000232)	30 (15.00)	60 (20.00)	30 (10.00)	30 (15.83)
92	71 (-0.000223)	73 (15.00)	73 (20.00)	49 (7.50)	73 (15.83)
93	52 (-0.000222)	74 (15.00)	79 (20.00)	59 (7.50)	76 (15.83)
94	30 (-0.000215)	76 (15.00)	52 (17.50)	60 (7.50)	59 (15.00)
95	75 (-0.000208)	79 (15.00)	68 (17.50)	68 (7.50)	60 (13.33)
96	95 (-0.000205)	60 (12.50)	74 (17.50)	8 (5.00)	49 (12.50)
97	60 (-0.000167)	68 (12.50)	75 (17.50)	12 (5.00)	68 (12.50)

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
98	49 (-0.000158)	49 (10.00)	76 (17.50)	75 (2.50)	75 (12.50)
99	76 (-0.000144)	95 (7.50)	95 (12.50)	95 (0.00)	95 (6.67)
	Pearson	0.5811 (2.9e-10)	0.5389 (8.7e-09)	0.5378 (9.5e-09)	0.5692 (7.9e-10)
	Spearman	0.7898 (2.6e-22)	0.7951 (8.7e-23)	0.8340 (8.6e-27)	0.8389 (2.3e-27)
	Kendall	0.6092 (4.1e-19)	0.6138 (2.2e-19)	0.6480 (2.0e-21)	0.6482 (2.0e-21)

Table 34: Comparison between sample-variance ranking and ASR rankings under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, and MIXED) for **CIFAR-100 + Adversarially Trained ResNet-18**. Due to page width constraints, class indices are used instead of class names.

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	MIXED
1	41 (1.69e-01)	80 (90.00)	80 (95.00)	80 (90.00)	96 (78.26)	80 (83.75)
2	86 (1.11e-01)	77 (85.19)	55 (93.33)	77 (85.19)	4 (74.07)	77 (79.63)
3	91 (1.07e-01)	96 (82.61)	77 (92.59)	4 (81.48)	50 (61.90)	96 (79.35)
4	9 (9.50e-02)	55 (80.00)	42 (85.71)	44 (76.19)	38 (61.76)	4 (78.70)
5	20 (9.44e-02)	4 (77.78)	2 (82.76)	42 (75.00)	42 (60.71)	55 (76.67)
6	29 (9.03e-02)	98 (76.47)	35 (82.61)	96 (73.91)	55 (60.00)	42 (74.11)
7	57 (8.57e-02)	79 (75.76)	96 (82.61)	55 (73.33)	80 (60.00)	38 (69.85)
8	40 (8.08e-02)	10 (75.00)	4 (81.48)	66 (71.43)	72 (56.25)	44 (67.86)
9	48 (7.93e-02)	42 (75.00)	72 (81.25)	34 (67.74)	77 (55.56)	34 (67.74)
10	24 (7.71e-02)	34 (74.19)	78 (80.65)	38 (67.65)	35 (52.17)	72 (67.19)
11	0 (7.47e-02)	35 (73.91)	38 (79.41)	27 (66.67)	27 (51.85)	50 (66.67)
12	32 (7.36e-02)	3 (73.68)	3 (78.95)	74 (66.67)	34 (51.61)	35 (66.30)
13	53 (6.62e-02)	44 (71.43)	79 (78.79)	3 (63.16)	44 (47.62)	3 (65.79)
14	82 (6.37e-02)	50 (71.43)	34 (77.42)	72 (62.50)	3 (47.37)	27 (64.81)
15	39 (6.19e-02)	74 (71.43)	98 (76.47)	63 (62.07)	2 (44.83)	74 (64.29)
16	8 (6.12e-02)	38 (70.59)	44 (76.19)	64 (60.87)	74 (42.86)	79 (62.88)
17	28 (6.08e-02)	46 (70.59)	50 (76.19)	65 (60.00)	10 (41.67)	10 (61.46)
18	16 (5.91e-02)	72 (68.75)	74 (76.19)	15 (58.82)	63 (41.38)	2 (60.34)
19	22 (5.88e-02)	27 (66.67)	33 (75.00)	10 (58.33)	46 (41.18)	63 (60.34)
20	76 (5.64e-02)	5 (65.71)	27 (74.07)	88 (58.06)	70 (40.00)	98 (58.82)
21	45 (5.35e-02)	2 (65.52)	63 (72.41)	79 (57.58)	79 (39.39)	46 (57.35)
22	26 (5.23e-02)	63 (65.52)	65 (72.00)	50 (57.14)	90 (39.39)	43 (56.08)
23	84 (5.14e-02)	78 (64.52)	10 (70.83)	43 (56.76)	18 (38.89)	78 (55.65)
24	14 (4.94e-02)	25 (62.50)	46 (70.59)	35 (56.52)	19 (37.93)	65 (54.00)
25	95 (4.76e-02)	37 (62.50)	18 (69.44)	89 (55.26)	43 (37.84)	5 (53.57)
26	4 (4.74e-02)	43 (62.16)	5 (68.57)	8 (54.05)	6 (35.90)	19 (53.45)
27	75 (4.53e-02)	6 (61.54)	43 (67.57)	37 (53.12)	98 (35.29)	64 (53.26)
28	99 (4.47e-02)	33 (60.71)	6 (66.67)	19 (51.72)	64 (34.78)	6 (53.21)
29	94 (4.40e-02)	11 (60.00)	25 (66.67)	25 (50.00)	93 (34.78)	25 (53.12)
30	87 (4.08e-02)	31 (59.38)	19 (65.52)	6 (48.72)	37 (34.38)	37 (53.12)
31	62 (3.80e-02)	92 (59.09)	64 (65.22)	5 (48.57)	89 (34.21)	33 (52.68)
32	58 (3.77e-02)	19 (58.62)	81 (64.86)	36 (48.39)	14 (33.33)	88 (51.61)
33	12 (3.74e-02)	18 (58.33)	83 (64.52)	78 (48.39)	25 (33.33)	89 (50.00)
34	85 (3.62e-02)	88 (58.06)	92 (63.64)	2 (48.28)	84 (33.33)	66 (50.00)
35	1 (3.53e-02)	65 (56.00)	31 (62.50)	46 (47.06)	15 (32.35)	18 (50.00)
36	25 (3.45e-02)	89 (55.26)	37 (62.50)	98 (47.06)	83 (32.26)	15 (49.26)
37	63 (3.41e-02)	70 (54.29)	88 (61.29)	31 (46.88)	92 (31.82)	70 (47.86)
38	69 (3.26e-02)	64 (52.17)	11 (60.00)	97 (46.88)	5 (31.43)	31 (47.66)
39	83 (3.03e-02)	14 (51.85)	7 (58.82)	33 (46.43)	32 (31.03)	92 (46.59)
40	61 (2.94e-02)	59 (51.72)	84 (57.58)	32 (44.83)	59 (31.03)	84 (46.21)
41	56 (2.93e-02)	83 (51.61)	15 (55.88)	7 (44.12)	73 (31.03)	90 (45.45)
42	59 (2.90e-02)	84 (51.52)	14 (55.56)	93 (43.48)	11 (30.00)	14 (45.37)
43	54 (2.88e-02)	66 (51.43)	89 (55.26)	69 (43.18)	7 (29.41)	11 (45.00)
44	15 (2.87e-02)	81 (51.35)	67 (54.55)	70 (42.86)	36 (29.03)	7 (44.85)
45	97 (2.86e-02)	15 (50.00)	66 (54.29)	84 (42.42)	78 (29.03)	83 (44.35)
46	60 (2.83e-02)	90 (48.48)	70 (54.29)	90 (42.42)	88 (29.03)	59 (43.97)
47	50 (2.64e-02)	93 (47.83)	49 (52.50)	45 (42.31)	33 (28.57)	93 (43.48)
48	13 (2.55e-02)	85 (47.37)	32 (51.72)	87 (41.67)	95 (28.57)	81 (43.24)
49	71 (2.49e-02)	7 (47.06)	59 (51.72)	59 (41.38)	65 (28.00)	32 (43.10)
50	5 (2.48e-02)	8 (45.95)	90 (51.52)	14 (40.74)	45 (26.92)	8 (42.57)
51	70 (2.37e-02)	67 (45.45)	8 (51.35)	85 (39.47)	85 (26.32)	36 (41.13)
52	7 (2.37e-02)	32 (44.83)	45 (50.00)	51 (38.24)	16 (25.71)	85 (40.13)
53	93 (2.36e-02)	87 (44.44)	93 (47.83)	23 (37.84)	30 (25.00)	49 (38.75)
54	47 (2.26e-02)	52 (42.86)	85 (47.37)	81 (37.84)	49 (25.00)	67 (38.64)
55	66 (2.26e-02)	49 (42.50)	87 (47.22)	1 (36.59)	57 (25.00)	45 (38.46)
56	73 (2.21e-02)	36 (41.94)	51 (47.06)	67 (36.36)	1 (24.39)	97 (38.28)
57	11 (2.18e-02)	13 (40.91)	30 (46.88)	21 (35.71)	66 (22.86)	87 (38.19)
58	19 (2.06e-02)	30 (40.62)	97 (46.88)	49 (35.00)	91 (22.86)	30 (36.72)
59	52 (1.99e-02)	95 (40.00)	52 (45.24)	40 (34.48)	69 (22.73)	51 (36.03)
60	46 (1.98e-02)	61 (39.47)	36 (45.16)	30 (34.38)	47 (22.22)	69 (34.09)

