Gradual Uncertainty Refinement via Noise-Driven Curriculum: A Post-Hoc Meta-Model for Robust Uncertainty Quantification

Charmaine Barker[*1] **Daniel Bethell**[*1] **Simos Gerasimou**[1,2]

[1]Department of Computer Science, University of York, York, UK
[2]Cyprus University of Technology, Limassol, Cyprus

Abstract

Reliable uncertainty quantification remains a major obstacle to the deployment of deep learning models under distributional shift. Existing post-hoc approaches that retrofit pretrained models either inherit misplaced confidence or merely reshape predictions, without advising the model when to be uncertain. We introduce GUIDE, a lightweight evidential learning meta-model that attaches to a frozen deep learning model and is explicitly guided on when and how to be uncertain. GUIDE identifies salient internal features via a calibration stage and then uses these features to construct a noise-driven curriculum that teaches the model when and how to express uncertainty. GUIDE requires no retraining, no architectural modifications, and no manual intermediate-layer selection to the base deep learning model, thus ensuring broad applicability and minimal user intervention. The resulting model avoids distilling overconfidence from the base model, improves out-of-distribution detection ($\approx 77\%$) and adversarial attack detection ($\approx 80\%$), while preserving in-distribution performance. Across diverse benchmarks, GUIDE consistently outperforms state-of-the-art approaches, evidencing the need for actively guiding uncertainty to close the gap between predictive confidence and reliability.

1 INTRODUCTION

Recent advances in artificial intelligence (AI), particularly in medical imaging [Li et al., 2020] and autonomous agents [Rudin et al., 2022], have enabled the deployment of deep learning models into numerous real-world applications. In high-stakes domains, however, such as healthcare [Miotto et al., 2018] and human-in-the-loop robotics [Retzlaff et al., 2024], the reliability and trustworthiness of these models remain a paramount concern. In such contexts, a model must not only recognise the limitations of its predictions but also produce well-calibrated estimates of predictive uncertainty. This requirement is particularly critical in the real world, where distributional shifts from out-of-distribution (OOD) inputs and adversarial attacks that aim to mislead the model are prevalent.

Uncertainty quantification (UQ) provides a principled framework for mitigating these reliability challenges [He and Jiang, 2023]. However, many established UQ approaches either require retraining or modifying the base model, introduce significant computational overhead, or remain vulnerable to misplaced confidence under OOD and adversarial conditions [Mukhoti et al., 2023, Postels et al., 2021]. Fully post-hoc methods are attractive because they can be applied to pretrained models without altering their weights, but existing non-intrusive approaches often estimate uncertainty from frozen representations without explicitly teaching the model when uncertainty should increase [Chen et al., 2019, Shen et al., 2023, Bethell et al., 2024].

We introduce **G**radual **U**ncertainty Ref**i**nement via Noise-**D**riv**e**n Curriculum (GUIDE), a non-intrusive post-hoc evidential meta-model approach that operates on top of a pretrained base model and is explicitly guided on when and how to be uncertain. GUIDE's two-stage approach (Figure 1) initially performs saliency calibration to determine the (frozen) pretrained model's salient intermediate features. Then, it leverages this knowledge to create a gradual noise-driven curriculum to teach the meta-model when to be uncertain and to what magnitude. GUIDE retains and, in some cases, improves the in-distribution (ID) predictive performance (e.g., +16% ID coverage on the CIFAR10 $\rightarrow$ SVHN dataset pairing) while significantly improving the OOD and adversarial attack robustness. We conduct extensive experiments across various ID/OOD datasets, near- and far-OOD scenarios, and both gradient- and non-gradient-based attacks at multiple perturbation strengths. GUIDE achieves state-of-

[*]Equal Contribution

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:413–447.

the-art results in our experimental evaluation, achieving a significant drop in OOD and adversarial predictions compared to state-of-the-art intrusive and non-intrusive post-hoc UQ methods.

Our key contributions are: (1) GUIDE is the first fully post-hoc meta-model approach that explicitly learns when and how to be uncertain by leveraging saliency calibration and noise-driven curriculum learning; (2) theoretical guarantees for GUIDE's soundness and convergence; and (3) extensive experiments demonstrating that GUIDE retains informative structure from the pretrained model while achieving state-of-the-art robustness to OOD and adversarial inputs compared to both intrusive and non-intrusive post-hoc baselines. To the best of our knowledge, GUIDE is the first non-intrusive, fully post-hoc model that reliably estimates predictive uncertainty and explicitly teaches the model when and how to be uncertain within its predictions.

2 RELATED WORK

Uncertainty Quantification. Uncertainty quantification (UQ) is essential for reliable deep learning, especially in safety-critical settings [He and Jiang, 2023]. Epistemic uncertainty, due to limited or biased training data, is typically addressed with Bayesian neural networks [Goan and Fookes, 2020], variational inference [Blei et al., 2017], or Laplace approximations [Fortuin, 2022], though at high cost. More scalable methods, including Monte Carlo Dropout [Gal and Ghahramani, 2016], deep ensembles [Lakshminarayanan et al., 2017], adversarial perturbations [Schweighofer et al., 2023], and distance-aware models [Liu et al., 2020, Van Amersfoort et al., 2020], trade efficiency for expressiveness. In contrast, aleatoric uncertainty, arising from data noise, is typically managed with input-dependent predictors [Kendall and Gal, 2017, Guo et al., 2017], prediction intervals [Khosravi et al., 2010, Tagasovska and Lopez-Paz, 2019, Marco et al., 2025], or generative models [Kingma et al., 2013, Goodfellow et al., 2014], while test-time augmentation [Ayhan and Berens, 2018] is simple but type-agnostic. Since both forms often coexist [He and Jiang, 2023], hybrid methods such as ensembles with prediction intervals [Pearce et al., 2018] or conformal prediction [Angelopoulos et al., 2023, Bethell et al., 2024] aim to capture both uncertainty types, albeit with added cost or calibration overheads.

Post-hoc Methods. Most of the previous approaches require modifying or retraining the base model, limiting their applicability in large-scale or black-box settings [Mukhoti et al., 2023, Postels et al., 2021]. To overcome this, recent work introduced post-hoc uncertainty quantification approaches that act directly on pretrained deterministic models. Notable examples include Monte Carlo Dropout [Gal and Ghahramani, 2016] and intrusive evidential extensions [Sensoy et al., 2018, Franchi et al., 2024, Barker et al., 2025], which attach uncertainty estimation layers and perform light fine-tuning on top of the pretrained backbone.

Non-Intrusive Post-hoc Methods. Recent work avoids retraining or modifying the base model by devising auxiliary models on frozen pretrained representations. Representative examples include Whitebox [Chen et al., 2019] and EMM [Shen et al., 2023], which attach lightweight uncertainty estimators to intermediate or final pretrained features. These non-intrusive methods are broadly applicable and efficient, suitable even in black-box settings without training pipelines. Their efficiency, however, often yields a robustness cost. Their uncertainty estimates are brittle under out-of-distribution (OOD) or adversarial inputs, frequently assigning unwarranted confidence to harmful samples.

GUIDE, unlike existing UQ research, is a fully post-hoc non-intrusive approach that not only produces effective predictive uncertainty estimates but also explicitly advises the model when and how to be uncertain. GUIDE's design yields significantly improved robustness to both OOD shifts and adversarial attacks, distinguishing GUIDE from prior UQ approaches.

3 PRELIMINARIES

We study the standard supervised classification setting, where the objective is to train a predictive model over a finite labelled set. The input domain is $X \subseteq \mathbb{R}^d$ and the output space is $Y = \{1, \ldots, K\}$, corresponding to K discrete classes. The training (ID) data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ comprises samples $(x_i, y_i) \in X \times Y$ drawn independently and identically from the joint distribution $p(x, y) = p(x)\, p(y \mid x)$. The task is to estimate the conditional distribution $p(y \mid x)$.

Beyond accuracy, we also focus on assessing predictive uncertainty. A pretrained deterministic model f_θ trained on ID data, though often accurate, is typically overconfident, especially on OOD and adversarial inputs. This occurs since its softmax outputs $\sigma(f_\theta(x))$ combined with cross-entropy loss promote high confidence even for out-of-distribution samples. GUIDE reduces such overconfidence in a fully post-hoc manner, using only the base model's outputs and features without modifying the structure of the model f or its weights θ, as intrusive changes may be inaccessible, infeasible for large models, or harmful to performance.

4 GUIDE

Our **G**radual **U**ncertainty Refi**n**ement via Noise-**D**riv**e**n Curriculum (GUIDE) meta-model approach (Figure 1) enhances uncertainty calibration and robustness to OOD and adversarial inputs. GUIDE adaptively identifies both salient intermediate layers of the pretrained model to connect them to evidential linear layers and salient weight maps. These weight maps are then exploited to construct a mono-

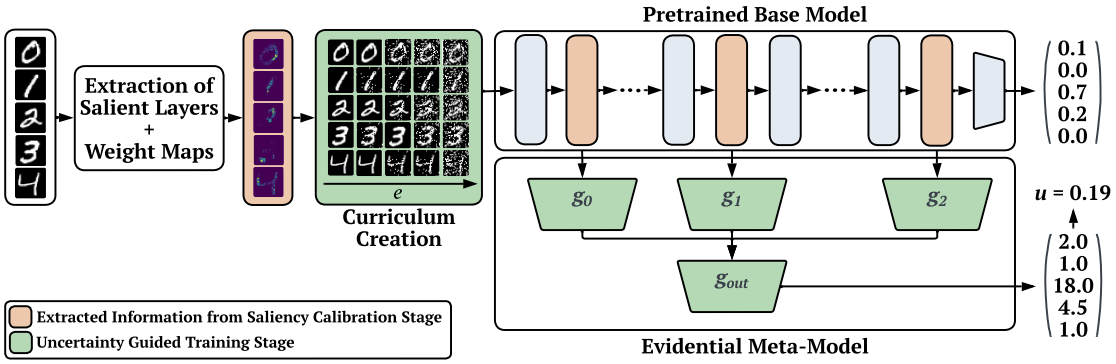


Figure 1: Overview of the GUIDE meta-model approach, showing the Saliency Calibration and Uncertainty Guided Training stages. GUIDE extracts salient features and weight maps from a pretrained model in a fully post-hoc manner, then generates a noise-driven curriculum to teach the meta-model when to be uncertain. In the figure, g_0, g_1, g_2 are evidential projection branches from salient layers, and u is the predicted uncertainty from the Dirichlet evidence.

tonic noise-driven curriculum with progressively perturbed inputs. This curriculum enables GUIDE to learn *when to be uncertain* via monotonic soft-target supervision, which mimics increasing OOD shift and trains the Dirichlet evidence to become less confident as perturbations intensify. This calibration step mitigates the propagation of overconfidence, typically inherited from the pretrained model, and improves robustness under distributional shift. GUIDE combines relevance-based saliency, evidential meta-modelling, and curriculum noise, but does so through a novel soft-target mechanism that directly trains the Dirichlet evidence to become less confident as reliability decreases.

