Robustness Meets Uncertainty: Evidential Adversarial Training for Robust Selective Classification

Nicolas Sournac[†], Ahmed Baha Ben Jmaa[†], and Bertrand Braeckeveldt

Multitel Research & Innovation Center, Artificial Intelligence Department, Belgium
TRAIL – Trusted AI Labs, Belgium

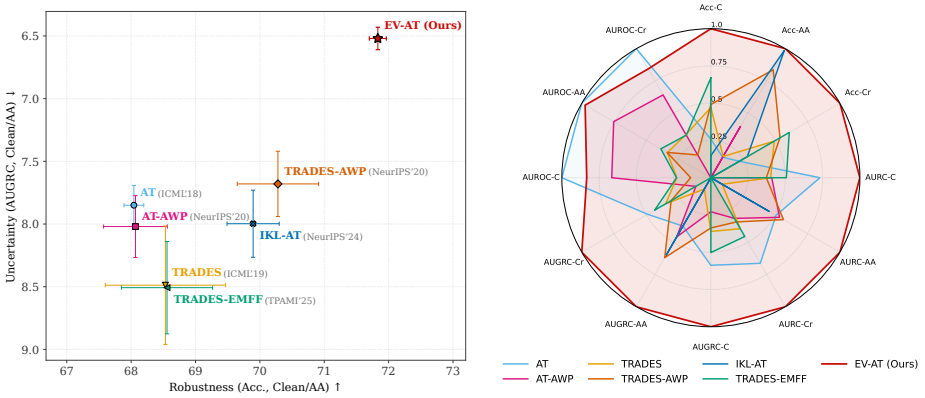
Abstract. Safety-critical applications require classifiers that are both robust and reliable. Adversarial training is a widely adopted defense for improving robustness in deep neural networks; however, its effect on the reliability of predictive uncertainty remains underexplored. We investigate this gap through the lens of selective classification, which has rarely been systematically analyzed alongside adversarial robustness. We introduce a unified benchmark for the robustness–uncertainty trade-off. It standardizes architectures, augmentations, threat models, and evaluation metrics across clean, adversarial, and common-corruption settings. Across a wide range of state-of-the-art adversarial training methods, we uncover a recurring failure mode: several approaches improve robust accuracy while degrading uncertainty ranking, leading to poorer selective behavior. To address this, we propose Evidential Adversarial Training (EV-AT), which models uncertainty through a Dirichlet distribution and combines (i) an evidence-based loss promoting clean accuracy and reliable uncertainty with (ii) a robust evidence-alignment loss matching clean and adversarial predictions in log Dirichlet-parameter space. Extensive experiments show that EV-AT shifts the Pareto frontier of robustness–uncertainty trade-offs beyond prior state-of-the-art adversarial training methods. Our source code is publicly available at `https://github.com/NicolasSournac/Robustness_Meets_Uncertainty.EV-AT`.

Keywords: Adversarial robustness · Adversarial training · Uncertainty estimation · Selective classification · Evidential deep learning

1 Introduction

Deep neural networks have become the backbone of modern vision systems, delivering remarkable performance across recognition tasks [24, 55]. In safety-critical deployments, recent efforts in trustworthy machine learning emphasize that reliable systems must be both robust to perturbations and uncertainty-aware [5, 19]. Yet their deployment in such settings remains constrained by two tightly coupled weaknesses: susceptibility to adversarial perturbations [43] and unreliable

[†] These authors contributed equally to this work.



(a) Robustness–uncertainty trade-off. **(b)** Multi-metric comparison.

Fig. 1: Robustness–uncertainty trade-off on CIFAR-10 with WRN-34-10. EV-AT shifts the Pareto frontier toward higher robustness and lower uncertainty. **(a)** The x-axis reports *robustness* as the average accuracy over clean and AutoAttack (AA) (higher is better), and the y-axis reports *uncertainty* as the average AUGRC over clean and AA (lower is better; the y-axis is inverted so better performance appears higher). Each point corresponds to a method's mean over four data augmentations (Cutout, Basic, AutoAug, AugMix), and error bars denote the standard error across these augmentations. **(b)** Radar comparison of robustness and uncertainty metrics across clean, adversarial, and common-corruption settings. Each point corresponds to a method's mean over four data augmentations and three random seeds.

confidence estimates [27, 29]. Imperceptible input changes can induce incorrect predictions, often persisting in restricted-access and transfer settings, making adversarial robustness a central concern. Adversarial training has therefore emerged as the dominant paradigm to improve robustness [31, 34, 44]. However, robustness alone is insufficient for risk-sensitive decision making: when errors are costly, a system must also know when not to predict. This requirement is naturally captured by selective classification, where the model can abstain on inputs it considers unreliable [40]. Selective prediction is a pragmatic interface for safety, deferring to a fallback policy when uncertain. Crucially, selective behavior depends not only on accuracy under attack, but on the integrity of uncertainty under perturbations: errors, especially adversarial ones, must be assigned high uncertainty so they can be rejected [19, 40]. Robustness evaluations [11], which primarily report robust accuracy, can hide a damaging failure mode: a method may become harder to fool while simultaneously degrading uncertainty ranking, producing confident adversarial mistakes that bypass rejection and manifest as silent failures at practically relevant operating points.

A second challenge is methodological. Both robustness and uncertainty are highly sensitive to experimental confounders, making it difficult to draw controlled conclusions about the robustness–uncertainty trade-off. To enable principled comparisons, we introduce a unified benchmark that standardizes these

factors and evaluates models as selective systems across clean, adversarial, and common-corruption regimes. Across a broad set of state-of-the-art adversarial training methods, this benchmark reveals a consistent pattern: robustness gains do not reliably translate to robust uncertainty, and several strong defenses improve adversarial accuracy while worsening risk-coverage behavior due to degraded uncertainty ranking.

Motivated by these findings, we introduce Evidential Adversarial Training (EV-AT), a novel adversarial-training objective that explicitly targets predictive uncertainty and regularizes the predictive posterior under attack, without incurring additional computational overhead. EV-AT represents class predictions using a Dirichlet distribution, whose parameters jointly encode (i) the predictive mean used for classification and (ii) the total concentration governing evidential strength and confidence. Rather than aligning only the logits, EV-AT promotes distribution-level robustness by explicitly constraining adversarial drift in log-Dirichlet-parameter space.

Contributions. Our key contributions are summarized as follows:

1. We introduce a standardized benchmark for evaluating the trade-off between robustness and uncertainty in selective classification, using unified and reproducible protocols across training methods, architectures, and data augmentations.
2. We propose **Evidential Adversarial Training (EV-AT)**, a new adversarial training objective that learns Dirichlet distributions to jointly improve robustness and uncertainty quality, thereby enabling more reliable selective classification under adversarial perturbations.
3. Extensive experiments across datasets and threat models show that EV-AT improves adversarial robustness while enhancing uncertainty quality over state-of-the-art adversarial training baselines.

Paper Organization. Section 2 reviews prior work, Sec. 3 introduces EV-AT and its underlying principles, Sec. 4 presents the benchmark and experimental evaluation, and Sec. 5 concludes the paper.

2 Related Work

Adversarial Attacks. Adversarial attacks add small, ℓ_p-bounded perturbations to induce misclassification [43]. Standard gradient-based attacks such as FGSM and projected gradient descent (PGD) [21, 34] generate such perturbations through single-step and iterative updates, respectively. Since robustness can be overestimated due to weak attack settings or gradient masking [4, 8], AutoAttack (AA) [12], a strong parameter-free ensemble evaluation framework, provides a reliable assessment. Benchmarks such as RobustBench [11] and AttackBench [10] promote consistent evaluation.