Rank	Sample Variance	FGSM	PGD	C&W	SQUARE	MIXED
61	6 (1.92e-02)	51 (38.24)	40 (44.83)	18 (33.33)	31 (21.88)	57 (33.59)
62	51 (1.89e-02)	40 (37.93)	54 (44.44)	26 (33.33)	97 (21.88)	95 (33.57)
63	68 (1.82e-02)	54 (37.78)	61 (42.11)	92 (31.82)	52 (21.43)	1 (33.54)
64	36 (1.79e-02)	57 (37.50)	13 (40.91)	57 (31.25)	26 (20.83)	40 (32.76)
65	49 (1.76e-02)	97 (37.50)	47 (40.74)	11 (30.00)	51 (20.59)	13 (31.82)
66	23 (1.75e-02)	21 (35.71)	57 (40.62)	17 (29.73)	71 (20.51)	54 (31.67)
67	90 (1.73e-02)	45 (34.62)	17 (40.54)	83 (29.03)	53 (20.45)	21 (31.55)
68	34 (1.69e-02)	73 (34.48)	95 (40.00)	16 (28.57)	87 (19.44)	26 (31.25)
69	21 (1.68e-02)	56 (34.21)	1 (39.02)	39 (28.57)	8 (18.92)	73 (31.03)
70	92 (1.61e-02)	1 (34.15)	99 (38.89)	9 (27.50)	81 (18.92)	52 (30.95)
71	89 (1.56e-02)	69 (34.09)	21 (38.10)	13 (27.27)	56 (18.42)	61 (30.26)
72	67 (1.51e-02)	26 (33.33)	26 (37.50)	54 (26.67)	13 (18.18)	16 (30.00)
73	17 (1.49e-02)	47 (33.33)	16 (37.14)	61 (26.32)	67 (18.18)	71 (28.85)
74	74 (1.45e-02)	71 (33.33)	56 (36.84)	95 (25.71)	86 (18.18)	47 (28.70)
75	30 (1.31e-02)	17 (32.43)	69 (36.36)	58 (25.64)	54 (17.78)	17 (28.38)
76	2 (1.30e-02)	58 (30.77)	58 (35.90)	71 (25.64)	9 (17.50)	56 (28.29)
77	10 (1.26e-02)	16 (28.57)	71 (35.90)	28 (25.00)	76 (17.07)	23 (27.70)
78	31 (1.21e-02)	91 (28.57)	73 (34.48)	29 (24.24)	21 (16.67)	58 (26.92)
79	64 (1.19e-02)	12 (28.12)	39 (33.33)	86 (24.24)	94 (16.28)	91 (26.43)
80	81 (1.09e-02)	99 (27.78)	91 (31.43)	73 (24.14)	23 (16.22)	39 (25.60)
81	42 (1.02e-02)	9 (27.50)	12 (31.25)	56 (23.68)	58 (15.38)	9 (25.00)
82	72 (9.33e-03)	29 (27.27)	28 (30.00)	75 (23.40)	29 (15.15)	99 (24.31)
83	98 (8.49e-03)	23 (27.03)	75 (29.79)	91 (22.86)	39 (14.29)	29 (23.48)
84	80 (8.47e-03)	62 (26.83)	23 (29.73)	12 (21.88)	40 (13.79)	86 (22.73)
85	18 (8.42e-03)	39 (26.19)	62 (29.27)	22 (21.21)	68 (13.64)	12 (22.66)
86	79 (8.18e-03)	94 (25.58)	9 (27.50)	94 (20.93)	61 (13.16)	28 (22.50)
87	78 (8.05e-03)	28 (25.00)	41 (27.50)	41 (20.00)	60 (13.04)	94 (22.09)
88	3 (8.00e-03)	86 (24.24)	22 (27.27)	76 (19.51)	41 (12.50)	75 (21.81)
89	77 (7.57e-03)	75 (23.40)	29 (27.27)	99 (19.44)	22 (12.12)	41 (20.62)
90	88 (7.38e-03)	68 (22.73)	94 (25.58)	47 (18.52)	99 (11.11)	22 (20.45)
91	43 (7.29e-03)	41 (22.50)	86 (24.24)	48 (15.91)	17 (10.81)	76 (20.12)
92	38 (7.25e-03)	76 (21.95)	53 (22.73)	53 (15.91)	75 (10.64)	62 (20.12)
93	65 (5.80e-03)	22 (21.21)	68 (22.73)	68 (15.91)	28 (10.00)	53 (19.89)
94	55 (5.60e-03)	48 (20.45)	76 (21.95)	62 (14.63)	62 (9.76)	68 (18.75)
95	37 (5.47e-03)	53 (20.45)	60 (21.74)	52 (14.29)	12 (9.38)	48 (15.91)
96	44 (5.28e-03)	60 (19.57)	48 (20.45)	24 (13.95)	82 (9.09)	60 (14.67)
97	33 (5.15e-03)	82 (15.91)	24 (18.60)	82 (13.64)	24 (6.98)	82 (13.64)
98	35 (4.53e-03)	24 (11.63)	82 (15.91)	20 (11.36)	48 (6.82)	24 (12.79)
99	27 (4.03e-03)	20 (11.36)	20 (11.36)	0 (5.00)	0 (5.00)	20 (9.66)
100	96 (3.48e-03)	0 (10.00)	0 (10.00)	60 (4.35)	20 (4.55)	0 (7.50)
	Pearson	-0.6486 (3.0e-13)	-0.6573 (1.1e-13)	-0.5419 (5.8e-09)	-0.5095 (6.2e-08)	-0.6201 (6.0e-12)
	Spearman	-0.7314 (5.6e-18)	-0.7437 (7.8e-19)	-0.6524 (1.9e-13)	-0.6256 (3.4e-12)	-0.7251 (1.5e-17)
	Kendall	-0.5403 (1.7e-15)	-0.5484 (6.3e-16)	-0.4630 (8.8e-12)	-0.4475 (4.2e-11)	-0.5300 (5.6e-15)