4.1 SALIENCY CALIBRATION

The saliency calibration stage of GUIDE identifies salient components at both the layer and instance levels of the pretrained model. Given the logits of a forward pass $z = f_\theta(x)$ and one-hot vector e_y, we initialise the relevance at the output as $\mathcal{R}_L(x) = z \odot e_y$. Employing the LRP-$\epsilon$ rule [Montavon et al., 2019], we propagate relevance to each hidden layer of the pretrained model for $\ell = L, L-1, \ldots, 1$:

$$\mathcal{R}_{\ell-1}(x) \quad = \quad \Big(\frac{a_{\ell-1}W_\ell^\top}{\langle a_{\ell-1}, W_\ell^\top\rangle + \epsilon}\Big) \quad \odot \quad \mathcal{R}_\ell(x) \quad (1)$$

where $a_{\ell-1}$ are the activations, W_ℓ are the weights, and $\epsilon > 0$ is a small stabilisation term that partially absorbs relevance when there is a contradiction between consecutive layers [1].

This formulation yields relevance maps $\{\mathcal{R}_\ell(x)\}_{l=1}^{L}$ describing how each layer contributes to the final prediction. Each $R_\ell(x)$ is a vector of size $\dim(\ell)$ with each element signifying the relevance attributed to each neuron of the ℓ-th layer. This propagation rule preserves class-relevant evidence and is applicable across standard deep learning components (e.g., convolutional, dense, pooling). We then quantify each hidden layer's global importance by computing the average relevance magnitude over the N input samples:

$$M_\ell = \frac{1}{N}\sum_{i=1}^{N}\frac{||\mathcal{R}_\ell(x_i)||_1}{|\mathcal{R}_\ell(x_i)|} \quad (2)$$

which enables selecting hidden layers that contain sufficient information and can be used to train the evidential meta-model. Specifically, sorting layers by M_ℓ, we select the smallest subset L_{sal} that covers at least a fraction η of the total relevance mass:

$$\sum_{\ell\in L_{\text{sal}}} M_\ell \geq \eta \sum_{\ell} M_\ell \quad (3)$$

where $\eta \in (0, 1]$ is the cumulative relevance coverage threshold. For instance, $\eta = 0.9$ selects the set of layers that collectively account for 90% of relevance. This principled criterion avoids arbitrary layer selection [Shen et al., 2023] and ensures that the evidential meta-model operates on semantically informative representations. (See Appendix B.5 for an attribution ablation showing that Grad-CAM and Integrated Gradients do not preserve η-calibration, unlike LRP.)

Theorem 1. *Let $\eta \in (0, 1]$ denote the saliency coverage threshold, $\kappa \in (0, 1]$ the alignment of saliency scores with the Fisher information mass, and $\rho \in (0, 1]$ the information preserved by the projection. Under local linearisation and restricting to cross-depth correlated blocks, GUIDE's saliency calibration retains at least $\rho\kappa\eta$ of the pretrained Fisher trace with respect to β. Moreover, if the saliency normalisation is stable and the projections are well-conditioned,*

[1] (1) is expressed in vectorised notation for compactness and should be interpreted as an abbreviation of the per-neuron redistribution rule, consistent with prior work [Montavon et al., 2019].

then $\kappa \approx 1$ and $\rho \approx 1$, so that the retained Fisher fraction is essentially determined by η.

The above theoretical statement relies on local linearisation of the frozen model around the calibration data, stable saliency normalisation, and information-preserving projections. These assumptions are used only for the theoretical bound, not as additional requirements of the practical algorithm. The corresponding proof is available in Appendix A. To construct a noise-driven curriculum aligned with salient input features, we define per-input weight maps from the input-level relevance $\mathcal{R}_0(x)$:

$$\mathcal{W}(x)[h,w] = \frac{|\sum_{c=1}^{C} \mathcal{R}_0(x)[h,w,c]|}{\max_{h',w'} |\sum_{c=1}^{C} \mathcal{R}_0(x)[h',w',c]|} \quad (4)$$

Here, C denotes channels and (h, w) spatial indices. The normalisation yields a weight map where higher values correspond to stronger contributions to the model's prediction [Bach et al., 2015], and which can be used to prioritise input feature perturbations.

Both global layer relevance scores M_ℓ and input-level weight maps $\mathcal{W}(x)$ are obtained in a single backward pass: as relevance propagates from output to input, we accumulate per-layer relevance for M_ℓ and retain $\mathcal{R}_0(x)$ for $\mathcal{W}(x)$. This reuse adds no extra forward or backward passes beyond standard relevance propagation [Montavon et al., 2019].

4.2 UNCERTAINTY GUIDED TRAINING

The uncertainty-guided training stage of GUIDE constructs and trains a small Dirichlet-based [Sensoy et al., 2018] meta-model g_ϕ exclusively on features extracted from the selected layers L_{sal} and guides it on when to be uncertain using a monotonic soft-target curriculum. For each selected layer $\ell \in L_{\text{sal}}$, the feature map is flattened $\Phi_\ell(x)$ and projected to $\mathbb{R}^K$ using a branch of multiple linear layers g_ℓ, such that $g_\ell(\Phi_\ell(x)) \in \mathbb{R}^K$. The resulting meta-features from all selected layers, $\{g_\ell(\Phi_\ell(x))\}_{\ell \in L_{\text{sal}}}$, are concatenated into a single vector $\Psi(x)$ and passed through a final linear head g_{out}, yielding $\alpha(x) = \exp(g_{\text{out}}(\Psi(x))) + 1$. Here $\alpha(x)$ parametrises a log-concentrated Dirichlet distribution $\text{Dir}(\pi|\alpha)$ over the categorical label probabilities $\pi = [\pi_1, \ldots, \pi_K] \in \Delta^{K-1}$, where an additive 1 enforces strictly positive parameters as standard in evidential deep learning [Sensoy et al., 2018]. Following standard practice, we compute the scalar uncertainty as $u = K/S$, where $S = \sum_{k=1}^{K} \alpha_k$ is the total Dirichlet strength and K is the number of classes. This strategy enables the meta-model to represent both predictions and epistemic uncertainty. High total evidence S implies confident predictions, while low S reflects uncertainty. All ϕ parameters are learned from the beginning, while the pretrained model's parameters θ are frozen throughout.

To induce robustness and *guide* the meta-model when to be uncertain, we create a targeted noise-driven curriculum. Firstly, a monotonic exponential schedule is constructed $s_t = 1 - e^{-\gamma t}$ for $t \in \{0, 1, \ldots, T\}$, where $s_t \in [0, 1]$ is the target fraction of noise corrupted pixels at stage t, and $\gamma > 0$ is the rate of noise corruption. The first image ($t = 0$) represents the clean view; then, the exponential form ensures fine-grained corruption at early stages (where decision boundaries are sensitive) and coarser granularity at higher noise levels (where decision boundaries should be more sharply defined).

To ensure salient features are targeted first, we leverage the weight maps $\mathcal{W}(x)$ from the saliency calibration stage. A global saliency budget $\tilde{\mathcal{W}} = \frac{1}{HW} \sum_{h,w} \mathcal{W}(x)[h,w]$ is defined to construct a per-pixel corruption probability $p_t(h, w)$ satisfying two conditions: the expected global corruption matches the budget s_t, and pixel-wise probabilities are monotonic in s_t. This is defined as:

$$p_t(h,w) = \begin{cases} \frac{s_t}{\tilde{\mathcal{W}}} \cdot \mathcal{W}(x)[h,w], & s_t \le \tilde{\mathcal{W}} \\ \mathcal{W}(x)[h,w] + \frac{s_t - \tilde{\mathcal{W}}}{1-\tilde{\mathcal{W}}} & \\ \cdot(1 - \mathcal{W}(x)[h,w]), & s_t > \tilde{\mathcal{W}} \end{cases} \quad (5)$$

This ensures that low-noise perturbations are focused on high-saliency regions, while high-noise settings affect the entire image. To actually apply the noise, we sample a single base mask $m \sim U[0,1]^{H \times W}$, reused across all t for a given data point. For each stage t, we modify x and obtain $\tilde{x}_t$:

$$\tilde{x}[h,w] = \begin{cases} 0, & m[h,w] < p_t(h,w)/2, \\ 1, & m[h,w] > 1 - p_t(h,w)/2, \\ x[h,w], & \text{otherwise.} \end{cases} \quad (6)$$

This procedure creates stochastic binary corruption, preserving the expected noise budget per image and ensuring that $\tilde{x}_{t+1}$ is never less noisy than $\tilde{x}_t$. Furthermore, this corruption monotonic in t while $p_t(h, w)$ controls the expected saliency-weighted corruption budget.

To guide the meta-model when to express uncertainty, we construct soft targets that respond to both input corruption and the model's confidence. This part is important to steer the model to be confident for the clean input ($t = 0$) and very uncertain for the noisiest input ($t = T$). For each corrupted view, the base model's predicted confidence in the true class $c_t = \langle \sigma(f_\theta(\tilde{x}_t)), y \rangle$ is used along with the noise level to define an uncertainty target $\tilde{s}_t = s_t \cdot (1 - \frac{1}{2} {c_t}^2)$. The soft target for the meta-model is a convex combination of the uniform distribution and the true label $\tilde{y}_t = \frac{\tilde{s}_t}{K} \cdot \mathbf{1}_K + (1 - \tilde{s}_t) \cdot y$. As formalised in Proposition 2 (Appendix A), this encourages high certainty for confident, low-noise inputs, and near-uniform predictions for highly corrupted or uncertain examples. The soft targets remain monotonic in t in expectation, ensuring consistency in the learned uncertainty behaviour.

Next, we generate a curriculum over corruption strengths. Instead of training the meta-model on all strengths uniformly, we define an epoch-dependent difficulty index $\rho_e = \left(\frac{e}{E-1}\right)^2$ for every epoch $e \in \{0, 1, \ldots, E\}$. At each epoch e, the sampling distribution over the discrete noise level $\{s_0, \ldots, s_T\}$ is given by $\kappa_e(s) \propto (1-\rho_e)(1-s)+\rho_e \cdot s$ and determines which corruption levels populate the training batches. In early epochs ($\rho_e \approx 0$) the sampling distribution concentrates mass on small s, so a higher proportion of clean or mildly corrupted views are trained upon. In late epochs ($\rho_e \to 1$), the distribution shifts towards large s, evidently, strongly corrupted views dominate the training data. Thus, we generate a curriculum that forces a monotonic progression of views and soft targets, from clean to fully corrupted, in which we can guide the meta-model on how and when to be uncertain.