Adversarial Defenses. Adversarial training (AT) minimizes the worst-case risk over bounded input perturbations, commonly approximated using PGD [31,

34,44]. Although effective, AT often induces a trade-off between clean and robust accuracy. TRADES [57] addresses this trade-off by balancing clean classification performance with consistency between clean and adversarial predictions, while MART [49], AWP [54], SEAT [47], and Generalist [48] improve robust generalization through example-aware optimization, weight perturbation, self-ensembling, and decoupled learners, respectively. Complementary directions improve robustness through augmentation with synthetic data [6, 22, 50] and refined training objectives, including Improved KL divergence (IKL) [14] and Evidence-based Multi-Feature Fusion (EMFF) [51].

Uncertainty Estimation. Uncertainty methods broadly fall into sampling-based and sampling-free paradigms [19]. Sampling-based approaches approximate a posterior over parameters or functions but often require multiple evaluations [17, 23, 33, 38, 52]. Sampling-free alternatives provide single-pass estimates through conformal prediction [3, 42], representation distances for out-of-distribution (OOD) detection [2, 46], or directly learned predictive distributions [18, 41].

Evidential Deep Learning (EDL). EDL parameterizes a higher-order distribution (e.g., a Dirichlet) to represent evidence over class probabilities and has been applied beyond classification [9, 18, 41]. Recent theory highlights limitations of learning second-order distributions (e.g., absence of strictly proper scoring rules), implying potential unfaithful estimates [7, 28]. Nevertheless, EDL uncertainties are often well-ranked in practice, making them effective for selective prediction and OOD detection [1, 28].

Robustness–Uncertainty Interaction. Robustness and uncertainty are often studied separately despite being closely related [5, 19]. Under adversarial perturbations, conformal and representation-distance methods may become unreliable, motivating robust variants and additional smoothness constraints [2, 20, 27, 36, 39, 56]. Conventional and EDL classifiers can also produce overconfident adversarial errors [29]. Yet the effect of adversarial training on uncertainty ranking and selective performance remains comparatively underexplored [15, 32].

3 Proposed Approach: EV-AT

3.1 Problem formulation and overview

We consider supervised C-class classification with selective prediction under ℓ_p-bounded adversarial perturbations. Let P be a distribution over the input-target space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, \dots, C\}$. We consider a parametric predictor $f_\theta : \mathcal{X} \to \Delta^{C-1}$ that maps inputs to the probability simplex. This predictor induces a classification rule $\hat{y} = \arg\max_{c \in \mathcal{Y}} f_c(\mathbf{x}; \theta)$ and is equipped with an uncertainty estimator $u : \mathcal{X} \to \mathbb{R}$. This score is compared against a threshold $\tau \in \mathbb{R}$ to define the selective rule $g_\tau(\mathbf{x}; u) = \mathbb{1}[u(\mathbf{x}) \leq \tau]$, which determines whether a prediction is accepted or rejected. Adversarial examples $\mathbf{x}^{\text{adv}}$ are generated by an explicit attack procedure and constrained to the ℓ_p-ball $\mathcal{B}_p(\mathbf{x}; \varepsilon) = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_p \leq \varepsilon\}$.

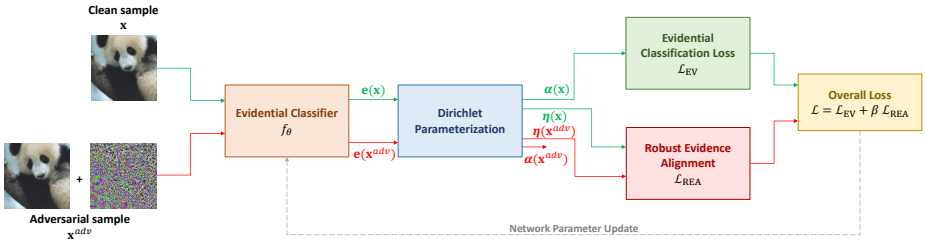


Fig. 2: Overview of the proposed evidential adversarial training (EV-AT) framework. Given a clean sample $\mathbf{x}$, the evidential classifier f_θ produces evidence $e(\mathbf{x})$, which is converted into Dirichlet parameters for prediction and uncertainty estimation. An adversarial sample $\mathbf{x}^{adv}$ is generated from $\mathbf{x}$ using K-step PGD within an ε-bounded perturbation set, where the attack maximizes the divergence in the evidence space. Training combines an evidential classification loss $\mathcal{L}_{\text{EV}}$ for clean samples and a robust evidence alignment loss $\mathcal{L}_{\text{REA}}$ between clean and adversarial evidential representations. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{EV}} + \beta\mathcal{L}_{\text{REA}}$. Green arrows denote the clean path, while red arrows denote the adversarial path.

The objective is to learn the predictor parameters θ that simultaneously maximize **(i)** accuracy on clean data, **(ii)** robustness of the predicted label within $\mathcal{B}_p(\mathbf{x}, \varepsilon)$, and **(iii)** integrity of uncertainty under perturbations. Specifically, label-preserving perturbations should remain confidently accepted, whereas perturbations that induce errors should not become confident errors that remain accepted and should instead be ranked as more uncertain. We achieve this objective by minimizing adversarial selective risk at a target adversarial coverage level $\gamma_{\texttt{cov}} \in [0, 1]$, i.e., $\min_\theta \ \text{risk}^{\text{adv}}(\theta; \tau) \quad \text{s.t.} \quad \text{cov}^{\text{adv}}(\tau; \theta) \geq \gamma_{\texttt{cov}}$.

Adversarial training addresses **(i)** and **(ii)** by minimizing a worst-case classification loss over the ℓ_p-ball, but it does not explicitly enforce **(iii)**. As a result, robust-accuracy gains can coincide with degraded uncertainty ranking, increasing selective risk at relevant coverages. EV-AT instead regularizes uncertainty under adversarial perturbations via posterior-level robustness. We model predictions with a Dirichlet posterior and penalize adversarial drift by aligning clean and adversarial evidential representations in log-parameter space. This jointly stabilizes the predictive mean (labels) and concentration (confidence/abstention). Figure 2 illustrates the method, while the corresponding training procedure is detailed in Algorithm S1 of the supplementary material.

3.2 Evidential classifier

To treat uncertainty as a primary learning objective, we utilize an evidential framework where the network outputs concentration parameters $\boldsymbol{\alpha}$ of a Dirichlet distribution. This parameterization provides a structured representation that naturally decomposes into **(i)** a predictive mean for classification, and **(ii)** a total concentration, which serves as a principled confidence proxy for selective prediction.

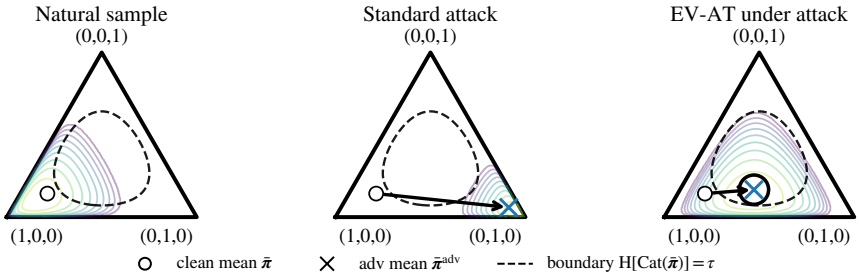


Fig. 3: Posterior-level view of selective robustness (3-class illustration). Dirichlet posteriors $\mathrm{Dir}(\boldsymbol{\alpha})$ are shown on the simplex. The clean mean $\bar{\boldsymbol{\pi}}$ (circle) and adversarial mean $\bar{\boldsymbol{\pi}}^{\mathrm{adv}}$ (cross) illustrate attack-induced drift; the dashed curve is the entropy accept/reject boundary $\mathrm{H}[\mathrm{Cat}(\bar{\boldsymbol{\pi}})] = \tau$. Standard attacks can move predictions to a wrong corner while staying accepted, whereas EV-AT pushes adversarial inputs toward higher-entropy regions, increasing rejection and reducing selective risk.