Table 35: Comparison between sample-covariance ranking and ASR rankings under multiple targeted attacks (I-FGSM, PGD, C&W, and MIXED) when the source class is 0 (Apple) for **CIFAR-100 + Adversarially Trained ResNet-18**. For correlation analysis, the absolute value of the sample covariance is used since it is non-positive while ASR is non-negative.

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
1	53 (-0.005892)	83 (97.50)	83 (97.50)	26 (100.00)	83 (98.33)
2	57 (-0.005256)	10 (90.00)	10 (92.50)	70 (100.00)	10 (91.67)
3	70 (-0.003634)	70 (77.50)	22 (80.00)	83 (100.00)	70 (85.83)
4	92 (-0.003249)	22 (70.00)	70 (80.00)	10 (92.50)	22 (79.17)
5	62 (-0.002044)	53 (70.00)	82 (75.00)	36 (92.50)	26 (78.33)
6	26 (-0.001952)	57 (70.00)	26 (72.50)	57 (92.50)	57 (77.50)
7	45 (-0.001885)	25 (62.50)	36 (72.50)	22 (87.50)	36 (75.00)
8	54 (-0.001630)	26 (62.50)	53 (72.50)	25 (87.50)	53 (75.00)
9	1 (-0.001521)	47 (62.50)	62 (72.50)	32 (87.50)	25 (72.50)
10	84 (-0.001484)	61 (62.50)	57 (70.00)	45 (87.50)	62 (70.83)
11	2 (-0.001368)	62 (62.50)	86 (70.00)	11 (82.50)	32 (69.17)
12	77 (-0.001135)	32 (60.00)	25 (67.50)	53 (82.50)	61 (69.17)
13	36 (-0.001097)	36 (60.00)	47 (67.50)	16 (80.00)	45 (67.50)
14	61 (-0.001073)	82 (60.00)	61 (65.00)	61 (80.00)	47 (66.67)
15	51 (-0.001063)	16 (57.50)	94 (62.50)	62 (77.50)	16 (65.83)
16	7 (-0.001025)	94 (57.50)	11 (60.00)	94 (75.00)	82 (65.83)
17	32 (-0.001003)	45 (55.00)	16 (60.00)	43 (72.50)	11 (65.00)