Finally, the meta-model is trained using an uncertainty regularised evidence lower bound (ELBO) combined with a self-rejecting evidence (SRE) penalty to discourage overconfident predictions when the predictive mean disagrees with the soft target. For a mini-batch $\mathcal{B}$, the loss is defined as:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(\tilde{x},\tilde{y})\in\mathcal{B}} \Big[-\sum_{k=1}^{K} \tilde{y}_k (\psi(\alpha_k) - \psi(S)) + \lambda_{\text{kl}} \cdot \text{KL}(\text{Dir}(\alpha)||\text{Dir}(\beta)) \Big] + \underbrace{\frac{S}{K} \cdot (1 - \langle \tilde{y}, \hat{p} \rangle)}_{\text{SRE penalty}} \quad (7)$$

where $\psi(\cdot)$ is the digamma function, $\hat{p}$ is the predictive mean. By weighting the expected log-likelihood by the soft target $\tilde{y}$, it makes evidence sharp for near-one-hot labels, flat/uniform for highly corrupted inputs, and intermediate for moderate noise, giving a graded uncertainty response. The self-rejecting evidence penalty then suppresses any large S that is not supported by target agreement, eliminating misplaced confidence while leaving well-aligned, high-evidence predictions untouched. Combined, these two terms guide the meta-model to be confident only when justified and gradually/monotonically uncertain everywhere else, tightening the gap between in- and out-of-distribution behaviour. The complete GUIDE procedure is detailed in Appendix C.5.

5 EVALUATION

We assess the performance of GUIDE through an extensive set of experiments, benchmarking it against state-of-the-art post-hoc UQ methods across 10 independent trials. The evaluation examines both predictive performance and the quality of uncertainty estimates under OOD shifts and adversarial perturbations. All code and replication material are publicly available in our open-source repository [2].

[2]https://github.com/team-daniel/guide

Datasets. Adopting the procedure on UQ-based evaluation from recent research [Shen et al., 2023, Sensoy et al., 2018], we evaluate all approaches on the MNIST [LeCun et al., 1998], FashionMNIST [Xiao et al., 2017], KMNIST [Clanuwat et al., 2018], EMNIST [Cohen et al., 2017], CIFAR10 [Krizhevsky et al., 2009], CIFAR100 [Krizhevsky et al., 2009], SVHN [Netzer et al., 2011], Oxford Flowers [Netzer et al., 2011], Deep Weeds [Olsen et al., 2019], MNIST-C Mu and Gilmer [2019], and CIFAR-10-C Hendrycks and Dietterich [2019] datasets which were selected to cover a diverse set of domains and challenges. In the following experiments, near-OOD datasets are those that share some degree of class overlap with the ID dataset [Yang et al., 2022].

Comparative Approaches. We compare GUIDE against state-of-the-art post-hoc UQ methods, both intrusive (that change the architecture of the pretrained model) and fully post-hoc: a pretrained deterministic network, ABNN [Franchi et al., 2024], an evidential head, replacing the softmax head for Dirichlet evidence [Sensoy et al., 2018], Whitebox [Chen et al., 2019], and EMM [Shen et al., 2023]. We focus on methods that retrofit uncertainty estimation to a pretrained model. Methods such as Deep Ensembles and MC Dropout Gal and Ghahramani [2016], Bethell et al. [2024] are therefore not included as direct baselines, since they require multiple trained models or dropout-enabled stochastic inference, respectively, and address a different deployment setting from GUIDE's frozen-model formulation. In addition to GUIDE, we evaluate two additional variants to better ablate the effects of each stage of GUIDE. EMM(+cal) refers to EMM that is trained on the salient layers found in GUIDE's uncertainty calibration stage instead of arbitrarily selected layers. EMM(+curric) refers to EMM that is trained upon the curriculum data generated by GUIDE using the standard EDL loss [Shen et al., 2023]. Experimental details, training procedures, hyperparameter choices, adversarial attack configurations, and dataset information are provided in Appendix C.

5.1 CORE RESULTS

For our core set of experiments, we evaluate the accuracy, ID coverage, OOD detection, including both near- and far-OOD settings, adversarial robustness, AUROC, and adversarial AUROC of both GUIDE and its variants, and comparative baselines across a diverse suite of ID/OOD dataset pairs, in order to assess overall reliability under distributional shift and attack. ID/OOD/Adv coverage is computed using a binary rejection threshold based upon the AUROC; see Appendix C.2 for further details. Coverage denotes the fraction of accepted inputs: high ID coverage is desirable, whereas low OOD and adversarial coverage indicate effective rejec-

[3]Details of the maximum L2PGD perturbation for each dataset are discussed in Appendix C.7

Table 1: Mean accuracy, OOD detection, and adversarial attack detection performance with standard deviation of the comparative approaches with a variety of ID and OOD datasets in order of dataset difficulty. The adversarial attack is an L2PGD attack [3]For compactness, we denote adversarial AUROC as AAUROC. Highlighted cells denote the best performance for each metric. * indicates datasets classed as Near-OOD.

	Pretrained	ABNN	EDL-Head	Whitebox	EMM	EMM(+cal)	EMM(+curric)	GUIDE
Type	-	Intrusive	Intrusive	Post-hoc	Post-hoc	Post-hoc	Post-hoc	Post-hoc
MNIST → FashionMNIST								
ID Acc (%) ↑	99.80 ± 0.07	99.78 ± 0.06	99.43 ± 0.17	99.92 ± 0.08	99.55 ± 0.08	99.69 ± 0.09	99.21 ± 0.25	99.87 ± 0.04
ID Cov (%) ↑	79.26 ± 4.36	74.88 ± 7.98	77.44 ± 3.24	76.38 ± 5.96	90.58 ± 4.02	92.32 ± 1.57	79.46 ± 10.25	87.05 ± 1.63
OOD Cov (%) ↓	24.89 ± 9.71	18.23 ± 8.60	22.51 ± 5.50	13.50 ± 2.72	31.87 ± 26.23	35.69 ± 17.35	40.18 ± 14.28	7.34 ± 3.42
Adv Cov (%) ↓	25.84 ± 7.84	24.35 ± 9.67	13.86 ± 2.87	9.04 ± 4.21	22.18 ± 22.49	55.47 ± 13.92	62.11 ± 15.18	4.60 ± 2.69
AUROC (%) ↑	84.18 ± 5.93	85.93 ± 2.11	83.23 ± 3.88	88.38 ± 1.95	77.68 ± 20.60	74.54 ± 14.51	72.85 ± 8.03	94.85 ± 1.27
AAUROC (%) ↑	83.67 ± 3.37	81.81 ± 3.95	88.73 ± 2.64	90.26 ± 0.59	83.43 ± 17.95	53.56 ± 13.23	53.62 ± 9.93	95.72 ± 0.91
MNIST → KMNIST								
ID Acc (%) ↑	99.81 ± 0.06	99.76 ± 0.05	99.57 ± 0.11	99.87 ± 0.04	99.51 ± 0.12	99.56 ± 0.10	99.15 ± 0.20	99.75 ± 0.08
ID Cov (%) ↑	80.69 ± 2.47	85.91 ± 1.67	75.03 ± 2.90	85.93 ± 0.85	92.48 ± 1.07	94.49 ± 1.36	86.67 ± 4.09	89.80 ± 1.10
OOD Cov (%) ↓	18.23 ± 2.14	22.90 ± 2.60	18.54 ± 4.31	8.63 ± 1.61	18.47 ± 7.35	12.19 ± 5.75	32.31 ± 4.48	6.78 ± 1.23
Adv Cov (%) ↓	38.35 ± 3.37	46.27 ± 3.75	17.78 ± 4.47	8.60 ± 1.65	50.97 ± 24.98	49.89 ± 16.60	67.51 ± 4.79	9.00 ± 2.66
AUROC (%) ↑	88.27 ± 0.93	86.79 ± 1.33	85.38 ± 2.08	94.64 ± 0.66	89.67 ± 4.94	94.07 ± 3.59	82.62 ± 2.77	96.68 ± 0.44
AAUROC (%) ↑	77.20 ± 1.29	74.87 ± 2.17	85.63 ± 2.53	94.42 ± 0.62	56.29 ± 24.20	61.42 ± 14.67	52.33 ± 5.29	95.80 ± 0.84
MNIST → EMNIST*								
ID Acc (%) ↑	99.80 ± 0.04	99.71 ± 0.06	99.49 ± 0.09	99.89 ± 0.05	99.61 ± 0.11	99.65 ± 0.14	99.01 ± 0.21	99.89 ± 0.05
ID Cov (%) ↑	81.64 ± 1.13	86.79 ± 0.93	76.29 ± 3.31	84.10 ± 1.60	89.58 ± 3.14	92.65 ± 1.69	87.59 ± 4.23	84.94 ± 1.30
OOD Cov (%) ↓	24.09 ± 1.55	28.65 ± 2.06	22.98 ± 2.97	17.59 ± 1.05	35.97 ± 8.39	25.80 ± 6.25	41.71 ± 5.17	15.82 ± 2.41
Adv Cov (%) ↓	31.60 ± 1.52	43.40 ± 3.01	17.54 ± 2.66	12.23 ± 2.12	51.21 ± 19.84	42.90 ± 11.06	59.73 ± 6.29	10.05 ± 3.87
AUROC (%) ↑	85.46 ± 0.97	83.57 ± 1.02	83.25 ± 1.67	89.47 ± 0.68	77.37 ± 6.66	84.72 ± 4.61	75.11 ± 4.28	91.08 ± 1.09
AAUROC (%) ↑	82.02 ± 1.08	76.47 ± 1.65	86.13 ± 1.83	91.96 ± 1.03	56.72 ± 19.78	65.67 ± 11.07	57.48 ± 5.52	93.76 ± 1.22
CIFAR10 → SVHN								
ID Acc (%) ↑	93.27 ± 1.56	96.02 ± 0.37	88.62 ± 1.37	90.14 ± 1.13	76.30 ± 16.34	92.07 ± 2.19	91.56 ± 3.18	88.46 ± 2.55
ID Cov (%) ↑	64.48 ± 6.65	63.94 ± 2.02	66.26 ± 8.93	69.71 ± 4.37	66.52 ± 17.63	60.54 ± 13.51	37.27 ± 22.46	80.65 ± 9.43
OOD Cov (%) ↓	12.03 ± 3.38	10.85 ± 2.12	23.55 ± 5.40	17.89 ± 4.08	16.43 ± 9.18	21.62 ± 8.01	13.70 ± 13.55	19.34 ± 6.03
Adv Cov (%) ↓	66.01 ± 15.98	71.05 ± 10.63	24.78 ± 5.48	37.06 ± 6.61	42.27 ± 23.38	63.74 ± 9.66	39.64 ± 22.33	48.04 ± 12.76
AUROC (%) ↑	81.96 ± 3.13	79.58 ± 1.98	77.57 ± 5.95	81.95 ± 3.77	78.76 ± 14.66	74.86 ± 7.81	56.02 ± 15.52	89.36 ± 5.76
AAUROC (%) ↑	57.95 ± 4.60	41.44 ± 4.58	76.78 ± 6.24	71.07 ± 3.98	69.86 ± 5.17	50.97 ± 11.69	42.66 ± 8.81	78.07 ± 6.74
CIFAR10 → CIFAR100*								
ID Acc (%) ↑	92.27 ± 1.48	93.48 ± 1.14	87.22 ± 1.44	83.49 ± 3.40	81.09 ± 17.11	89.20 ± 0.94	65.96 ± 32.04	88.90 ± 0.88
ID Cov (%) ↑	51.94 ± 3.23	53.69 ± 4.42	51.69 ± 4.72	56.06 ± 4.74	59.95 ± 2.09	57.42 ± 5.01	50.58 ± 26.52	56.67 ± 5.50
OOD Cov (%) ↓	20.97 ± 2.59	20.17 ± 4.33	28.70 ± 3.91	25.29 ± 3.70	27.05 ± 3.25	24.50 ± 3.81	40.52 ± 30.37	24.01 ± 4.18
Adv Cov (%) ↓	24.79 ± 4.37	28.12 ± 7.77	22.75 ± 4.73	17.17 ± 3.54	27.85 ± 4.71	23.81 ± 4.66	37.41 ± 29.41	17.09 ± 3.99
AUROC (%) ↑	70.41 ± 1.34	71.81 ± 0.63	65.66 ± 1.16	70.09 ± 2.68	71.81 ± 1.98	72.09 ± 1.29	55.55 ± 4.54	71.61 ± 0.97
AAUROC (%) ↑	65.24 ± 1.07	64.50 ± 0.95	69.90 ± 0.71	75.96 ± 2.78	70.86 ± 1.98	71.81 ± 0.82	57.89 ± 3.87	76.01 ± 0.68
Oxford Flowers (low-shot) → Deep Weeds								
ID Acc (%) ↑	99.14 ± 0.41	98.86 ± 0.69	84.11 ± 3.25	97.22 ± 3.00	87.69 ± 9.58	92.70 ± 13.03	79.66 ± 7.82	90.84 ± 1.58
ID Cov (%) ↑	45.12 ± 13.73	74.24 ± 7.00	57.57 ± 34.86	33.80 ± 28.70	49.10 ± 15.84	62.44 ± 12.40	48.52 ± 13.89	74.98 ± 6.10
OOD Cov (%) ↓	18.95 ± 15.01	19.58 ± 13.66	35.60 ± 37.80	13.62 ± 17.18	10.50 ± 7.71	13.71 ± 7.13	10.49 ± 4.38	19.52 ± 8.68
Adv Cov (%) ↓	7.92 ± 8.56	11.78 ± 9.47	32.33 ± 39.15	9.35 ± 13.48	2.76 ± 2.39	3.71 ± 2.41	3.40 ± 1.32	9.74 ± 5.61
AUROC (%) ↑	66.77 ± 9.15	82.44 ± 9.28	59.52 ± 19.29	53.86 ± 21.07	69.61 ± 10.05	78.72 ± 5.85	66.00 ± 10.78	81.63 ± 9.34
AAUROC (%) ↑	75.76 ± 7.38	86.31 ± 6.90	63.55 ± 18.24	57.53 ± 21.53	78.67 ± 10.35	87.62 ± 4.50	73.53 ± 8.79	88.56 ± 6.94