Dirichlet parameterization. For a given input $\mathbf{x} \in \mathcal{X}$, the network outputs a non-negative evidence vector $\mathbf{e}_\theta(\mathbf{x})$. It parameterizes a Dirichlet distribution over the categorical probabilities $\boldsymbol{\pi} \in \Delta^{C-1}$ via the concentration parameters $\boldsymbol{\alpha} = \mathbf{e}_\theta(\mathbf{x}) + \mathbf{1}$. This formulation induces a posterior predictive distribution $\mathrm{Cat}(\bar{\boldsymbol{\pi}})$, where the predictive mean $\bar{\boldsymbol{\pi}} = \boldsymbol{\alpha}/S$ corresponds to the Dirichlet expectation. Here, $S = \sum_{c=1}^{C} \alpha_c$ represents the total concentration (or strength), whose magnitude inversely quantifies the model uncertainty. The posterior mean $\bar{\boldsymbol{\pi}}$ is the standard choice for prediction, with the predicted label given by $\hat{y} = \arg\max_{c \in \mathcal{Y}} \bar{\pi}_c$.

Uncertainty and selective prediction. The evidential framework treats the Dirichlet strength S as a principled proxy for epistemic uncertainty. A large strength S yields a concentrated, confident posterior over $\boldsymbol{\pi}$, whereas a small S results in a diffuse distribution reflecting high uncertainty. Total predictive uncertainty is quantified by the entropy of the posterior predictive distribution as

$$u(\mathbf{x}) = \mathrm{H}[\mathrm{Cat}(\bar{\boldsymbol{\pi}})] = -\sum_{c=1}^{C} \frac{\alpha_c}{S} \log \frac{\alpha_c}{S}. \tag{1}$$

This score enables selective classification by rejecting predictions that exceed a predefined threshold τ. The resulting selective rule $g_\tau(\mathbf{x}; u) = \mathbb{1}[u(\mathbf{x}) \leq \tau]$ accepts a sample only if its uncertainty is sufficiently bounded.

3.3 Evidential Adversarial Training

EV-AT follows a min-max formulation: for each input $\mathbf{x}$, we generate an adversarial example $\mathbf{x}^{\mathrm{adv}}$ that maximally perturbs the evidential representation, then update the model to **(i)** fit clean labels with an evidential loss and **(ii)** minimize the resulting worst-case clean-adversarial drift. Following Sec. 3.4, we operate in

log-concentration space $\boldsymbol{\eta} = \log(\boldsymbol{\alpha})$, which acts as a "pseudo-logit" parameterization: additive changes in $\boldsymbol{\eta}$ correspond to multiplicative changes in $\boldsymbol{\alpha}$. This provides a well-scaled notion of drift that captures shifts in both the predictive mean and the evidence strength.

Evidence Learning. We first train an evidential predictor on clean data. Given $(\mathbf{x}, y)$, we learn Dirichlet concentrations $\boldsymbol{\alpha}$ so that the predictive mean $\bar{\boldsymbol{\pi}}$ matches y while discouraging spurious evidence. Following EDL [41], we use the negative log marginal likelihood

$$\mathcal{L} = -\log \int_{\Delta^{C-1}} p(y|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\pi}, \tag{2}$$

and add a KL regularizer toward the uniform Dirichlet prior to encourage vacuity when evidence is weak:

$$\mathcal{L}_{\mathrm{EV}} = \mathcal{L} + \lambda \cdot \mathrm{KL}[\mathrm{Dir}(\boldsymbol{\pi} \mid \tilde{\boldsymbol{\alpha}}) \,\|\, \mathrm{Dir}(\boldsymbol{\pi} \mid \mathbf{1})]\,, \tag{3}$$

where $\lambda > 0$ and $\tilde{\boldsymbol{\alpha}}$ removes ground-truth evidence as in [41].

Evidence-Targeted Adversary. Standard adversarial training typically constructs $\mathbf{x}^{\mathrm{adv}}$ by maximizing a classification loss (e.g., cross-entropy) within an ℓ_p-ball. In contrast, EV-AT constructs adversarial examples that explicitly target the evidential representation: the attacker seeks perturbations that maximally distort the Dirichlet parameters induced by the network, thereby directly stress-testing uncertainty integrity. Given a clean input $\mathbf{x}$, we define an evidence drift objective through a discrepancy D over log-Dirichlet parameters, $\mathcal{L} \triangleq D\big(\boldsymbol{\eta}, \boldsymbol{\eta}^{adv}\big)$, with $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$, $\boldsymbol{\eta}^{adv} = \log \boldsymbol{\alpha}^{adv}$. Starting from a random initialization $\mathbf{x}^{(0)} = \mathbf{x} + \boldsymbol{\delta}$, $\boldsymbol{\delta} \sim \mathcal{U}(-\varepsilon, \varepsilon)^d$, we construct adversarial examples via K-step PGD [34] and set $\mathbf{x}^{\mathrm{adv}} = \mathbf{x}^{(K)}$. This adversary is uncertainty-aware: it explicitly searches for perturbations that maximally corrupt the evidential posterior, rather than only flipping the argmax label.

Robust Evidence Alignment. In the min-max game of EV-AT, the outer minimization aims to simultaneously **(i)** fit the clean labels with the evidential loss and **(ii)** reduce the worst-case evidence drift found by the evidence-targeted adversary. To that end, we define the robust evidence alignment term as $\mathcal{L}_{\mathrm{REA}} \triangleq D(\boldsymbol{\eta}, \boldsymbol{\eta}^{\mathrm{adv}})$ and optimize the EV-AT objective

$$\min_{\theta} \; \mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\mathcal{L}_{\mathrm{EV}}\big(\boldsymbol{\alpha}, y\big) + \beta\, \mathcal{L}_{\mathrm{REA}}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\mathrm{adv}})\Big], \tag{4}$$

where $\beta > 0$ trades off clean evidential fitting and robust alignment.

Instantiating the discrepancy D. EV-AT is compatible with any differentiable discrepancy on $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$. In our implementation, we use the Improved Kullback-Leibler (IKL) Divergence loss, $D(\boldsymbol{\eta}, \boldsymbol{\eta}^{adv}) = \mathrm{IKL}(\boldsymbol{\eta}, \boldsymbol{\eta}^{adv})$, introduced in [14]. It measures drift directly in the log-parameter space and yields stable gradients for both the attack and the alignment objective. We discuss alternative divergences in the ablation study (Section 4.4).

Relation to adversarial training. Objective (4) parallels robust consistency training [14, 57], but enforces posterior-level (evidential) consistency rather than logit-level consistency. $\mathcal{L}_{\text{EV}}$ provides clean supervision, while $\mathcal{L}_{\text{REA}}$ penalizes worst-case discrepancies between clean and adversarial evidential representations. This directly discourages confident adversarial errors by limiting attack-induced changes in the predictive mean and evidence. Table S1 in the supplementary material summarizes the objective components of representative robustness methods and the proposed EV-AT formulation.

3.4 Theoretical Motivation & Analysis

We motivate two EV-AT design choices: **(i)** measuring adversarial drift at the Dirichlet-posterior level (rather than logits) and **(ii)** aligning in $\boldsymbol{\eta}$-space. Since selection thresholds the predictive entropy $u(\mathbf{x}) = \mathrm{H}[\bar{\boldsymbol{\pi}}]$, robust selective behavior requires limiting attack-induced changes in the predictive mean $\bar{\boldsymbol{\pi}}$ and in confidence. We show that controlling drift in $\boldsymbol{\eta}$ stabilizes both the predictive mean $\bar{\boldsymbol{\pi}}$, which directly determines the entropy-based selection score, and the Dirichlet strength S, which governs posterior concentration.