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
18	25 (-0.000945)	11 (52.50)	32 (60.00)	46 (72.50)	94 (65.00)
19	80 (-0.000891)	13 (52.50)	45 (60.00)	84 (72.50)	86 (61.67)
20	50 (-0.000878)	92 (52.50)	13 (57.50)	92 (72.50)	92 (60.00)
21	35 (-0.000842)	4 (50.00)	6 (55.00)	98 (72.50)	43 (59.17)
22	11 (-0.000815)	43 (50.00)	43 (55.00)	7 (70.00)	46 (56.67)
23	65 (-0.000812)	77 (50.00)	92 (55.00)	47 (70.00)	77 (55.83)
24	78 (-0.000792)	86 (50.00)	77 (52.50)	2 (67.50)	84 (55.83)
25	5 (-0.000788)	2 (47.50)	81 (52.50)	4 (65.00)	4 (55.00)
26	10 (-0.000772)	5 (47.50)	4 (50.00)	77 (65.00)	98 (54.17)
27	98 (-0.000690)	40 (47.50)	5 (50.00)	86 (65.00)	2 (54.17)
28	86 (-0.000689)	46 (47.50)	46 (50.00)	54 (62.50)	6 (53.33)
29	43 (-0.000672)	54 (47.50)	54 (50.00)	82 (62.50)	54 (53.33)
30	64 (-0.000651)	6 (45.00)	84 (50.00)	6 (60.00)	7 (52.50)
31	88 (-0.000648)	63 (45.00)	88 (50.00)	35 (60.00)	63 (50.83)
32	6 (-0.000634)	84 (45.00)	1 (47.50)	63 (60.00)	13 (50.00)
33	34 (-0.000629)	87 (45.00)	2 (47.50)	50 (57.50)	35 (49.17)
34	3 (-0.000616)	1 (42.50)	14 (47.50)	40 (52.50)	40 (49.17)
35	4 (-0.000615)	7 (42.50)	28 (47.50)	65 (52.50)	50 (48.33)
36	14 (-0.000610)	50 (42.50)	35 (47.50)	28 (50.00)	88 (47.50)
37	16 (-0.000601)	81 (42.50)	40 (47.50)	88 (50.00)	5 (46.67)
38	22 (-0.000598)	88 (42.50)	63 (47.50)	93 (50.00)	28 (45.83)
39	18 (-0.000597)	98 (42.50)	87 (47.50)	3 (47.50)	3 (44.17)
40	28 (-0.000589)	3 (40.00)	89 (47.50)	72 (45.00)	87 (43.33)
41	72 (-0.000579)	28 (40.00)	98 (47.50)	78 (45.00)	93 (43.33)
42	19 (-0.000578)	35 (40.00)	3 (45.00)	80 (45.00)	78 (42.50)
43	33 (-0.000569)	14 (37.50)	7 (45.00)	5 (42.50)	81 (41.67)
44	42 (-0.000561)	29 (37.50)	9 (45.00)	15 (42.50)	1 (40.83)
45	55 (-0.000536)	64 (37.50)	20 (45.00)	51 (42.50)	65 (40.83)
46	44 (-0.000531)	78 (37.50)	29 (45.00)	13 (40.00)	29 (39.17)
47	15 (-0.000521)	93 (37.50)	50 (45.00)	64 (40.00)	64 (39.17)
48	47 (-0.000520)	19 (35.00)	78 (45.00)	34 (37.50)	80 (39.17)
49	67 (-0.000519)	37 (35.00)	37 (42.50)	42 (37.50)	15 (38.33)
50	24 (-0.000518)	41 (35.00)	39 (42.50)	55 (37.50)	51 (37.50)
51	93 (-0.000515)	58 (35.00)	90 (42.50)	87 (37.50)	89 (37.50)
52	79 (-0.000498)	65 (35.00)	93 (42.50)	96 (37.50)	34 (36.67)
53	27 (-0.000495)	67 (35.00)	15 (40.00)	21 (35.00)	21 (35.00)
54	99 (-0.000495)	80 (35.00)	19 (40.00)	29 (35.00)	31 (35.00)
55	74 (-0.000484)	89 (35.00)	21 (40.00)	31 (35.00)	67 (35.00)
56	63 (-0.000484)	9 (32.50)	31 (40.00)	1 (32.50)	96 (35.00)
57	38 (-0.000472)	15 (32.50)	34 (40.00)	67 (32.50)	72 (34.17)
58	94 (-0.000469)	18 (32.50)	64 (40.00)	18 (30.00)	37 (34.17)
59	46 (-0.000464)	20 (32.50)	41 (37.50)	81 (30.00)	42 (34.17)
60	9 (-0.000459)	34 (32.50)	42 (37.50)	89 (30.00)	14 (34.17)
61	89 (-0.000456)	39 (32.50)	48 (37.50)	74 (27.50)	19 (33.33)
62	66 (-0.000456)	51 (32.50)	51 (37.50)	85 (27.50)	18 (31.67)
63	81 (-0.000434)	71 (32.50)	58 (37.50)	99 (27.50)	9 (30.83)
64	39 (-0.000427)	90 (32.50)	67 (37.50)	19 (25.00)	90 (30.83)
65	40 (-0.000425)	91 (32.50)	80 (37.50)	33 (25.00)	85 (30.83)
66	29 (-0.000392)	96 (32.50)	56 (35.00)	37 (25.00)	20 (30.00)
67	87 (-0.000392)	21 (30.00)	65 (35.00)	44 (25.00)	39 (30.00)
68	90 (-0.000388)	27 (30.00)	66 (35.00)	79 (25.00)	55 (30.00)
69	37 (-0.000386)	31 (30.00)	85 (35.00)	27 (22.50)	99 (30.00)
70	91 (-0.000379)	33 (30.00)	91 (35.00)	38 (22.50)	91 (29.17)
71	58 (-0.000378)	48 (30.00)	96 (35.00)	66 (22.50)	58 (29.17)
72	23 (-0.000377)	85 (30.00)	99 (35.00)	23 (20.00)	27 (28.33)
73	82 (-0.000371)	42 (27.50)	18 (32.50)	24 (20.00)	33 (28.33)
74	97 (-0.000364)	44 (27.50)	27 (32.50)	91 (20.00)	44 (28.33)
75	48 (-0.000359)	56 (27.50)	44 (32.50)	14 (17.50)	66 (28.33)
76	41 (-0.000342)	66 (27.50)	71 (32.50)	71 (17.50)	71 (27.50)
77	96 (-0.000339)	72 (27.50)	17 (30.00)	90 (17.50)	41 (26.67)
78	31 (-0.000331)	99 (27.50)	23 (30.00)	9 (15.00)	56 (25.83)
79	13 (-0.000330)	8 (25.00)	33 (30.00)	39 (15.00)	23 (25.00)