tion of shifted or attacked samples [Geifman and El-Yaniv, 2017, 2019]. These results are summarised in Table 1.

Firstly, across all datasets, every approach (including GUIDE and its variants) maintains near-ceiling accuracy across all datasets (e.g., 90-99%), showing that none of the UQ approaches compromise classification performance on clean ID data downstream from the pretrained model. This finding concretely shows that any observed gains in robustness through GUIDE are not due to degradation in performance, with the comparatively lower CIFAR10→CIFAR100 accuracy reflecting the harder natural-image setting and the post-hoc objective's stronger evidence tempering rather than a general loss of ID performance.

In terms of ID coverage (the proportion of ID inputs retained after abstention), the pretrained base model rarely exceeds 80% (e.g., 79.26% on MNIST → FashionMNIST). Intrusive methods such as ABNN raise coverage (up to 85.91%) but retain high OOD ($\geq 20\%$) and adversarial inputs. Post-hoc baselines like EMM push ID coverage above 90%, yet still admit excessive adversarial examples ($\geq 20\%$). In contrast, GUIDE achieves a better balance: strong ID coverage ($\sim 87\%$) while cutting OOD to $\leq 8\%$ and adversarial to $\leq 5\%$, delivering reliable abstention without sacrificing retention.

Intrusive methods such as ABNN often fail to reject harmful inputs, with OOD coverage exceeding 20% and adversarial coverage surpassing 40% across multiple benchmarks (e.g., 46.27% adversarial coverage on MNIST → KMNIST). Post-hoc baselines like EMM show some improvement, but still retain a substantial fraction of adversarial inputs ($\geq 20\%$). In contrast, GUIDE consistently reduces OOD coverage below 10% (e.g., 6.78% on MNIST → KMNIST) and adversarial coverage below 5% (e.g., 4.60% on MNIST →

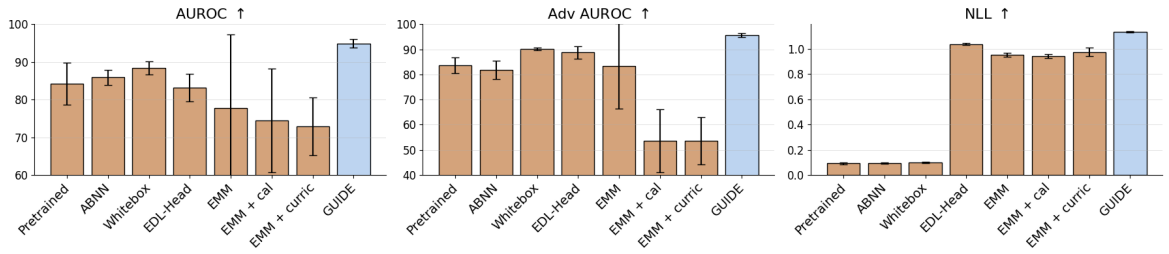


Figure 2: AUROC, adversarial AUROC, and negative log likelihood across comparative approaches where the ID → OOD dataset is MNIST and FashionMNIST. High AUROC and NLL is desirable, to show better ID → OOD/Adv separation and reduced overconfidence. GUIDE is in blue.

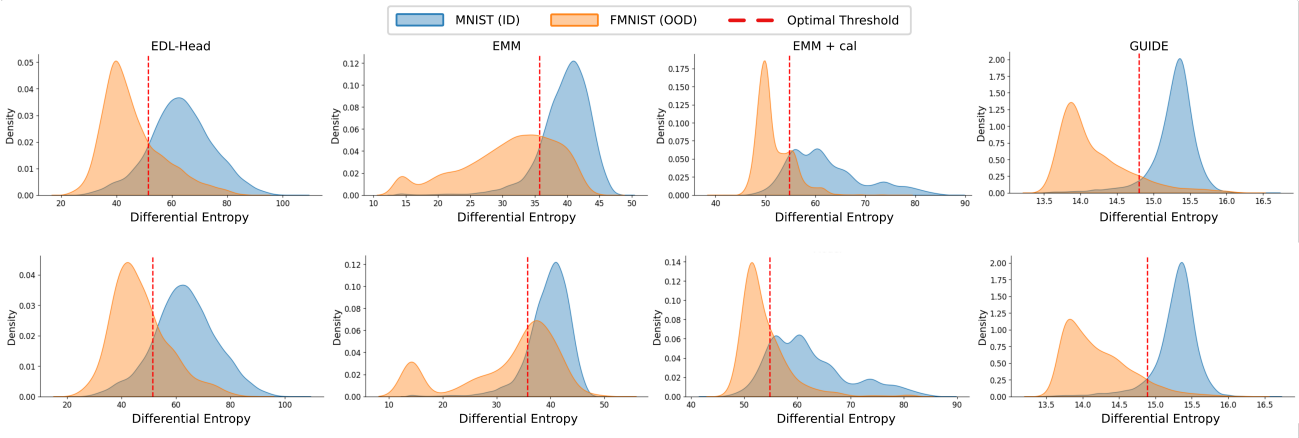


Figure 3: Visualised AUROC plots (top row) and Adv AUROC plots (bottom row), including the binary decision threshold for OOD/Adv rejection, for comparative methods where the ID dataset is MNIST and the OOD dataset is FashionMNIST.

FashionMNIST), highlighting its ability to reliably abstain on uncertain or adversarial inputs while preserving ID performance.

In terms of AUROC (measure of ID/OOD separability), post-hoc baselines such as Whitebox and EMM achieve moderate separation between ID and OOD inputs (e.g., $\approx 77\%$ AUROC on MNIST → FashionMNIST), but their adversarial AUROC remains inconsistent ($\leq 85\%$). Intrusive methods like ABNN perform similarly, rarely exceeding 86% AUROC. In contrast, GUIDE consistently delivers near-perfect discrimination, achieving 94.85% AUROC and 95.72% adversarial AUROC on MNIST → FashionMNIST, and above 95% on MNIST → KMNIST. These results highlight GUIDE's ability to sharply distinguish in-distribution from both OOD and adversarial inputs, substantially outperforming prior intrusive and post-hoc methods. GUIDE is also competitive on the two comparatively weaker pairs, CIFAR10→CIFAR100 and Oxford Flowers→DeepWeeds, where the best standard AUROC is achieved by another baseline but the margin is small. On CIFAR10→CIFAR100, EMM(+cal) obtains the highest AUROC (72.09%), while GUIDE remains close (71.61%) and achieves stronger adversarial separation, with higher adversarial AUROC (76.01% vs. 71.81%) and lower adversarial coverage (17.09% vs. 23.81%). Similarly, on Oxford Flowers→DeepWeeds, ABNN attains the highest AUROC (82.44%), but GUIDE remains close (81.63%) and achieves the strongest adversarial AUROC (88.56%). These pairs are difficult because the OOD samples are visually or semantically close to the ID data, compressing the uncertainty-score distributions. Thus, even where GUIDE is not best on standard AUROC, it preserves a favourable robustness trade-off by remaining near-best on OOD separation while improving adversarial separation and rejection.

We further ablate GUIDE with two variants: EMM(+cal), which uses GUIDE's saliency-based calibration for layer selection, and EMM(+curric), which applies the noise-driven curriculum without soft targets or custom loss. EMM(+cal) yields modest gains over EMM, showing that informed feature selection helps moderately. EMM(+curric) often degrades performance, as corrupted views without guidance induce over- or underconfidence. GUIDE, combining all components, consistently outperforms both, demonstrating that all components are essential.