Posterior alignment. Logits alignment is a common robustness heuristic, but it only indirectly constrains the quantities used for selective prediction. Logits are not identifiable (e.g., additive shifts leave softmax unchanged, and scaling affects confidence implicitly), and standard alignment targets a single classification vector rather than the confidence structure needed for abstention. In an evidential model, the Dirichlet posterior separates these roles: the mean $\bar{\boldsymbol{\pi}}$ determines the predicted class, while the strength S controls uncertainty. Aligning the (log-)posterior therefore directly stabilizes both the predictive distribution and its confidence, reducing overconfident errors under perturbation.

Log-concentration space. Mapping concentrations to $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$ yields three useful properties.

(a) Log-parameters induce the predictive mean. With $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$,

$$\bar{\boldsymbol{\pi}} \triangleq \mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\boldsymbol{\pi}] \;=\; \frac{\boldsymbol{\alpha}}{\sum_c \alpha_c} \;=\; \frac{\exp(\boldsymbol{\eta})}{\sum_c \exp(\eta_c)} \;=\; \mathrm{softmax}(\boldsymbol{\eta}). \tag{5}$$

Consequently, constraining drift in $\boldsymbol{\eta}$ directly constrains drift in $\bar{\boldsymbol{\pi}}$. Lemma 1 formalizes why log-Dirichlet alignment is appropriate for posterior robustness: it simultaneously limits changes in the predictive mean (class probabilities) and in the evidence scale (strength), two levers that adversarial perturbations can exploit to produce confident errors.

Lemma 1 (Multiplicative stability under log-parameter drift). *Let $\boldsymbol{\alpha} \in \mathbb{R}^C_{>1}$ be Dirichlet concentration parameters, with $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$. Define the total strength $S = \sum_{c=1}^C \alpha_c$, and the predictive mean $\bar{\boldsymbol{\pi}} = \frac{\boldsymbol{\alpha}}{S}$. Let $\boldsymbol{\alpha}'$, $\boldsymbol{\eta}'$, S', and $\bar{\boldsymbol{\pi}}'$ denote the corresponding perturbed quantities. If $\|\boldsymbol{\eta}' - \boldsymbol{\eta}\|_\infty \le \rho$, then for all classes c,*

$$e^{-\rho} \le \frac{\alpha'_c}{\alpha_c} \le e^{\rho}, \qquad e^{-\rho} \le \frac{S'}{S} \le e^{\rho}, \qquad e^{-2\rho} \le \frac{\bar{\pi}'_c}{\bar{\pi}_c} \le e^{2\rho}. \tag{6}$$

(b) Log-drift gives multiplicative evidence stability. Additive changes in $\boldsymbol{\eta}$ correspond to multiplicative changes in $\boldsymbol{\alpha}$: for $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta} + \Delta\boldsymbol{\eta}$, $\tilde{\boldsymbol{\alpha}} = \exp(\tilde{\boldsymbol{\eta}}) = \boldsymbol{\alpha} \odot \exp(\Delta\boldsymbol{\eta})$, providing a well-scaled notion of evidence drift.

(c) Strength controls posterior precision. For a fixed mean, the strength $S = \sum_c \alpha_c$ controls posterior concentration. Lemmas 1 and 2 formalize that controlling $\boldsymbol{\eta}$ jointly stabilizes the mean $\bar{\boldsymbol{\pi}}$ and precision S (Figure 3).

Lemma 2 (Strength upper-bounds posterior variance). *For any Dirichlet distribution* $\mathrm{Dir}(\boldsymbol{\alpha})$ *with strength* $S = \sum_c \alpha_c$,

$$\max_{c \in \{1,\dots,C\}} \mathrm{Var}[\pi_c] \leq \frac{1}{4(S+1)}. \tag{7}$$

Guarantees of alignment in entropy-based selection. Our selector thresholds predictive entropy $u(\mathbf{x}) = \mathrm{H}[\mathrm{Cat}(\bar{\boldsymbol{\pi}})]$, so controlling adversarial drift of $\bar{\boldsymbol{\pi}}$ stabilizes both uncertainty and the accept/reject decision.

(a) Entropy continuity. If $\bar{\boldsymbol{\pi}}^{\mathrm{adv}}$ remains close to $\bar{\boldsymbol{\pi}}$, then u^{adv} remains close to u by continuity of Shannon entropy (Lemma 3).

Lemma 3 (Entropy continuity). *Let* $\mathbf{p}, \mathbf{q} \in \Delta^{C-1}$ *and define*

$$\delta = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1. \tag{8}$$

Then

$$|\mathrm{H}(\mathbf{p}) - \mathrm{H}(\mathbf{q})| \leq \delta \log(C-1) + h(\delta), \tag{9}$$

where

$$h(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta) \tag{10}$$

is the binary entropy.

(b) Stability of selection. Let the selection margin be $\tau - u$. If $|u^{\mathrm{adv}} - u| \leq \epsilon_H$ and $|\tau - u| > \epsilon_H$, then the decision is unchanged: $g_\tau(\mathbf{x}^{\mathrm{adv}}; u^{\mathrm{adv}}) = g_\tau(\mathbf{x}; u)$. Thus, alignment reduces brittleness except near the threshold boundary. Proofs of Lemmas 1 to 3 and of the selection-stability consequence are provided in supplementary Sec. S1.4.

4 Experiments and Results

4.1 Benchmark Design and Protocol

We design a controlled benchmark for robustness–uncertainty trade-offs in selective classification. It comprises four layers: data, methods, evaluation, and reporting. See supplementary Fig. S1 and Sec. S2.1 for details.

Data Layer. We evaluate all methods on CIFAR-10/100 [30] under clean, adversarial, and common-corruption conditions (CIFAR-10/100-C [25]). For each

setting, we consider four augmentation strategies: Basic [54], Cutout [16, 58], AutoAugment [13], and AugMix [26].

Method Layer. We compare EV-AT with AT [34], TRADES [57], AT-AWP, TRADES-AWP [54], IKL-AT [14], and TRADES-EMFF [51]. These baselines cover standard adversarial optimization, consistency-based regularization, adversarial weight perturbation, and recent uncertainty-aware robust-training objectives. We evaluate all methods using WideResNet-34-10 [55] and PreActResNet-18 [24], representing complementary high- and moderate-capacity regimes. For each comparison, we fix the architecture, augmentation regime, optimizer family, training budget, while varying only the training objective and method-specific coefficients. All models are trained independently under identical experimental conditions using three shared random seeds, and we report the mean and standard deviation across runs.

Evaluation Layer. We evaluate robustness under ℓ_∞ ($\varepsilon = 8/255$) and ℓ_2 ($\varepsilon = 128/255$) threat models, using AutoAttack (AA) [12] as the primary adversarial evaluation and PGD-20/PGD-100 [34] as diagnostic checks. We report accuracy under clean, adversarial, and corruption conditions, following standard robustness benchmarks [11]. For selective prediction, samples are rejected according to an uncertainty score $u(x)$. We report retention and risk-coverage curves, with scalar summaries AURC [35], AUGRC [45] (primary), and AUROC for error detection [37]. Metrics are computed consistently across clean, adversarial, and corruption settings.

Reporting Layer. We report per-regime results and clean/AA aggregates. Following [53], we define $\mathrm{Acc.}_{\mathrm{avg}} = \frac{1}{2}(\mathrm{Acc.}_{\mathrm{clean}} + \mathrm{Acc.}_{\mathrm{AA}})$ and compute AURC, AUGRC, and AUROC aggregates analogously.

Implementation and Reproducibility. All training and evaluation configurations are shared across methods through the same benchmark implementation. The complete hyperparameter specification is provided in supplementary Tab. S2.