Rank	Sample Covariance	I-FGSM	PGD	C&W	MIXED
80	56 (-0.000320)	23 (25.00)	72 (30.00)	56 (15.00)	38 (25.00)
81	8 (-0.000314)	38 (25.00)	8 (27.50)	58 (15.00)	79 (25.00)
82	21 (-0.000305)	55 (25.00)	12 (27.50)	20 (12.50)	48 (24.17)
83	20 (-0.000277)	12 (22.50)	24 (27.50)	52 (12.50)	74 (23.33)
84	17 (-0.000275)	17 (22.50)	38 (27.50)	97 (12.50)	24 (22.50)
85	85 (-0.000269)	79 (22.50)	55 (27.50)	69 (10.00)	8 (20.00)
86	73 (-0.000268)	24 (20.00)	69 (27.50)	8 (7.50)	69 (18.33)
87	30 (-0.000242)	74 (20.00)	79 (27.50)	30 (7.50)	97 (18.33)
88	59 (-0.000241)	75 (20.00)	97 (25.00)	41 (7.50)	17 (17.50)
89	12 (-0.000231)	30 (17.50)	49 (22.50)	59 (7.50)	12 (17.50)
90	69 (-0.000222)	59 (17.50)	59 (22.50)	48 (5.00)	59 (15.83)
91	75 (-0.000202)	69 (17.50)	73 (22.50)	73 (5.00)	30 (15.00)
92	71 (-0.000201)	97 (17.50)	74 (22.50)	76 (5.00)	52 (14.17)
93	68 (-0.000180)	49 (15.00)	30 (20.00)	12 (2.50)	49 (13.33)
94	49 (-0.000175)	52 (15.00)	75 (20.00)	49 (2.50)	73 (13.33)
95	52 (-0.000160)	68 (15.00)	76 (20.00)	68 (2.50)	75 (13.33)
96	60 (-0.000147)	73 (12.50)	68 (17.50)	17 (0.00)	76 (12.50)
97	83 (-0.000146)	76 (12.50)	52 (15.00)	60 (0.00)	68 (11.67)
98	95 (-0.000144)	60 (7.50)	60 (15.00)	75 (0.00)	60 (7.50)
99	76 (-0.000143)	95 (7.50)	95 (7.50)	95 (0.00)	95 (5.00)
	Pearson	0.5118 (6.2e-08)	0.4702 (9.1e-07)	0.5467 (4.8e-09)	0.5333 (1.3e-08)
	Spearman	0.6792 (1.1e-14)	0.6368 (1.4e-12)	0.7709 (1.0e-20)	0.7327 (6.7e-18)
	Kendall	0.5302 (7.5e-15)	0.4902 (6.5e-13)	0.6080 (4.8e-19)	0.5694 (6.8e-17)

D.5 The Choice of Ideal-Centered Second-Order Statistics in SOIV

In the main paper we adopt the ideal-centered formulation as the definition of SOIV. On CIFAR-10 with ResNet-18, both the ideal-centered statistics and the conditional sample statistics produce similar structural patterns and correlations with ASR. On the adversarially trained CIFAR-100 models, however, the sample-centered statistics no longer track the vulnerability ordering in the untargeted setting.

The main reason for this choice is conceptual consistency. In sample-based statistics, variance is measured around the conditional mean confidence, which may vary across classes depending on dataset difficulty or model calibration. Consequently, the baseline for measuring variability becomes class-dependent. In contrast, ideal-centered statistics measure deviations from the ideal prediction confidence (i.e., probability equal to 1), providing a unified reference point across all classes.

This formulation is also naturally aligned with the objective of adversarial attacks. Since most attacks aim to reduce the probability assigned to the true class, measuring the deviation from the ideal confidence level provides an interpretable measure of how unstable the model's prediction becomes under perturbations. For these reasons, the ideal-centered formulation provides a conceptually consistent and interpretable basis for defining SOIV.

D.6 Sensitivity to Temperature Scaling

Since SOIV is computed from clean softmax probabilities, changes in the softmax confidence scale may affect the resulting vulnerability estimates. To examine this sensitivity, we conduct a post-hoc temperature scaling analysis on **CIFAR-10 + ResNet-18**. Given the original logits z, temperature scaling computes the predictive probability as

$$p_i^{(T)} = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)},$$

where $T < 1$ sharpens the predictive distribution, whereas $T > 1$ flattens it. Since positive temperature scaling preserves the ordering of the logits, it does not change the predicted class. We therefore keep the attack-based ASRs fixed and recompute SOIV from the temperature-scaled clean probabilities.

We first present the detailed class-wise results at $T = 0.5$ as a representative confidence-sharpening setting. Table 36 compares the SOIV-U ranking with ASR rankings under multiple untargeted at-

Table 31: Comparison between sample-covariance ranking and ASR rankings under multiple targeted attacks (FGSM, PGD, C&W, and MIXED) when the source class is *automobile* for **CIFAR-10 + ResNet-18**. For correlation analysis, the absolute value of the sample covariance is used since it is non-positive while ASR is non-negative.