As shown in Figure 2, existing methods vary across AUROC, adversarial AUROC, and NLL (higher NLL indicates less overconfidence). Intrusive models like ABNN improve over

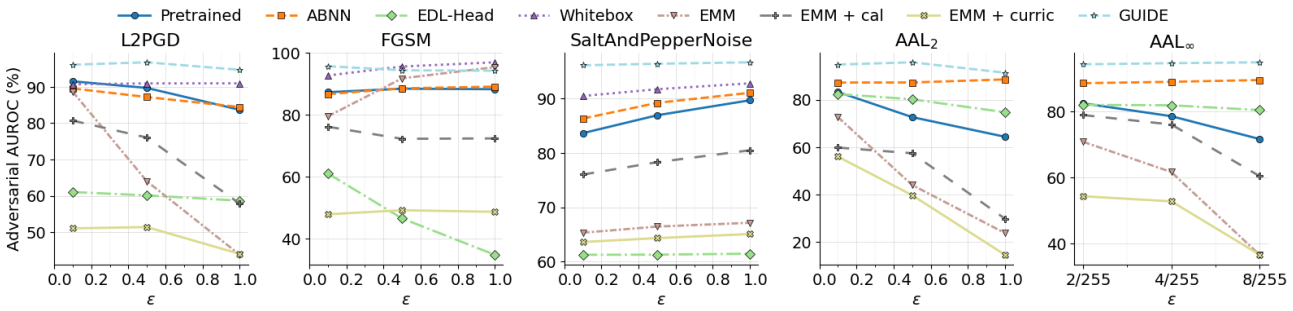


Figure 4: Adversarial AUROC (%) across varying perturbation strengths (ϵ) for three attack types (L2PGD, FGSM, Salt and Pepper noise, AutoAttack L_2, and AutoAttack L_∞) where the ID $\rightarrow$ OOD dataset is MNIST and FashionMNIST. Higher AUROC indicates better robustness.

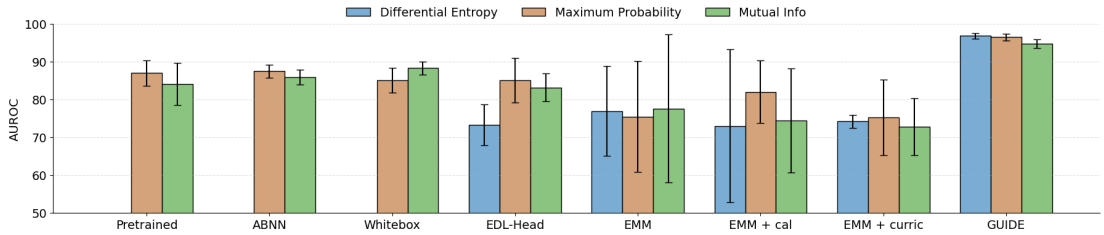


Figure 5: AUROC for different ID-OOD threshold metrics (differential entropy, maximum probability, mutual information) where the ID dataset is MNIST, and the OOD dataset is FashionMNIST.

baseline but plateau below 90%, while post-hoc methods such as Whitebox achieve competitive AUROC yet mainly reduce overconfidence, reflected in higher NLL. GUIDE reaches near-95% AUROC and adversarial AUROC with the highest NLL, offering reliable uncertainty. Figure 3 confirms this; GUIDE shows the clearest ID/OOD separation, unlike the overlapping curves of EDL-Head and EMM.

Calibration curves illustrating the limits of existing methods can be seen in Appendix B.1, Fig. 11. The pretrained model and EDL-Head fall below the diagonal on ID data, indicating overconfidence (smECE $= 0.317$ and 0.265), and assign high probabilities to OOD inputs. EMM, by contrast, lies above the diagonal, underconfident on ID data (smECE $= 0.193$) yet still overly confident on OOD inputs. GUIDE instead aligns closely with the diagonal (smECE $= 0.061$) while keeping OOD probabilities low, demonstrating reliable calibration and effective OOD rejection. Plots showing the drop in predictive confidence between ID and shifted inputs can be seen in Appendix B.1, Fig. 12. The pretrained model, ABNN, and Whitebox exhibit almost no drop, remaining overconfident on OOD and adversarial data. EDL-Head and EMM achieve larger drops but are unstable. In contrast, GUIDE yields the largest and most consistent drops across all datasets, demonstrating its ability to retain confidence on ID samples while reducing it on harmful inputs.

Our evaluation highlights a key distinction between intrusive and post-hoc methods. Intrusive approaches like ABNN and EDL-Head can achieve strong AUROC but require modifying or retraining the base model, which is often impractical in large-scale or black-box settings. Post-hoc methods are more general since they operate on frozen networks but tend to be unstable. GUIDE bridges this gap, retaining post-hoc accessibility while matching or surpassing intrusive baselines in robustness.

5.2 ADVERSARIAL ATTACK ANALYSIS

Figure 4 shows adversarial AUROC across perturbation strengths ϵ for L2PGD, FGSM, Salt and Pepper noise, AutoAttack L_2, and AutoAttack L_∞ attacks. GUIDE sustains high AUROC ($\geq 90\%$) across all perturbations, while pretrained, ABNN, and EMM models often drop below 70%. Robustness holds even under non-gradient Salt-and-Pepper noise, highlighting the effectiveness of GUIDE's saliency-calibrated curriculum in preserving adversarial separability across both gradient-based and gradient-free settings. Here, the evaluated attacks stress different failure modes: L2PGD and FGSM use gradient information, Salt and Pepper noise tests sparse non-gradient corruption, and AutoAttack provides a stronger combined robustness check under both L_2 and L_∞ constraints. GUIDE remains stable across these settings, suggesting that its adversarial detection performance is not tied to a single perturbation type or attack objective. The full adversarial results are repor-

ted in Appendix B.2, where GUIDE achieves the strongest or near-strongest adversarial AUROC across the evaluated perturbation strengths.

5.3 THRESHOLD ANALYSIS

To assess robustness across uncertainty metrics, we compare AUROC under differential entropy, maximum probability, and mutual information (Figure 5). Pretrained, ABNN, and Whitebox models show moderate but inconsistent performance, while EDL-Head and EMM are even less stable, often dropping below 80% with wide error bars. In contrast, GUIDE maintains consistently high AUROC ($\geq 95\%$) with minimal variance, demonstrating reliable uncertainty estimates. This stability shows that GUIDE's gains are not tied to a single rejection metric. Since maximum probability captures predictive confidence, mutual information reflects epistemic uncertainty, and differential entropy measures Dirichlet dispersion, strong performance across all three indicates robust separation of ID, OOD, and adversarial inputs. Full threshold results are provided in Appendix B.3.

5.4 ADDITIONAL ANALYSIS OVERVIEW

The remaining appendix analyses support the central conclusion that GUIDE's gains arise from the interaction between calibrated feature access, uncertainty-guided supervision, and lightweight post-hoc deployment, rather than from a single favourable experimental setting. First, the saliency-calibration results show that GUIDE's layer selection is not arbitrary: the achieved relevance coverage closely tracks or exceeds the target η, while the selected layers differ from manually chosen EMM layers in ways that better reflect the pretrained model's decision-relevant structure (Appendix B.1) This supports the role of saliency calibration as a principled mechanism for deciding which (frozen) representations should be exposed to the evidential meta-model, rather than treating intermediate-layer choice as a manual design decision.

Second, we illustrate how the uncertainty-guided curriculum behaves in practice. The qualitative curriculum examples and soft-target surface show that weakly corrupted, high-confidence inputs retain sharp targets, while increasingly corrupted or ambiguous inputs move toward uncertainty-aware supervision without collapsing into uninformative labels (Appendix B.1). This is consistent with the endpoint behaviour formalised in Appendix A, Proposition 2: GUIDE learns high evidence when the target is reliable and lower evidence as supervision approaches the uniform distribution.

Third, the ablation and attribution-comparison studies indicate that GUIDE is robust to reasonable design choices while still benefiting from the full pipeline. The ablations over curriculum length T, corruption rate γ, and saliency threshold η show stable performance across a range of settings, with the chosen defaults providing a strong balance between ID retention, OOD rejection, adversarial rejection, and computational cost (Appendix B.4, Tables 6, 7, and 8). At the same time, replacing LRP-ϵ with Grad-CAM or Integrated Gradients weakens coverage-controlled layer selection, showing that GUIDE's calibration stage depends on having a relevance signal that is comparable across layers (Appendix B.5, Figure 16).

Finally, the corrupted-distribution and runtime analyses clarify both the difficulty and practicality of the setting. On MNIST-C and CIFAR-10-C, the shifted samples remain visually close to the ID distribution, making separation more difficult than in standard ID→OOD transfers; nevertheless, GUIDE remains competitive and often strongest, particularly for adversarial AUROC (Appendix B.6, Table 9). The runtime analysis shows that GUIDE's main additional cost is the offline saliency-calibration stage, while inference only requires a forward pass through the frozen backbone and the lightweight evidential meta-model (Appendix C.8, Table 11). Overall, these results strengthen the main conclusion: GUIDE improves uncertainty reliability by combining calibrated feature selection, guided uncertainty supervision, and a practical post-hoc deployment path.

6 CONCLUSION AND FUTURE WORK

GUIDE is a post-hoc evidential meta-model that attaches to any pretrained classifier to explicitly learns when and how much to be uncertain. Through saliency-driven calibration, a noise-based curriculum, and an uncertainty-aware loss, GUIDE achieves reliable ID→OOD/Adversarial separation without retraining or altering the base model. Extensive experiments across diverse datasets and adversarial settings show that GUIDE consistently outperforms both intrusive and post-hoc baselines in AUROC, coverage, and calibration. Due to its lightweight and architecture-agnostic foundation, GUIDE is practical for real-world deployment. Future work includes extending GUIDE to regression and structured prediction tasks, integrating it with active learning and safe decision-making, exploring more scalable aggregation mechanisms, such as relevance-weighted or attention-based pooling for larger backbones, and evaluating its applicability to ImageNet-scale datasets and transformer-based architectures such as ViTs.

Acknowledgements

This work was supported the European Union's Horizon Europe research and innovation programme under grant agreement 101168067 (GuardAI - Enhancing Robustness and Security of Edge AI Systems for Safety-Critical Applications).

References

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

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Charmaine Barker, Daniel Bethell, and Simos Gerasimou. Quantifying adversarial uncertainty in evidential deep learning using conflict resolution. *arXiv preprint arXiv:2506.05937*, 2025.

Daniel Bethell, Simos Gerasimou, and Radu Calinescu. Robust uncertainty quantification using conformalised monte carlo prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 20939–20948, 2024.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. Confidence scoring using whitebox meta-models with linear classifier probes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1467–1475. PMLR, 2019.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Coint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 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.