4.2 Benchmark Results & Findings

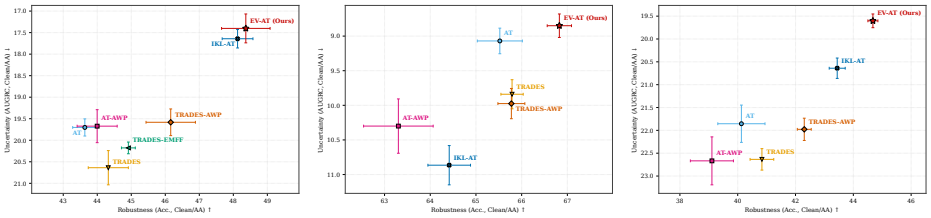
How do SOTA robust training methods perform as selective classifiers? Table 1 summarizes our benchmark on CIFAR-10/WRN-34-10 across four augmentation regimes, reporting accuracy and AUGRC on clean, adversarial (AA), and corruption (CIFAR-C) inputs, as well as the clean+AA aggregate (Clean/AA). Figure 1b provides a complementary multi-metric comparison. Additional metrics, datasets, and architectures are reported in supplementary Sec. S3. A key finding is that no SOTA baseline dominates as a selective system: higher AA accuracy does not reliably imply lower adversarial AUGRC. For example, under Basic, IKL-AT and TRADES-AWP achieve the strongest AA accuracies among baselines yet remain comparatively weak in $\mathrm{AUGRC}_{\mathrm{AA}}$, indicating degraded uncertainty ranking under attack. Conversely, AT-AWP attains the best Clean/AA AUGRC among baselines under Basic despite lower AA

Table 1: Robustness-uncertainty benchmark on CIFAR-10 with WRN-34-10 across data augmentations. We report mean±std over three seeds for robustness and uncertainty metrics under clean / adversarial / corruption shifts. Within each augmentation block, **best**, second-best, and third-best results are highlighted per metric.

Method	Venue	Robustness (Acc. ↑)				Uncertainty & Selective Classification (AUGRC ↓)			
		Clean	AA	Corr.	Clean/AA	Clean	AA	Corr.	Clean/AA
Aug.: Basic									
AT [34]	ICML'18	85.11 ± 1.60	51.63 ± 0.49	76.10 ± 1.77	68.37 ± 0.93	2.79 ± 0.45	12.32 ± 0.25	5.58 ± 0.72	7.56 ± 0.33
AT-AWP [54]	NeurIPS'20	85.28 ± 0.81	53.57 ± 0.65	76.00 ± 1.13	69.42 ± 0.73	2.96 ± 0.22	11.64 ± 0.31	6.00 ± 0.44	7.30 ± 0.26
TRADES [57]	ICML'19	84.53 ± 0.33	52.92 ± 0.35	75.88 ± 0.37	68.73 ± 0.02	3.83 ± 0.13	12.82 ± 0.18	6.90 ± 0.19	8.32 ± 0.03
TRADES-AWP [54]	NeurIPS'20	84.89 ± 0.56	55.85 ± 0.51	76.57 ± 0.57	70.37 ± 0.51	3.58 ± 0.30	11.26 ± 0.42	6.49 ± 0.39	7.42 ± 0.36
IKL-AT [14]	NeurIPS'24	85.04 ± 0.21	**56.18 ± 0.30**	76.69 ± 0.21	70.61 ± 0.25	3.48 ± 0.06	11.27 ± 0.12	6.40 ± 0.07	7.38 ± 0.09
TRADES-EMFF [51]	TPAMI'25	84.97 ± 0.60	50.35 ± 0.19	75.97 ± 0.50	67.66 ± 0.25	3.67 ± 0.17	14.03 ± 0.13	6.82 ± 0.21	8.85 ± 0.13
EV-AT (Ours)	-	**88.16 ± 0.28**	55.38 ± 0.25	**79.91 ± 0.43**	**71.77 ± 0.14**	**2.20 ± 0.02**	**10.70 ± 0.08**	**4.58 ± 0.15**	**6.45 ± 0.03**
Aug.: Cutout									
AT [34]	ICML'18	84.42 ± 0.14	52.07 ± 0.38	75.31 ± 0.10	68.24 ± 0.26	3.09 ± 0.12	12.20 ± 0.21	6.05 ± 0.07	7.64 ± 0.16
AT-AWP [54]	NeurIPS'20	82.65 ± 0.46	52.70 ± 0.47	73.63 ± 0.51	67.67 ± 0.47	3.94 ± 0.14	12.30 ± 0.23	7.16 ± 0.19	8.12 ± 0.18
TRADES [57]	ICML'19	85.63 ± 0.19	53.28 ± 0.37	76.97 ± 0.19	69.46 ± 0.13	3.32 ± 0.02	12.46 ± 0.19	6.13 ± 0.08	7.89 ± 0.11
TRADES-AWP [54]	NeurIPS'20	85.40 ± 0.22	56.39 ± 0.27	76.71 ± 0.28	70.90 ± 0.16	3.57 ± 0.11	11.17 ± 0.11	6.65 ± 0.14	7.37 ± 0.10
IKL-AT [14]	NeurIPS'24	82.67 ± 0.24	55.85 ± 0.18	74.28 ± 0.18	69.26 ± 0.04	4.73 ± 0.09	12.14 ± 0.04	8.06 ± 0.12	8.43 ± 0.04
TRADES-EMFF [51]	TPAMI'25	87.04 ± 0.08	51.81 ± 0.10	78.09 ± 0.25	69.42 ± 0.06	2.85 ± 0.02	13.10 ± 0.11	5.68 ± 0.10	7.98 ± 0.05
EV-AT (Ours)	-	**87.37 ± 0.15**	**56.49 ± 0.08**	**78.84 ± 0.19**	**71.93 ± 0.09**	**2.64 ± 0.06**	**10.42 ± 0.02**	**5.32 ± 0.10**	**6.53 ± 0.04**
Aug.: AutoAug									
AT [34]	ICML'18	84.91 ± 0.79	50.68 ± 0.64	75.65 ± 1.41	67.80 ± 0.71	3.27 ± 0.31	13.24 ± 0.29	6.19 ± 0.61	8.25 ± 0.30
AT-AWP [54]	NeurIPS'20	83.86 ± 0.40	52.31 ± 0.44	75.17 ± 0.52	68.08 ± 0.32	3.83 ± 0.07	12.69 ± 0.17	6.75 ± 0.11	8.26 ± 0.10
TRADES [57]	ICML'19	87.37 ± 0.07	52.76 ± 0.04	79.59 ± 0.16	70.06 ± 0.06	2.99 ± 0.05	12.78 ± 0.04	5.26 ± 0.06	7.88 ± 0.04
TRADES-AWP [54]	NeurIPS'20	87.03 ± 0.03	55.74 ± 0.18	79.44 ± 0.39	71.39 ± 0.10	3.31 ± 0.07	11.61 ± 0.09	5.71 ± 0.12	7.46 ± 0.08
IKL-AT [14]	NeurIPS'24	85.28 ± 0.22	**55.91 ± 0.16**	78.16 ± 0.31	70.60 ± 0.10	3.79 ± 0.12	11.64 ± 0.14	6.17 ± 0.17	7.72 ± 0.11
TRADES-EMFF [51]	TPAMI'25	**88.90 ± 0.11**	51.23 ± 0.12	**81.37 ± 0.20**	70.07 ± 0.03	2.47 ± 0.01	13.18 ± 0.18	4.60 ± 0.06	7.83 ± 0.09
EV-AT (Ours)	-	88.40 ± 0.11	55.86 ± 0.16	81.12 ± 0.05	**72.13 ± 0.11**	**2.24 ± 0.04**	**10.43 ± 0.08**	**4.32 ± 0.03**	**6.34 ± 0.05**
Aug.: AugMix									
AT [34]	ICML'18	83.38 ± 0.59	52.17 ± 0.18	76.93 ± 0.62	67.78 ± 0.23	3.51 ± 0.17	12.35 ± 0.08	5.67 ± 0.23	7.93 ± 0.08
AT-AWP [54]	NeurIPS'20	81.41 ± 0.65	52.76 ± 0.60	74.46 ± 0.98	67.09 ± 0.62	4.37 ± 0.22	12.45 ± 0.34	6.93 ± 0.42	8.41 ± 0.28
TRADES [57]	ICML'19	84.06 ± 1.04	47.71 ± 7.76	77.33 ± 1.77	65.88 ± 3.37	4.06 ± 0.76	15.64 ± 3.56	6.45 ± 1.21	9.85 ± 1.40
TRADES-AWP [54]	NeurIPS'20	84.64 ± 1.42	52.37 ± 2.98	78.01 ± 1.64	68.50 ± 0.80	3.90 ± 0.85	13.00 ± 0.60	6.27 ± 1.10	8.45 ± 0.16
IKL-AT [14]	NeurIPS'24	82.85 ± 0.20	55.38 ± 0.10	76.18 ± 0.12	69.11 ± 0.11	4.65 ± 0.05	12.27 ± 0.02	7.23 ± 0.08	8.46 ± 0.03
TRADES-EMFF [51]	TPAMI'25	84.23 ± 0.45	49.90 ± 0.14	76.84 ± 0.41	67.07 ± 0.18	4.25 ± 0.21	14.52 ± 0.13	6.88 ± 0.26	9.38 ± 0.17
EV-AT (Ours)	-	**87.10 ± 0.30**	**55.93 ± 0.24**	**80.81 ± 0.14**	**71.52 ± 0.05**	**2.78 ± 0.05**	**10.73 ± 0.10**	**4.73 ± 0.07**	**6.75 ± 0.07**