Rank	Sample Covariance	FGSM	PGD	C&W	MIXED
1	truck (-0.002723)	truck (65.78)	truck (57.78)	truck (64.93)	truck (62.83)
2	ship (-0.001111)	ship (25.27)	ship (31.24)	ship (37.42)	ship (31.31)
3	horse (-0.000412)	frog (17.38)	airplane (25.37)	airplane (23.67)	airplane (22.00)
4	airplane (-0.000381)	airplane (16.95)	frog (19.08)	frog (23.35)	frog (19.94)
5	frog (-0.000153)	bird (10.77)	bird (15.78)	cat (14.82)	bird (13.58)
6	dog (-0.000112)	horse (7.89)	cat (15.78)	bird (14.18)	cat (12.65)
7	cat (-0.000102)	cat (7.36)	horse (12.79)	dog (13.01)	horse (10.98)
8	bird (-0.000083)	deer (7.25)	deer (12.69)	horse (12.26)	dog (10.84)
9	deer (-0.000024)	dog (6.82)	dog (12.69)	deer (10.45)	deer (10.13)
Pearson		0.972451 (1.1e-05)	0.968653 (1.7e-05)	0.966364 (2.2e-05)	0.973555 (9.6e-06)
Spearman		0.733333 (2.5e-02)	0.722715 (2.8e-02)	0.700000 (3.6e-02)	0.733333 (2.5e-02)
Kendall		0.555556 (4.5e-02)	0.628828 (1.8e-02)	0.611111 (2.5e-02)	0.611111 (2.5e-02)

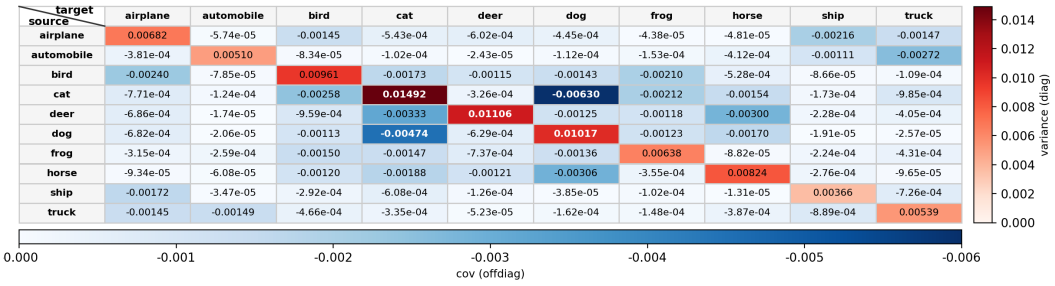


Figure 7: Unified variance–covariance matrix computed from conditional sample second-order statistics on **CIFAR-10 + ResNet-18**.

Table 36: Comparison between SOIV-U ranking and ASR ranking under multiple untargeted attacks (FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED) for **CIFAR-10 + ResNet-18** with temperature scaling at $T = 0.5$.

Rank	SOIV-U	FGSM	PGD	C&W	SQUARE	AUTOATTACK	MIXED
1	cat (0.007965)	cat (0.750663)	cat (0.782493)	cat (0.755968)	cat (0.866048)	cat (0.811671)	cat (0.793369)
2	deer (0.005699)	deer (0.620413)	deer (0.675459)	deer (0.732798)	deer (0.780963)	deer (0.711009)	deer (0.704128)
3	dog (0.005102)	dog (0.608748)	dog (0.625759)	dog (0.620899)	dog (0.718104)	dog (0.637910)	dog (0.642284)
4	horse (0.004729)	bird (0.513189)	bird (0.557554)	bird (0.588729)	bird (0.666667)	bird (0.587530)	bird (0.582734)
5	airplane (0.003506)	frog (0.455135)	frog (0.508108)	frog (0.534054)	frog (0.607568)	frog (0.526487)	frog (0.526270)
6	bird (0.003476)	horse (0.409890)	airplane (0.430449)	airplane (0.443593)	airplane (0.553122)	airplane (0.452355)	airplane (0.455860)
7	truck (0.002752)	airplane (0.399781)	horse (0.424176)	horse (0.407692)	horse (0.523077)	horse (0.441758)	horse (0.441319)
8	frog (0.002690)	truck (0.337323)	truck (0.339499)	ship (0.357143)	ship (0.465885)	ship (0.364606)	ship (0.365672)
9	automobile (0.002196)	automobile (0.312367)	ship (0.337953)	truck (0.321001)	truck (0.439608)	truck (0.356910)	truck (0.358868)
10	ship (0.001278)	ship (0.302772)	automobile (0.303838)	automobile (0.281450)	automobile (0.373134)	automobile (0.317697)	automobile (0.317697)
Pearson		0.911871 (2.4e-04)	0.878695 (8.2e-04)	0.813341 (4.2e-03)	0.853558 (1.7e-03)	0.873273 (9.7e-04)	0.869136 (1.1e-03)
Spearman		0.866667 (1.2e-03)	0.842424 (2.2e-03)	0.806061 (4.9e-03)	0.806061 (4.9e-03)	0.806061 (4.9e-03)	0.806061 (4.9e-03)
Kendall		0.777778 (9.5e-04)	0.688889 (4.7e-03)	0.644444 (9.1e-03)	0.644444 (9.1e-03)	0.644444 (9.1e-03)	0.644444 (9.1e-03)

tacks. SOIV-U identifies *cat*, *deer*, and *dog* as the three most vulnerable classes, which are also ranked as the top three classes under FGSM, PGD, C&W, SQUARE, AUTOATTACK, and MIXED. The correlations with MIXED ASR are 0.869, 0.806, and 0.644 for Pearson, Spearman, and Kendall, respectively. Positive correlations are also consistently observed for the individual attacks.