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

Gianni Franchi, Olivier Laurent, Maxence Leguéry, Andrei Bursuc, Andrea Pilzer, and Angela Yao. Make me a bnn: A simple strategy for estimating bayesian uncertainty from pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12194–12204, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international cCnference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf`.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR, 2019.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87, 2020.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Wenchong He and Z Jiang. A survey on uncertainty quantification methods for deep neural networks: An uncertainty source perspective. *Perspective*, 1:88, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F Atiya. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Adriel Sosa Marco, John Daniel Kirwan, Alexia Toumpa, and Simos Gerasimou. Uncertainty quantification for deep regression using contextualised normalizing flows. *arXiv preprint arXiv:2512.00835*, 2025.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24384–24394, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, page 4. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, 9(1):2058, 2019.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning*, pages 4075–4084. PMLR, 2018.

Janis Postels, Mattia Segu, Tao Sun, Luca Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649*, 2021.

Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. In *AAAI*, volume 34, pages 11890–11898, 2020.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E Taylor, and Andreas Holzinger. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *Journal of Artificial Intelligence Research*, 79:359–415, 2024.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on robot learning*, pages 91–100. PMLR, 2022.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. *Advances*

in Neural Information Processing Systems, 36:19446–19484, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9772–9781, 2023.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

A THEORETICAL ANALYSIS

This appendix provides the proofs to complement Theorem 1 presented in Section 4. More specifically, we provide a proof for how much information is retained from the pretrained model during GUIDE's saliency-based calibrations stage.

A.1 THEOREM 1

Theorem 1. Let $\eta \in (0, 1]$ denote the saliency coverage threshold, $\kappa \in (0, 1]$ the alignment of saliency scores with the Fisher information mass, and $\rho \in (0, 1]$ the information preserved by the projection. Under local linearisation and restricting to cross-depth correlated blocks, GUIDE's saliency calibration retains at least $\rho\kappa\eta$ of the pretrained Fisher trace with respect to β. Moreover, if the saliency normalisation is stable and the projections are well-conditioned, then $\kappa \approx 1$ and $\rho \approx 1$, so that the retained Fisher fraction is essentially determined by η.

Proof. Let $f_\theta : \mathcal{X} \to \mathbb{R}^K$ be a pretrained network with frozen parameters θ. On the train dataset $\mathcal{D}$, work in the standard local linear regime:

$$z(x) = b + \sum_{\ell=1}^{L} W_\ell \Phi_\ell(x) \quad \text{in } L^2(\mathcal{D}) \tag{8}$$

where $\Phi_\ell(x)$ are layer features (larger ℓ is closer to the networks output). Two layers are said to be cross-depth correlated if their feature covariance is non-zero. Form the undirected graph whose vertices are layers and whose edges connect pairs (i, j) with $\mathrm{Cov}(\Phi_i, \Phi_j) \neq 0$. Let $\mathcal{I}_{\text{corr}}$ be the union of connected components that contain at least one edge spanning different depths. All "sum over all layers" denominators below sum only over $\mathcal{I}_{\text{corr}}$, so purely uncorrelated blocks are ignored by design.

GUIDE computes non-negative saliencies M_ℓ, sorts layers by M_ℓ in descending order, and selects the smallest prefix L_{sal} that satisfies the coverage constraint in Equation 3. For each selected layer $\ell \in \mathcal{S} := L_{\text{sal}} \cap \mathcal{I}_{\text{corr}}$ in the correlated set, the branch $P_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^K$ produces $\psi_\ell = P_\ell \Phi_\ell$, and Ψ concatenates $\{\psi_\ell\}_{\ell \in \mathcal{S}}$. For analysis, we replace Φ_ℓ by its innovation $\tilde{\Phi}_\ell$ and study $\psi_\ell = P_\ell \tilde{\Phi}_\ell$; this removes components that are linearly predictable from lower-depth information and therefore can only decrease Fisher, yielding a valid lower bound. [4]

Next is to ensure that the features from different correlated layers contribute non-redundant information to the Fisher calculation. To do this, we construct innovations by processing the correlated layers in order of depth. We work in the Hilbert space $L^2(\Omega, \mathbb{P})$ with inner product $\langle A, B \rangle = \mathbb{E}[A^\top B]$. Let the indices in $\mathcal{I}_{\text{corr}}$ be ordered from lower to higher depth, $i_1 > i_2 > \ldots > i_m$ (so i_1 is closest to the output). At each step t, we remove from Φ_{i_t} the part that can be explained as a linear combination of the innovations from all later depths (i.e. those processed earlier in the ordering). Formally, if $\prod_{t-1}$ denotes the orthogonal projector in L^2 onto $\mathrm{span}\{\tilde{\Phi}_{i_s} : s < t\}$, then the innovation at depth i_t is defined by $\tilde{\Phi}_{i_t} := \Phi_{i_t} - \prod_{t-1} \Phi_{i_t}$. By construction, innovations from different depths are orthogonal in L^2, meaning $\mathbb{E}[\tilde{\Phi}_{i_s} \tilde{\Phi}_{i_t}^\top] = 0$ whenever $s \neq t$. We denote by $\sum_{i_t}^{\text{inn}} = \mathrm{Cov}(\tilde{\Phi}_{i_t}) \succeq 0$ the covariance of the innovation at depth i_t. This orthogonalisation isolates the new, linearly independent contribution of each correlated layer after accounting for deeper ones. It makes the Fisher matrix block-diagonal, so the total trace is just the sum over innovations; without it, cross-layer covariances would produce off-diagonal terms and break this additivity.

We model the GUIDE setting as a local Gaussian parametric experiment, in which the output vector U depends linearly on the depth-ordered innovations, with additive Gaussian noise of fixed covariance:

$$U \mid \tilde{\Phi} \sim \mathcal{N}\big(b + \sum_{t=1}^{m} B_{i_t} \tilde{\Phi}_{i_t}, \Sigma_\epsilon\big) \qquad \Sigma_\epsilon \succeq 0 \tag{9}$$

The unknown parameter $\beta := \{B_{i_t}\}_{t=1}^m$ i.e., the collection of block matrices mapping each innovation $\tilde{\Phi}_{i_t}$ to the output. This model makes the classical Fisher information exactly computable from the innovation covariances $\sum_{i_t}^{\text{inn}}$, and constant

[4] Each branch $g_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^K$ is linear in our method; we therefore write $g_\ell(\Phi_\ell) = P_\ell \Phi_\ell, \quad P_\ell \in \mathbb{R}^{K \times d_\ell}$. If g_ℓ includes non-linearities, all Fisher bounds below apply to its local linearisation, with P_ℓ denoting the corresponding linear operator.

factors from Σ_ϵ cancel in all Fisher ratios of interest. Because the innovations are uncorrelated across t, the Fisher matrix is block diagonal $\mathcal{I}(\beta) = \text{blkdiag}_t(\Sigma_\epsilon^{-1} \otimes \sum_{i_t}^{\text{inn}})$ whose trace gives the total Fisher from all correlated blocks:

$$\mathfrak{F}_{\text{all}} = \text{tr}\, \mathcal{I}(\beta) = \text{tr}(\Sigma_\epsilon^{-1}) \sum_{\ell \in \mathcal{I}_{\text{corr}}} \text{tr}(\Sigma_\ell^{\text{inn}}) \tag{10}$$

The meta-model only receives features from the selected layers $\mathcal{S} = L_{\text{sal}} \cap \mathcal{I}_{\text{corr}}$. For each such layer ℓ, the innovation $\tilde{\Phi}_\ell$ is first projected through the learned branch P_ℓ to give $\psi_\ell = P_\ell \tilde{\Phi}_\ell$ for $\ell \in \mathcal{S}$. Parametrising:

$$U \mid \psi \sim \mathcal{N}\big(b + \sum_{\ell \in \mathcal{S}} B_{i_t} C_\ell \psi_\ell, \Sigma_\epsilon\big) \qquad \Sigma_\epsilon \succeq 0 \tag{11}$$

with $\{C_\ell\}_{\ell \in \mathcal{S}}$ in place of $\{B_\ell\}$- one matrix per projected branch- gives:

$$\mathcal{I}_\Psi(C) = \text{blkdiag}_{\ell \in \mathcal{S}}(\Sigma_\epsilon^{-1} \otimes P_\ell \Sigma_\ell^{\text{inn}} P_\ell^\top), \qquad \mathfrak{F}_\Psi = \text{tr}\, \mathcal{I}_\Psi(C) = \text{tr}(\Sigma_\epsilon^{-1}) \sum_{\ell \in \mathcal{S}} \text{tr}(P_\ell \Sigma_\ell^{\text{inn}} P_\ell^\top) \tag{12}$$

where $\mathfrak{F}_\Psi$ is the total Fisher content accessible to the meta-model. It is important to measure, in worst-case relative terms, how much of the per-layer Fisher content is preserved by the projection. We define the projection–retention factor:

$$\rho := \inf_{\ell \in \mathcal{S}} \frac{\text{tr}(P_\ell \Sigma_\ell^{\text{inn}} P_\ell^\top)}{\text{tr}(\Sigma_\ell^{\text{inn}})} \quad \in (0, 1] \tag{13}$$

which quantifies the minimum per-branch retention of Fisher information after projection. By construction, ρ is the smallest fraction of Fisher (in trace form) that any selected branch retains compared to the unprojected innovation. From this definition, the total projected Fisher content can be bounded as:

$$\sum_{\ell \in \mathcal{S}} \text{tr}(P_\ell \Sigma_\ell^{\text{inn}} P_\ell^\top) \geq \rho \sum_{\ell \in \mathcal{S}} \text{tr}(\Sigma_\ell^{\text{inn}})) \tag{14}$$

This inequality states that, in aggregate, the selected meta-model branches retain at least a fraction ρ of the Fisher content they would have had without projection.

For the purposes of the bound, the *true* per-layer contribution to the Fisher trace is measured by $M_\ell^* = \text{tr}(\Sigma_\ell^{\text{inn}})$, which we refer to as the ideal Fisher weight of layer ℓ. In practice, GUIDE does not have access to M_ℓ^* and instead uses computed emperical saliency scores M_ℓ to select the subset L_{sal} via the coverage rule described earlier. To connect the empirical selection to the ideal Fisher weighting, we assume a calibrated coverage condition:

$$\sum_{\ell \in \mathcal{S}} M_\ell^* \geq \kappa\, \eta \sum_{\ell \in \mathcal{I}_{\text{corr}}} M_\ell^* \qquad \kappa \in (0, 1] \tag{15}$$

where η is the coverage threshold from the selection rule, and κ quantifies any mismatch between the empirical saliencies and the ideal Fisher weights. When the base network is exactly linear (or locally linear with fixed gates and controlled rescalings), the empirical saliencies are proportional to M_ℓ^* and $\kappa = 1$. In more general settings, κ may be smaller; in that case, it can be treated as a one-off model-dependent calibration factor, absorbed into the effective coverage level η.