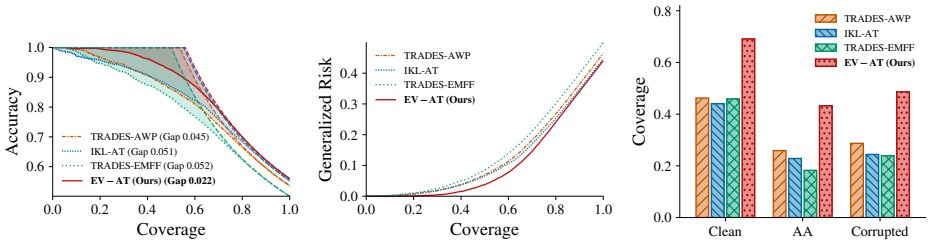
accuracy, revealing an internal robustness–uncertainty trade-off. Similar mismatches persist under shift (CIFAR-C), where TRADES-EMFF can achieve strong clean/corruption accuracy but poor adversarial AUGRC. Overall, baselines exhibit "spiky" profiles across robustness and selective metrics, motivating the analyses below.

Do robust-accuracy gains correlate with robust uncertainty, and which methods lie on the Pareto frontier? Figures 1a and 4 plot Acc._{avg} versus $\text{AUGRC}_{\text{avg}}$ to summarize the robustness–uncertainty trade-off. The correlation is weak: improved adversarial robustness does not reliably improve uncertainty for selective prediction. For instance, IKL-AT increases robustness over AT (69.90 vs. 68.05) but yields worse uncertainty (8.00 vs. 7.85), and AT-AWP matches AT in robustness (68.07 vs. 68.05) while degrading uncertainty (8.02 vs. 7.85). Among SOTA baselines on CIFAR-10/WRN-34-10, TRADES-AWP offers the strongest trade-off (e.g., $\text{Acc.}_{\text{avg}} = 70.29$, $\text{AUGRC}_{\text{avg}} = 7.68$), illustrating a non-trivial Pareto structure.



(a) WRN-34-10-CIFAR100 **(b)** PreActResNet18-CIFAR10 **(c)** PreActResNet18-CIFAR100

Fig. 4: Robustness vs Uncertainty trade-off across datasets and architectures.



(a) Retention curve **(b)** Gen. risk-coverage **(c)** Coverage @ 5% Risk

Fig. 5: Selective classification under AA perturbations. CIFAR-10/WRN-34-10 (AugMix): TRADES-AWP, IKL-AT, TRADES-EMFF vs. EV-AT. (a) Retention (Acc. vs. coverage) on clean/AA/Corr. with clean→shift gap. (b) Generalized risk–coverage (lower is better). (c) Coverage at 5% risk (higher is better). EV-AT reduces adversarial degradation and improves coverage at fixed risk.

Where does selective behavior fail under attack, and does it manifest as confident errors? Figure 5 analyzes selective behavior under adversarial perturbations for AugMix-trained CIFAR-10/WRN-34-10 models, comparing strong baselines and EV-AT. AA perturbations cause the largest degradation in retention and generalized-risk curves, more severe than common corruptions. Failures concentrate at moderate-to-high coverage: many adversarial errors remain accepted, indicating overconfident mistakes. This is reflected by faster risk accumulation and by low coverage at fixed risk (baselines coverage@5% under AA: TRADES-AWP 0.26, IKL-AT 0.23, TRADES-EMFF 0.18). Overall, under attack, uncertainty no longer separates correct from incorrect predictions, leading to silent failures at practical operating points.

Does EV-AT shift the robustness–uncertainty Pareto frontier beyond prior SOTA? EV-AT improves robustness and selective performance simultaneously, moving beyond the Pareto frontier formed by prior objectives (Figure 1a). Averaged over augmentations, EV-AT improves Acc._{avg} from 70.29 for TRADES-AWP to 71.84 and reduces $\text{AUGRC}_{\text{avg}}$ from 7.68 to 6.52. This is consistent across augmentation regimes in the Clean/AA summary (Table 1) and extends to corruption shift, where EV-AT also yields the best Corr. accuracy and Corr. AUGRC. The improvement is reflected across metrics in the radar

Table 2: Generalization across attack types and threat models. We report clean/adversarial metrics averaged across data augmentations and three random seeds.

Norm	Method	PGD-20		PGD-100		AA	
		Acc. ↑	AUGRC↓	Acc. ↑	AUGRC↓	Acc. ↑	AUGRC↓
L_∞	IKL-AT	72.05 ± 0.95	8.91 ± 0.47	71.95 ± 1.00	8.95 ± 0.47	69.89 ± 0.80	7.99 ± 0.51
	EV-AT (Ours)	**74.03** ± 0.49	**7.66** ± 0.18	**73.90** ± 0.51	**7.71** ± 0.19	**71.83** ± 0.27	**6.52** ± 0.18
L_2	IKL-AT	82.15 ± 0.74	4.35 ± 0.24	81.97 ± 0.73	4.42 ± 0.23	81.37 ± 0.66	3.56 ± 0.32
	EV-AT (Ours)	**83.84** ± 0.61	**3.50** ± 0.19	**83.60** ± 0.61	**3.60** ± 0.21	**82.96** ± 0.44	**2.49** ± 0.10

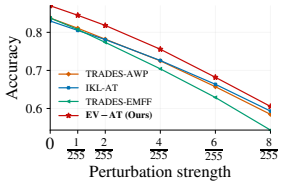


Fig. 6: Generalization across perturbation budgets.

view (Figure 1b) and in selective operating points: EV-AT reduces the clean-AA retention-gap and substantially increases coverage at fixed risk (Figure 5). Overall, EV-AT shifts the robustness–uncertainty frontier by improving robustness without degrading uncertainty ranking, strengthening end-to-end selective behavior under both adversarial and corruption shifts.

4.3 Generalization and Stress Tests

Diverse Attacks and Threat Models. We test whether EV-AT's robustness–uncertainty gains persist across stronger attacks and threat models. Table 2 reports accuracy and AUGRC averaged over clean and attacked inputs for CIFAR-10/WRN-34-10 and across augmentation regimes. Compared with the strongest baseline, IKL-AT, EV-AT consistently achieves higher accuracy and lower AUGRC under PGD-20, PGD-100, and AutoAttack (AA), for both ℓ_∞ and ℓ_2 threat models. Under ℓ_∞, it improves Clean/PGD-20, Clean/PGD-100, and Clean/AA accuracy from 72.05% to 74.03%, 71.95% to 73.90%, and 69.89% to 71.83%, respectively, while reducing Clean/AA AUGRC from 7.99 to 6.52. The trend also holds under ℓ_2, where Clean/AA accuracy increases from 81.37% to 82.96% and AUGRC decreases from 3.56 to 2.49. The consistent gains under more PGD iterations and AA indicate that they are unlikely to result from insufficient attack optimization.