Table 37 presents the corresponding SOIV-T analysis at $T = 0.5$, using *automobile* as the source class. SOIV-T ranks *truck* and *ship* as the two most vulnerable target classes, and these two classes are also ranked first and second under FGSM, PGD, C&W, and MIXED targeted ASR. The Pearson, Spearman, and Kendall correlations between SOIV-T and MIXED ASR are 0.927, 0.833, and 0.778, respectively. These results show that the directional vulnerability structure identified by SOIV-T remains strongly aligned with the attack-based target-class rankings in this setting.

To examine the overall trend across different confidence scales, we additionally evaluate $T \in \{0.5, 0.75, 1.0, 1.5, 2.0\}$. Rather than repeating the complete class-wise ranking tables for every

Table 37: Comparison between SOIV-T ranking and targeted ASR rankings under multiple attacks (FGSM, PGD, C&W, and MIXED) when the source class is *automobile* for **CIFAR-10 + ResNet-18** with temperature scaling at $T = 0.5$.

Rank	SOIV-T	FGSM	PGD	C&W	MIXED
1	truck (-0.001335)	truck (0.657783)	truck (0.577825)	truck (0.649254)	truck (0.628287)
2	ship (-0.000426)	ship (0.252665)	ship (0.312367)	ship (0.374200)	ship (0.313077)
3	horse (-0.000346)	frog (0.173774)	airplane (0.253731)	airplane (0.236674)	airplane (0.219972)
4	airplane (-0.000066)	airplane (0.169510)	frog (0.190832)	frog (0.233475)	frog (0.199360)
5	frog (-0.000010)	bird (0.107676)	bird (0.157783)	cat (0.148188)	bird (0.135750)
6	bird (-0.000006)	horse (0.078891)	cat (0.157783)	bird (0.141791)	cat (0.126510)
7	cat (-0.000003)	cat (0.073561)	horse (0.127932)	dog (0.130064)	horse (0.109808)
8	dog (-0.000003)	deer (0.072495)	deer (0.126866)	horse (0.122601)	dog (0.108387)
9	deer (-0.000000)	dog (0.068230)	dog (0.126866)	deer (0.104478)	deer (0.101279)
Pearson		0.938349 (1.8e-04)	0.916986 (5.0e-04)	0.912724 (5.9e-04)	0.927485 (3.1e-04)
Spearman		0.866667 (2.5e-03)	0.823558 (6.4e-03)	0.733333 (2.5e-02)	0.833333 (5.3e-03)
Kendall		0.722222 (5.9e-03)	0.743161 (5.3e-03)	0.666667 (1.3e-02)	0.777778 (2.4e-03)

Table 38: Correlation between SOIV-U/SOIV-T and MIXED ASR under different temperature scaling settings for **CIFAR-10 + ResNet-18**.

Metric	$T = 0.50$	$T = 0.75$	$T = 1.00$	$T = 1.50$	$T = 2.00$
SOIV-U Pearson	0.869 (1.1e-03)	0.920 (1.6e-04)	0.927 (1.1e-04)	0.911 (2.5e-04)	0.877 (8.6e-04)
SOIV-U Spearman	0.806 (4.9e-03)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.867 (1.2e-03)	0.818 (3.8e-03)
SOIV-U Kendall	0.644 (9.1e-03)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.644 (9.1e-03)	0.600 (1.7e-02)
SOIV-T Pearson	0.927 (3.1e-04)	0.957 (5.2e-05)	0.975 (7.8e-06)	0.989 (4.5e-07)	0.991 (2.3e-07)
SOIV-T Spearman	0.833 (5.3e-03)	0.750 (2.0e-02)	0.733 (2.5e-02)	0.850 (3.7e-03)	0.917 (5.0e-04)
SOIV-T Kendall	0.778 (2.4e-03)	0.667 (1.3e-02)	0.611 (2.5e-02)	0.722 (5.9e-03)	0.833 (8.5e-04)

temperature, we report the detailed rankings only for the representative $T = 0.5$ setting in Tables 36 and 37, and summarize the correlations with MIXED ASR for the remaining temperatures in Table 38.

As shown in Table 38, both SOIV-U and SOIV-T maintain positive correlations with MIXED ASR throughout the tested temperature range. For SOIV-U, Pearson correlations range from 0.869 to 0.927, Spearman correlations from 0.806 to 0.915, and Kendall correlations from 0.600 to 0.778. SOIV-T also maintains consistently positive correlations across the evaluated settings, with Pearson correlations ranging from 0.927 to 0.991.

These results indicate that, within the temperature range examined here, the relationship between SOIV and the corresponding adversarial vulnerability rankings is generally maintained under both confidence sharpening and moderate confidence flattening. In particular, SOIV-T exhibits stable directional alignment across the evaluated temperatures, while SOIV-U also maintains consistent class-wise alignment. However, this analysis does not establish invariance to arbitrary calibration transformations or stronger confidence rescaling. Accordingly, SOIV is primarily interpreted as an intra-model metric for relative class-wise and directional vulnerability analysis rather than as a calibration-invariant absolute score for direct inter-model comparison.

D.7 Sensitivity to Attack Strength

The theoretical motivation in the main paper Section 3.4 relies on a first-order Taylor approximation, which is valid for small perturbations. To examine how the alignment between SOIV-U and class-wise ASR changes with the attack strength, we evaluate a range of strengths on CIFAR-10 + ResNet-18. Table 39 reports the correlations together with the standard deviation of class-wise ASR at each strength.

The correlations generally decrease as the strength grows, and for the iterative attacks, this largely tracks the loss of class-wise variation in ASR rather than a breakdown of the metric. At the largest strength, PGD and C&W reach 100% ASR for every class. The standard deviation of class-wise ASR then falls to zero, leaving no ordering to compare against, so the correlation is undefined. SOIV is therefore informative only where class-wise vulnerability remains differentiated, and the

Table 39: Correlation between SOIV-U and class-wise ASR across different attack strengths for **CIFAR-10 + ResNet-18**. The C&W column correspond to $c = 0.01, 0.1, 1, 10$ with $\kappa = 3$. A dash indicates an undefined correlation, since every class reaches 100% ASR.