Finally, since both $\mathfrak{F}_\Psi$ and $\mathfrak{F}_{\text{all}}$ include the factor $\text{tr}(\Sigma_\epsilon^{-1})$, it cancels in the ratio. By the definition of ρ,

$$\frac{\sum_{\ell \in \mathcal{S}} \text{tr}(P_\ell \Sigma_\ell^{\text{inn}} P_\ell^\top)}{\sum_{\ell \in \mathcal{I}_{\text{corr}}} \text{tr}(\Sigma_\ell^{\text{inn}})} \geq \rho \cdot \frac{\sum_{\ell \in \mathcal{S}} \text{tr}(\Sigma_\ell^{\text{inn}})}{\sum_{\ell \in \mathcal{I}_{\text{corr}}} \text{tr}(\Sigma_\ell^{\text{inn}})} \tag{16}$$

The calibrated coverage condition ensures the second fraction is at least $\kappa\eta$, giving:

$$\frac{\mathfrak{F}_\Psi}{\mathfrak{F}_{\text{all}}} \geq \rho\kappa\eta \tag{17}$$

and hence $\mathfrak{F}_\Psi \geq (\rho\kappa\eta)\mathfrak{F}_{\text{all}}$.

This shows that, even after restricting to the η-coverage saliency set and applying potentially lossy K-dimensional projections, the GUIDE head retains at least a fraction $\rho\kappa\eta$ of the total classical Fisher information available from all correlated blocks. In other words, the selected layers preserve a guaranteed share of the information about the linearised parameter β.

□

A.2 PROPOSITION 2

Proposition 2. Fix a training pair $(\tilde{x}_t, y)$ at corruption stage t with K classes. GUIDE outputs Dirichlet parameters:

$$\alpha(\tilde{x}_t) = \exp\big(g_{\text{out}}(\Psi(\tilde{x}_t))\big) + 1, \qquad S = \sum_{k=1}^{K} \alpha_k(\tilde{x}_t), \tag{18}$$

and predictive mean $\hat{p} = \mathbb{E}[\pi \mid \alpha] = \alpha/S$. The curriculum constructs the uncertainty target $\tilde{s}_t$ and soft label $\tilde{y}_t$ as:

$$\tilde{s}_t = s_t\Big(1 - \frac{1}{2}c_t^2\Big), \qquad c_t = \langle \sigma(f_\theta(\tilde{x}_t)), y\rangle, \qquad \tilde{y}_t = \frac{\tilde{s}_t}{K} \cdot \mathbf{1}_K + (1 - \tilde{s}_t) \cdot y. \tag{19}$$

GUIDE minimises the loss (per sample):

$$\mathcal{L}(\alpha; \tilde{y}_t, y) = -\sum_{k=1}^{K} \tilde{y}_{t,k}\big(\psi(\alpha_k) - \psi(S)\big) + \lambda_{\text{kl}}\text{KL}\left(\text{Dir}(\alpha)||\text{Dir}(\beta)\right) + \frac{S}{K} \cdot \Big(1 - \langle y, \hat{p}\rangle\Big), \tag{20}$$

where $\psi(\cdot)$ is the digamma function and $\beta = \mathbf{1}_K$.

Then $\mathcal{L}(\alpha; \tilde{y}_t, y)$ exhibits the following endpoint behaviour:

1. **Uniform endpoint** ($\tilde{s}_t = 1$, so $\tilde{y}_t = \frac{1}{K}\mathbf{1}_K$): a global minimiser is:

$$\alpha^\star(\tilde{x}_t) = \beta = \mathbf{1}_K, \tag{21}$$

   i.e. the Dirichlet prior (minimal evidence under the constraint $\alpha_k \geq 1$).

2. **One-hot endpoint** ($\tilde{s}_t = 0$, so $\tilde{y}_t = y$): every minimiser satisfies:

$$\alpha_k^\star(\tilde{x}_t) = 1 \;\; (k \neq y), \qquad \alpha_y^\star(\tilde{x}_t) > 1, \tag{22}$$

   and as $\lambda_{\text{kl}} \to 0$ the optimiser drives $\alpha_y^\star(\tilde{x}_t)$ (hence $S^\star$) arbitrarily large, yielding high evidence concentrated on the true class.

Consequently, since GUIDE constructs $\tilde{y}_t$ as an interpolation between y and the uniform distribution via a corruption- and confidence-dependent mixing coefficient, these endpoints formally explain how GUIDE learns when to be uncertain: as $\tilde{y}_t$ moves toward uniform, the minimiser moves toward the prior and reduces evidence; as $\tilde{y}_t$ moves toward one-hot, the minimiser concentrates evidence on the true class.

Proof. Given a single training data point $(\tilde{x}_t, y)$, let $\alpha = \alpha(\tilde{x}_t)$, $S = \sum_k \alpha_k$, and $\hat{p} = \alpha/S$. By construction, $\alpha_k \geq 1$ for all k.

First, we consider the uniform endpoint. The ELBO term becomes:

$$\mathcal{L}_{\text{ELBO}} = -\frac{1}{K}\sum_{k=1}^{K} \psi(\alpha_k) + \psi(S) \tag{23}$$

Since the digamma function $\psi : (0, \infty) \to \mathbb{R}$ is concave, Jensen's inequality yields that for any $\alpha_1, \ldots, \alpha_K > 0$ with $\sum_{k=1}^{K} \alpha_k = S$:

$$\frac{1}{K}\sum_{k=1}^{K}\psi(\alpha_k) \le \psi\left(\frac{1}{K}\sum_{k=1}^{K}\alpha_k\right) = \psi(S/K) \tag{24}$$

with equality if and only if $\alpha_1 = \ldots = \alpha_K = S/K$. Consequently, among all α with fixed total evidence S, the quantity $\frac{1}{K}\sum_{k=1}^{K}\psi(\alpha_k)$ is maximised (and hence the corresponding ELBO term is minimised) by the symmetric choice $\alpha = (S/K)\mathbf{1}$.

Restricting to symmetric Dirichlet parameters of the form $\alpha = a\mathbf{1}_K$ (i.e., $\alpha_k = a$ for all k) with $a \ge 1$ enforced by the parameterisation, we have $S = Ka$ and $\hat{p} = \frac{1}{K}\mathbf{1}_K$. Under this restriction, the SRE penalty becomes:

$$\frac{S}{K}\Big(1 - \langle y, \hat{p}\rangle\Big) = \frac{Ka}{K}\Big(1 - \frac{1}{K}\Big) = a\Big(1 - \frac{1}{K}\Big) \tag{25}$$

which is strictly increasing in a. Since $a \ge 1$, this term is minimised at $a = 1$, i.e. $\alpha = \mathbf{1}_K$. The KL term is minimised at $\alpha = \beta$, and with $\beta = \mathbf{1}_K$ we conclude $\alpha^\star = \mathbf{1}_K$ is a global minimiser at the uniform endpoint.

Second, we consider the one-hot endpoint. Let y denote the correct class index, so $\tilde{y}_k = \mathbb{1}[k = y]$. The ELBO term reduces to:

$$\mathcal{L}_{\mathrm{ELBO}} = -(\psi(\alpha_y) - \psi(S)) \tag{26}$$

Fix $k \ne y$. Increasing α_k increases S; since ψ is increasing, this increases $\psi(S)$ and thus increases $\mathcal{L}_{\mathrm{ELBO}}$. Moreover, the SRE term simplifies to:

$$\frac{S}{K}\Big(1 - \langle y, \hat{p}\rangle\Big) = \frac{S}{K}\Big(1 - \frac{\alpha_y}{S}\Big) = \frac{S - \alpha_y}{K} = \frac{1}{K}\sum_{j \ne y}\alpha_j \tag{27}$$

which is strictly increasing in each α_k for $k \ne y$. The KL term is also minimised by staying close to $\beta = \mathbf{1}_K$. Hence, the loss is minimised by taking each off-class concentration as small as permitted, so $\alpha_k^\star = 1$ for all $k \ne y$.

With $\alpha_{k\ne y} = 1$, we have $S = \alpha_y + (K - 1)$ and the SRE term becomes $(K - 1)/K$, which does not depend on α_y. Thus α_y is controlled only by the ELBO term and the KL regulariser. The ELBO strictly prefers larger α_y: using that the trigamma $\psi'(\cdot)$ is positive and strictly decreasing:

$$\frac{d}{d\alpha_y}\Big[-(\psi(\alpha_y) - \psi(S))\Big] = -\psi'(\alpha_y) + \psi'(S) < 0 \tag{28}$$

since $S > \alpha_y$. Therefore, absent KL regularisation, the optimiser drives α_y upward. As $\lambda_{\mathrm{kl}} \to 0$, the KL restraint vanishes and $\alpha_y^\star$ (hence $S^\star$) can grow arbitrarily, yielding high evidence concentrated on the true class.

□

B ADDITIONAL EXPERIMENTS

This appendix provides additional experimental insights and analyses to complement the core results presented in Section 5. Specifically, we include extended analysis of metrics, adversarial attacks, thresholds, and ablations of hyperparameters introduced in GUIDE.

B.1 EXTENDED ANALYSIS OF CORE RESULTS

To empirically assess Theorem 1, we estimate the achieved cumulative relevance coverage (η) using Jacobian-based Fisher information on the MNIST → FashionMNIST. Figure 6 compares the achieved values to the desired targets. For target values of $\eta \in \{0.75, 0.80, 0.85\}$, the calibration procedure consistently selected {`pool1` and `conv1`} as the intermediate layers, yielding achieved η slightly above the desired level. At higher targets ($\eta = 0.90, 0.95$), the set expanded to include

pool2, which further increased coverage and again exceeded the bound predicted by Theorem 1. These results confirm that the saliency calibration reliably preserves the intended fraction of Fisher information and, if anything, errs on the side of retaining more information than required, thereby validating the practical soundness of our theoretical guarantee.