Perturbation Budgets. We stress-test robustness by sweeping the ℓ_∞ radius $\varepsilon \in \{0, 1, 2, 4, 6, 8\}/255$ on CIFAR-10/WRN-34-10, where $\varepsilon = 0$ corresponds to clean evaluation. Figure 6 reports PGD-20 accuracy (AugMix). EV-AT is consistently best across the full range, with gains that persist at both small ($\varepsilon \leq 2/255$) and large ($\varepsilon \geq 6/255$) budgets. This indicates improved robustness beyond the standard $\varepsilon = 8/255$ setting, rather than overfitting to a single evaluation radius.

Architectures and Augmentations. We test generalization across architectures and augmentation pipelines, two major confounders in robust training. Across all four augmentations (Basic/Cutout/AutoAugment/AugMix), EV-AT achieves the best Clean/AA accuracy while also attaining the lowest Clean/AA AUGRC (Table 1), whereas several baselines exhibit augmentation-dependent

Table 3: Objective component ablation on CIFAR-10 with WRN-34-10 using AugMix. We report averaged metrics across clean and adversarial (AA) conditions.

Variant	Clean loss	REA	AWP	Acc. ↑	AUGRC↓
EV-AT w/ CE	CE	✓	✓	69.56	7.88
EV-AT w/o REA (β=0)	$\mathcal{L}_{EV}$	✗	✓	20.11	37.67
EV-AT w/o AWP	$\mathcal{L}_{EV}$	✓	✗	68.46	8.21
EV-AT (Ours)	$\mathcal{L}_{EV}$	✓	✓	**71.51**	**6.67**

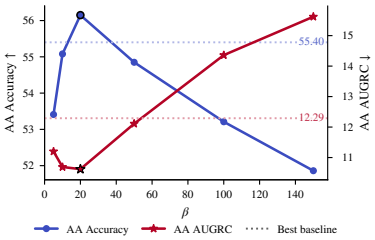


Fig. 7: Effect of the REA weight β.

Table 4: Design ablations on CIFAR-10 with WRN-34-10 using AugMix. We report averaged metrics across clean and adversarial (AA) conditions.

(a) Alignment space

Variant	Space	Acc. ↑	AUGRC↓
EV-AT align in α	α	67.32	10.26
EV-AT align in $\bar{\boldsymbol{\pi}}$ $\bar{\boldsymbol{\pi}} = \alpha/S$		64.64	10.43
EV-AT (Ours)	$\eta = \log \alpha$	**71.51**	**6.67**

(b) REA divergence D

Variant	D	Acc. ↑	AUGRC↓
EV-AT w/ L_2	L_2	69.81	7.64
EV-AT w/ KL	KL	70.61	7.60
EV-AT (Ours)	IKL	**71.51**	**6.67**

(c) Inner-max objective

Variant	Inner-max	Acc. ↑	AUGRC↓
EV-AT w/ CE	CE	71.26	**6.38**
EV-AT w/ KL	KL	**71.52**	6.78
EV-AT (Ours)	IKL	71.51	6.67

trade-offs. Across architectures and datasets (WRN-34-10 vs. PreActResNet-18; CIFAR-10/100), EV-AT remains on (or beyond) the Pareto frontier in the Acc._{avg}-$\text{AUGRC}_{\text{avg}}$ plane (Figure 4). Overall, EV-AT's advantage is stable across both model family and augmentation regime, suggesting it is not an artifact of a specific capacity or pipeline choice.

4.4 Ablations and Analysis of EV-AT

Objective components. Table 3 ablates EV-AT on CIFAR-10/WRN-34-10 with AugMix under ℓ_∞ AutoAttack ($\varepsilon = 8/255$), reporting clean+AA averages. REA is essential: removing it ($\beta = 0$) collapses performance (Acc. 71.51 → 20.11, AUGRC 6.67 → 37.67), indicating that the clean evidential loss alone is not robust and that posterior alignment in log-Dirichlet space prevents pathological evidence mismatch under attack. AWP provides an additional but non-substitutable gain (w/o AWP: Acc. 68.46, AUGRC 8.21). Replacing the evidential clean loss with cross-entropy also degrades the trade-off (Acc. 69.56, AUGRC 7.88). Overall, EV-AT's gains come from combining the evidential objective with REA, with AWP further improving optimization.

Design Choices. Table 4 ablates key EV-AT design choices on CIFAR-10 using WRN-34-10 and AugMix. We report clean+AA averages under ℓ_∞ AutoAttack with $\varepsilon = 8/255$. **(a) Alignment space:** aligning in log-Dirichlet space $\boldsymbol{\eta} = \log \boldsymbol{\alpha}$ is crucial; aligning in $\boldsymbol{\alpha}$ or in the Dirichlet mean $\bar{\boldsymbol{\pi}} = \boldsymbol{\alpha}/S$ substantially degrades the trade-off (Acc. 67.32/64.64, AUGRC 10.26/10.43 vs. $\boldsymbol{\eta}$: Acc. 71.51, AUGRC 6.67). **(b) Divergence for REA:** IKL performs best; ℓ_2 or KL reduces

performance (Acc. 69.81/70.61, AUGRC 7.64/7.60). **(c) Inner-max objective:** CE achieves the lowest AUGRC, KL attains the marginally highest accuracy, and IKL provides a balanced operating point close to the best value on both metrics. Overall, robust selective prediction benefits most from posterior alignment in $\boldsymbol{\eta}$, with IKL providing the most effective signal for both alignment and adversarial example generation.

Sensitivity & stability. We study sensitivity to the REA weight β using a sweep on CIFAR-10/WRN-34-10 (AugMix) under ℓ_∞ AutoAttack ($\varepsilon = 8/255$); Figure 7 reports AA accuracy and AA AUGRC (dotted lines: best baseline, IKL-AT). Small β under-emphasizes alignment, yielding weaker robustness and selective performance. Increasing β improves AA accuracy and reduces AA AUGRC up to a moderate range, where EV-AT outperforms IKL-AT on both metrics. For overly large β, over-regularization harms both AA accuracy and AUGRC. Overall, β controls the trade-off between clean evidential structure and adversarial posterior alignment, with an intermediate value giving the best joint robustness–uncertainty operating point.

5 Conclusion and Future Work

In this paper, we studied robustness–uncertainty interactions via robust selective classification. Under a unified benchmark, we show that robust-accuracy gains can degrade uncertainty ranking and risk-coverage behavior under attack. We then propose EV-AT, which learns Dirichlet posteriors and enforces posterior-level robustness via robust evidence alignment. Across datasets, architectures, augmentations, attacks, and threat models, EV-AT improves the robustness–uncertainty trade-off, shifting the Pareto frontier and strengthening selective performance. Despite these gains, several directions remain open. An immediate next step is to scale EV-AT to larger architectures and datasets and to extend the protocol to structured vision tasks (e.g., detection and segmentation), where selective decision-making is central. More broadly, evaluating robust selective classification in additional modalities (e.g., 3D point clouds) and spatiotemporal settings (e.g., video) may reveal new robustness–uncertainty failure modes.

Acknowledgements

This work was supported by the Walloon Public Service (Economy, Employment and Research) under Grant No. 2010235 (ARIAC – Applications and Research for Trusted Artificial Intelligence), within the DigitalWallonia4.ai program. The research also benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the Grant Agreement No. 1910247.