	FGSM	PGD	C&W	SQUARE	MIXED
Pearson					
$\epsilon = 2/255$	0.959 (1.2e-05)	0.901 (3.7e-04)	0.877 (8.6e-04)	0.961 (9.7e-06)	0.931 (9.1e-05)
$\epsilon = 4/255$	0.912 (2.4e-04)	0.752 (1.2e-02)	0.506 (1.4e-01)	0.917 (1.9e-04)	0.874 (9.4e-04)
$\epsilon = 8/255$	0.835 (2.6e-03)	0.470 (1.7e-01)	0.312 (3.8e-01)	0.789 (6.7e-03)	0.817 (3.9e-03)
$\epsilon = 16/255$	0.664 (3.6e-02)	–	–	0.630 (5.1e-02)	0.684 (2.9e-02)
Spearman					
$\epsilon = 2/255$	0.964 (7.0e-06)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)	0.915 (2.1e-04)
$\epsilon = 4/255$	0.952 (2.2e-05)	0.806 (4.9e-03)	0.649 (4.2e-02)	0.915 (2.1e-04)	0.927 (1.1e-04)
$\epsilon = 8/255$	0.891 (5.4e-04)	0.620 (5.6e-02)	0.351 (3.2e-01)	0.855 (1.6e-03)	0.891 (5.4e-04)
$\epsilon = 16/255$	0.455 (1.9e-01)	–	–	0.802 (5.3e-03)	0.539 (1.1e-01)
Kendall					
$\epsilon = 2/255$	0.911 (3.0e-05)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)	0.778 (9.5e-04)
$\epsilon = 4/255$	0.867 (1.2e-04)	0.689 (4.7e-03)	0.511 (4.7e-02)	0.778 (9.5e-04)	0.822 (3.6e-04)
$\epsilon = 8/255$	0.733 (2.2e-03)	0.450 (7.0e-02)	0.304 (2.2e-01)	0.689 (4.7e-03)	0.733 (2.2e-03)
$\epsilon = 16/255$	0.289 (2.9e-01)	–	–	0.584 (1.9e-02)	0.378 (1.6e-01)
ASR std (%)					
$\epsilon = 2/255$	14.24	14.46	16.04	12.82	14.26
$\epsilon = 4/255$	13.40	5.95	3.72	15.20	9.30
$\epsilon = 8/255$	8.92	1.01	0.07	6.48	4.00
$\epsilon = 16/255$	6.36	0.00	0.00	0.57	1.67

same holds for attack-based diagnosis. Accordingly, the attack strengths used throughout the paper are chosen so that class-wise ASRs remain spread out.

D.8 Sensitivity to Data Split and Random Seed

To examine whether the alignment depends on a particular data split or random seed, we re-train the model in the primary CIFAR-10 + ResNet-18 setting under a 3×3 grid of split seeds $\{1, 521, 884126\}$ and training seeds $\{0, 1, 2\}$. The 45,000/5,000 train/validation ratio, the training configuration, and the attack configuration are kept unchanged. The split seed determines which images form the validation set on which SOIV is computed, and the training seed determines the weight initialization, mini-batch order, and data augmentation. Table 40 reports the correlations of SOIV-U and SOIV-T with MIXED ASR for each model.

Across the nine models, the Pearson correlation with MIXED ASR is at least 0.817 for SOIV-U and 0.971 for SOIV-T. For the rank correlations, Spearman is at least 0.818 for SOIV-U and 0.783 for SOIV-T, and Kendall at least 0.644 and 0.592, respectively. The standard deviation across the nine models is at most 0.047 for SOIV-U and 0.114 for SOIV-T. These values are close to those of the main experiments in Tables 1 and 2.

Table 40: Correlations of SOIV-U and SOIV-T with MIXED ASR for **CIFAR-10 + ResNet-18** under different split and training seeds.

Split seed	Training seed	SOIV-U			SOIV-T		
		Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
1	0	0.947 (3.2e-05)	0.891 (5.4e-04)	0.778 (9.5e-04)	0.994 (5.5e-08)	0.933 (2.4e-04)	0.833 (8.5e-04)
1	1	0.890 (5.6e-04)	0.818 (3.8e-03)	0.644 (9.1e-03)	0.992 (1.5e-07)	0.933 (2.4e-04)	0.833 (8.5e-04)
1	2	0.928 (1.1e-04)	0.891 (5.4e-04)	0.689 (4.7e-03)	0.971 (1.3e-05)	0.783 (1.3e-02)	0.611 (2.5e-02)
521	0	0.854 (1.7e-03)	0.915 (2.1e-04)	0.778 (9.5e-04)	0.980 (3.7e-06)	0.933 (2.4e-04)	0.833 (8.5e-04)
521	1	0.827 (3.2e-03)	0.903 (3.4e-04)	0.733 (2.2e-03)	0.973 (1.0e-05)	0.787 (1.2e-02)	0.592 (2.6e-02)
521	2	0.830 (3.0e-03)	0.891 (5.4e-04)	0.778 (9.5e-04)	0.972 (1.2e-05)	0.933 (2.4e-04)	0.833 (8.5e-04)
884126	0	0.875 (9.2e-04)	0.915 (2.1e-04)	0.778 (9.5e-04)	0.978 (5.1e-06)	0.817 (7.2e-03)	0.611 (2.5e-02)
884126	1	0.817 (3.9e-03)	0.891 (5.4e-04)	0.733 (2.2e-03)	0.982 (2.5e-06)	0.833 (5.3e-03)	0.667 (1.3e-02)
884126	2	0.860 (1.4e-03)	0.879 (8.1e-04)	0.733 (2.2e-03)	0.985 (1.3e-06)	0.933 (2.4e-04)	0.833 (8.5e-04)
Mean		0.870	0.888	0.738	0.981	0.876	0.739
Std		0.045	0.029	0.047	0.009	0.069	0.114