References

1. Aguilar, E., Raducanu, B., Radeva, P.: Multi-label out-of-distribution detection via evidential learning. In: ECCV Workshops. vol. 15639, pp. 202–218 (2024). https://doi.org/10.1007/978-3-031-91585-7_13
2. van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: ICML. vol. 119, pp. 9690–9700 (2020)
3. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: ICLR. pp. 2600–2608 (2024)
4. Athalye, A., Carlini, N., Wagner, D.A.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: ICML. vol. 80, pp. 274–283 (2018)
5. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. In: IJCAI. pp. 4312–4321 (2021). https://doi.org/10.24963/ijcai.2021/591
6. Bartoldson, B.R., Diffenderfer, J., Parasyris, K., Kailkhura, B.: Adversarial robustness limits via scaling-law and human-alignment studies. In: ICML. vol. 235, pp. 3046–3072 (2024)
7. Bengs, V., Hüllermeier, E., Waegeman, W.: On second-order scoring rules for epistemic uncertainty quantification. In: ICML. vol. 202, pp. 2078–2091 (2023)
8. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A.: On evaluating adversarial robustness. arXiv preprint **abs/1902.06705** (2019)
9. Chen, M., Gao, J., Yang, S., Xu, C.: Dual-evidential learning for weakly-supervised temporal action localization. In: ECCV. vol. 13664, pp. 192–208 (2022). https://doi.org/10.1007/978-3-031-19772-7_12
10. Cinà, A.E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., Ayed, I.B., Roli, F.: AttackBench: Evaluating gradient-based attacks for adversarial examples. In: AAAI. pp. 2600–2608 (2025). https://doi.org/10.1609/AAAI.V39I3.32263
11. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: RobustBench: a standardized adversarial robustness benchmark. In: NeurIPS (2021)
12. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML. vol. 119, pp. 2206–2216 (2020)
13. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: AutoAugment: Learning augmentation strategies from data. In: CVPR. pp. 113–123 (2019). https://doi.org/10.1109/CVPR.2019.00020
14. Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., Zhang, H.: Decoupled Kullback-Leibler divergence loss. In: NeurIPS (2024)
15. Detavernier, A., Bock, J.D.: Robustness and uncertainty: two complementary aspects of the reliability of the predictions of a classifier. arXiv **abs/2512.15492** (2025)
16. Devries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. CoRR **abs/1708.04552** (2017)
17. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML. vol. 48, pp. 1050–1059 (2016)
18. Gao, J., Chen, M., Xiang, L., Xu, C.: A comprehensive survey on evidential deep learning and its applications. TPAMI **48**(3), 2118–2138 (2026). https://doi.org/10.1109/TPAMI.2025.3625258

19. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**, 1513–1589 (2023). https://doi.org/10.1007/s10462-023-10562-9
20. Gibbs, I., Candès, E.J.: Adaptive conformal inference under distribution shift. In: NeurIPS. pp. 1660–1672 (2021)
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
22. Gowal, S., Rebuffi, S., Wiles, O., Stimberg, F., Calian, D.A., Mann, T.A.: Improving robustness using generated data. In: NeurIPS. pp. 4218–4233 (2021)
23. Graves, A.: Practical variational inference for neural networks. In: NeurIPS. pp. 2348–2356 (2011)
24. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. vol. 9908, pp. 630–645 (2016). https://doi.org/10.1007/978-3-319-46493-0_38
25. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
26. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: AugMix: A simple data processing method to improve robustness and uncertainty. In: ICLR (2020)
27. Jeary, L., Kuipers, T., Hosseini, M., Paoletti, N.: Verifiably robust conformal prediction for probabilistic guarantees under adversarial attacks. PR **170**, 112051 (2026). https://doi.org/10.1016/J.PATCOG.2025.112051
28. Jürgens, M., Meinert, N., Bengs, V., Hüllermeier, E., Waegeman, W.: Is epistemic uncertainty faithfully represented by evidential deep learning methods? In: ICML. vol. 235, pp. 22624–22642 (2024)
29. Kopetzki, A.K., Charpentier, B., Zügner, D., Giri, S., Günnemann, S.: Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? In: ICML. pp. 5707–5718 (2021)
30. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto (2009)
31. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. In: ICLR (2017)
32. Ledda, E., Scodeller, G., Angioni, D., Piras, G., Cinà, A.E., Fumera, G., Biggio, B., Roli, F.: On the robustness of adversarial training against uncertainty attacks. PR **172**, 112519 (2026). https://doi.org/10.1016/J.PATCOG.2025.112519
33. Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D.P., Wilson, A.G.: A simple baseline for Bayesian uncertainty in deep learning. In: NeurIPS. pp. 13132–13143 (2019)
34. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
35. Moon, J., Kim, J., Shin, Y., Hwang, S.: Confidence-aware learning for deep neural networks. In: ICML. vol. 119, pp. 7034–7044 (2020)
36. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H.S., Gal, Y.: Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. In: ICML (2021)
37. Nandy, J., Hsu, W., Lee, M.: Towards maximizing the representation gap between in-domain & out-of-distribution examples. In: NeurIPS (2020)
38. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo, pp. 113–162. Chapman and Hall/CRC (2011)

39. Prach, B., Brau, F., Buttazzo, G.C., Lampert, C.H.: 1-Lipschitz layers compared: Memory, speed, and certifiable robustness. In: CVPR. pp. 24574–24583 (2024). https://doi.org/10.1109/CVPR52733.2024.02320
40. Rabanser, S., Papernot, N.: What does it take to build a performant selective classifier? In: NeurIPS (2025)
41. Sensoy, M., Kaplan, L.M., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: NeurIPS. pp. 3183–3193 (2018)
42. Shafer, G., Vovk, V.: A tutorial on conformal prediction. J. Mach. Learn. Res. **9**, 371–421 (2008)
43. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
44. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: Attacks and defenses. In: ICLR (2018)
45. Traub, J., Bungert, T.J., Lüth, C.T., Baumgartner, M., Maier-Hein, K.H., Maier-Hein, L., Jaeger, P.F.: Overcoming common flaws in the evaluation of selective classification systems. In: NeurIPS (2024)
46. Venkataramanan, A., Benbihi, A., Laviale, M., Pradalier, C.: Gaussian latent representations for uncertainty estimation using Mahalanobis distance in deep classifiers. In: ICCV. pp. 4490–4499 (2023). https://doi.org/10.1109/ICCVW60793.2023.00483
47. Wang, H., Wang, Y.: Self-ensemble adversarial training for improved robustness. In: ICLR (2022)
48. Wang, H., Wang, Y.: Generalist: Decoupling natural and robust generalization. In: CVPR. pp. 20554–20563 (2023). https://doi.org/10.1109/CVPR52729.2023.01969
49. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: ICLR (2020)
50. Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., Yan, S.: Better diffusion models further improve adversarial training. In: ICML. vol. 202, pp. 36246–36263 (2023)
51. Wang, Z., Xu, X., Zhu, L., Bin, Y., Wang, G., Yang, Y., Shen, H.T.: Evidence-based multi-feature fusion for adversarial robustness. TPAMI **47**(10), 8923–8937 (2025). https://doi.org/10.1109/TPAMI.2025.3582518
52. Wang, Z., Ku, A., Baldridge, J., Griffiths, T., Kim, B.: Gaussian process probes (GPP) for uncertainty-aware probing. In: NeurIPS (2023)
53. Waseda, F.K., Chang, C., Echizen, I.: Rethinking invariance regularization in adversarial training to improve robustness-accuracy trade-off. In: ICLR (2025)
54. Wu, D., Xia, S., Wang, Y.: Adversarial weight perturbation helps robust generalization. In: NeurIPS (2020)
55. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC. pp. 87.1–87.12 (2016). https://doi.org/10.5244/C.30.87
56. Zargarbashi, S.H., Akhondzadeh, M.S., Bojchevski, A.: Robust yet efficient conformal prediction sets. In: ICML. vol. 235, pp. 17123–17147 (2024)
57. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML. vol. 97, pp. 7472–7482 (2019)
58. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. pp. 13001–13008 (2020). https://doi.org/10.1609/AAAI.V34I07.7000