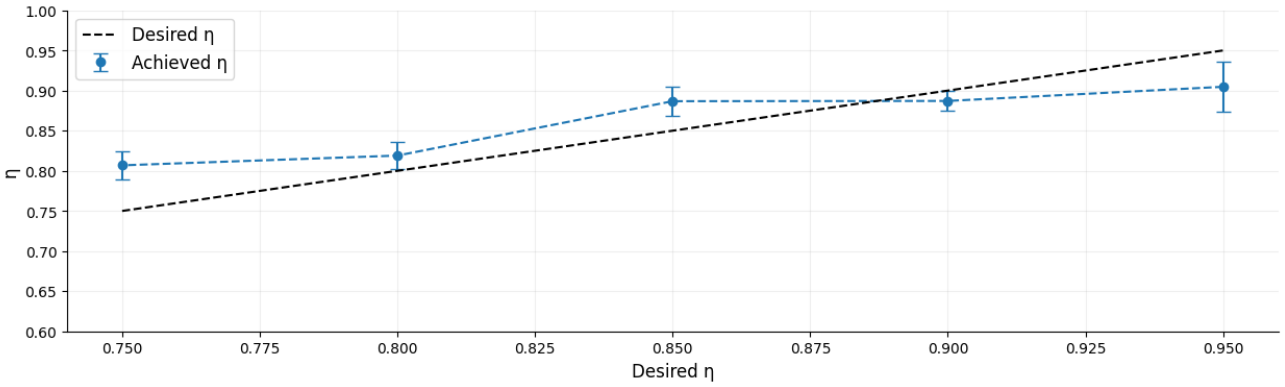
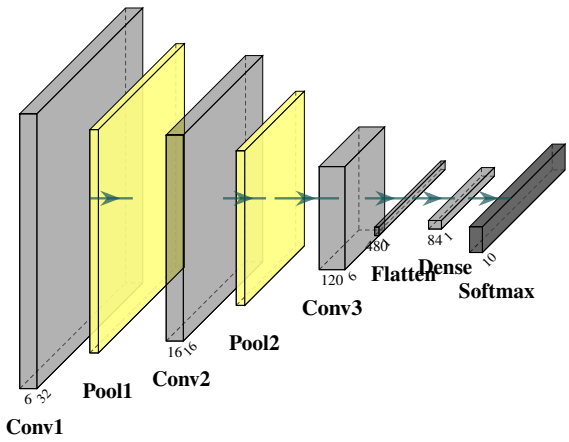
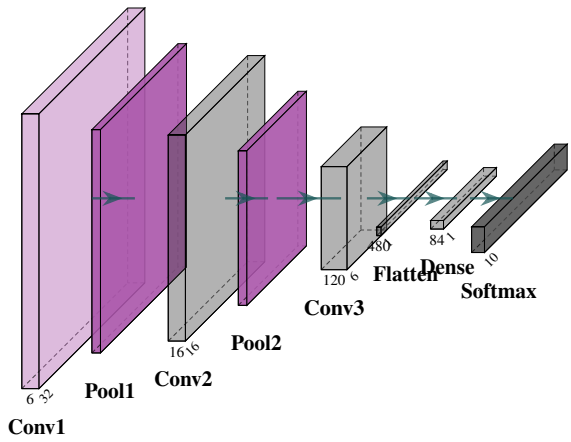


Figure 6: Visual approximation of Theorem 1: achieved cumulative relevance coverage (η) under saliency calibration versus the desired target. The empirical estimates, obtained via Jacobian-based Fisher information, closely track or exceed the theoretical bound.



(a) Evidential Meta-Model



(b) GUIDE meta-model

Figure 7: Qualitative comparison between the intermediate features selected in the LeNet model for (a) Evidential meta-model and (b) GUIDE meta-model. In the **Evidential meta-model**, intermediate layers are selected arbitrarily, shown in **yellow**, whereas in the **GUIDE meta-model**, the chosen features are calibrated to correspond to the most relevant and salient layers, highlighted in **purple**.

A clear distinction between EMM and GUIDE emerges when comparing the intermediate features each method relies on (Figure 7, Table 2). In the evidential meta-model, EMM depends on arbitrarily chosen layers (shown in **yellow**), which may not align with the most relevant or informative features. Moreover, selecting layers manually becomes increasingly impractical as models scale in depth and complexity, making an automated calibration procedure essential for general applicability. By contrast, GUIDE calibrates layer selection through cumulative relevance coverage, consistently identifying salient layers (highlighted in **purple**). This behaviour is evident across architectures: whereas EMM typically relies on handpicked layers, GUIDE selects a compact set of semantically meaningful layers (e.g., pool1, pool2, and conv1 in LeNet), ensuring the meta-model operates on features with maximal relevance. This principled calibration removes the arbitrariness of EMM's design and underpins GUIDE's improved stability and robustness across datasets.

Figure 8 illustrates representative curriculum samples generated by GUIDE, showing the relationship between the ground-truth label y, the base model confidence c, and the curriculum soft targets $\tilde{y}_k$. For clean or weakly corrupted digits, the base classifier is highly confident ($c \approx 1$), and the induced soft targets remain sharply peaked, closely approximating one-hot

Table 2: Comparison of layers selected by EMM and layers calibrated by GUIDE ($\eta = 0.9$, in order of relevancy) across different base networks. GUIDE consistently selects salient layers identified by relevance coverage, whereas EMM relies on arbitrary intermediate-layer choices.

Base Model	EMM Selected Layers	GUIDE Calibrated Layers
LeNet (MNIST)	{pool1, pool2}	{pool1, conv1, pool2}
ResNet-18 (CIFAR-10)	{block1_3_out, block2_3_out, block3_3_out}	{block1_3_out, block2_3_out, block3_3_out}
SENet (CIFAR-100, Flowers)	{pool1, pool2, gap }	{gap, pool2}

supervision. As corruption severity increases, visually ambiguous samples exhibit a controlled redistribution of probability mass across semantically similar classes (e.g., $8 \leftrightarrow 9$, $0 \leftrightarrow 9$, $1 \leftrightarrow 7$), while remaining anchored to the ground-truth label. Notably, even when the input is heavily degraded and confidence drops, the curriculum targets do not collapse to uniform noise, indicating that GUIDE preserves informative supervisory signal while reflecting increasing epistemic ambiguity. This behaviour supports the intended role of the curriculum as a smooth interpolation between hard labels and uncertainty-aware soft supervision.

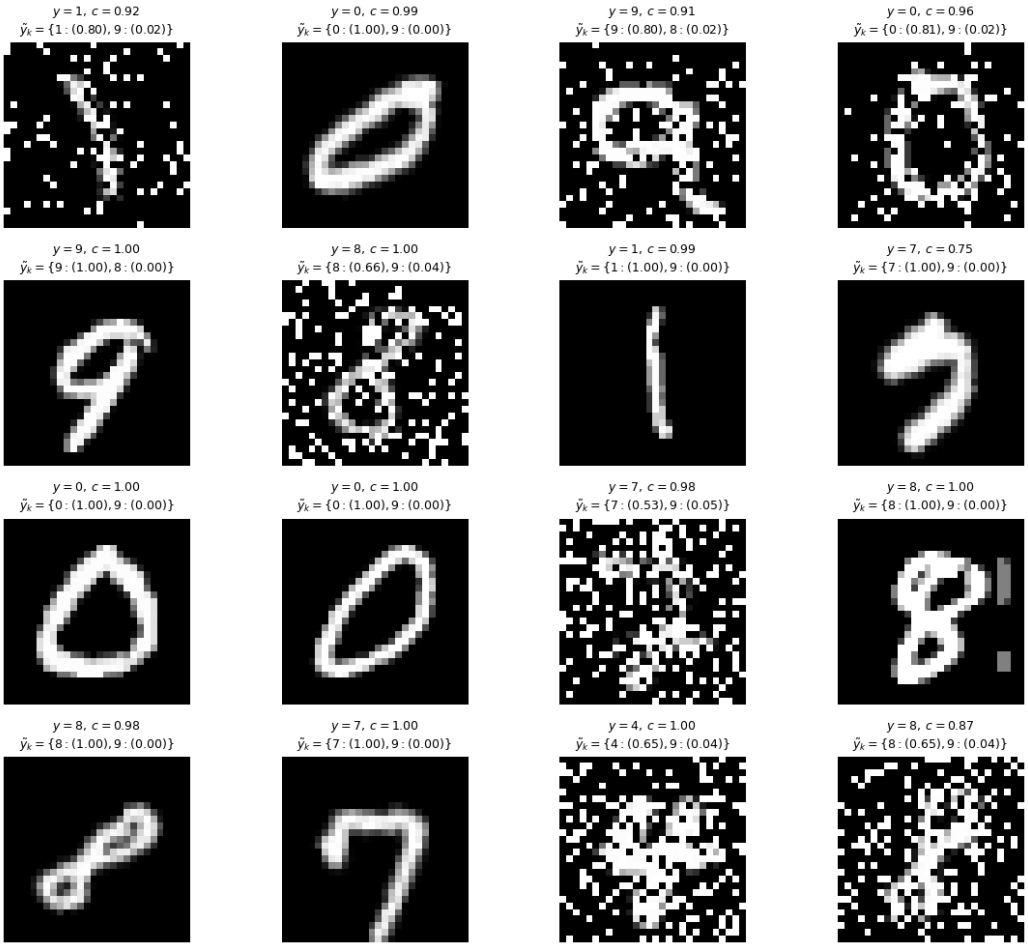


Figure 8: Qualitative examples of GUIDE curriculum on MNIST. Each panel shows a curriculum training sample with its ground-truth label y, base model confidence c, and the top-k entries of the corresponding soft target $\tilde{y}_k$.

Figure 11 provides calibration curves supporting the main-text discussion. The pretrained model and EDL-Head are overconfident on ID data (curves below the diagonal; smECE $= 0.317$ and 0.265) and remain overly confident on OOD inputs. EMM is underconfident on ID data (above-diagonal; smECE $= 0.193$) yet still assigns high confidence to OOD samples. GUIDE best matches the diagonal (smECE $= 0.061$) and maintains low confidence on OOD inputs, consistent with robust uncertainty calibration and OOD rejection.

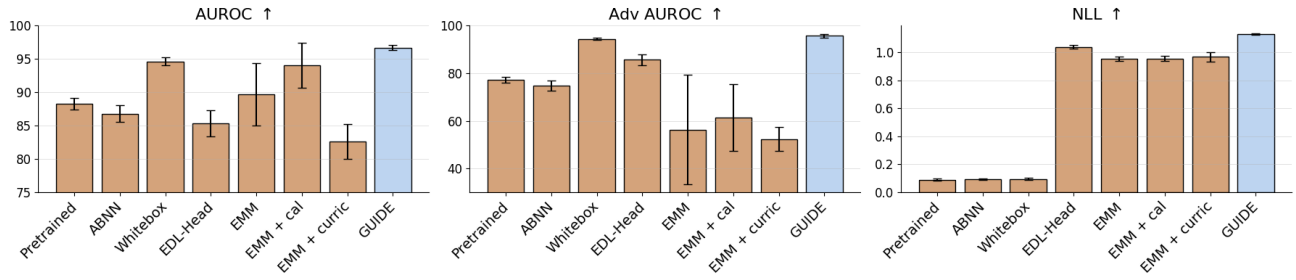
Table 3: The mean NLL, drop in confidence between ID → OOD data, and ID → Adv data with standard deviation of the comparative approaches with a variety of ID and OOD datasets. Highlighted cells denote the best performance for each metric. * indicates datasets classed as Near-OOD.

	Pretrained	**ABNN**	**EDL-Head**	**Whitebox**	**EMM**	**EMM(+cal)**	**EMM(+curric)**	**GUIDE**
Type	-	Intrusive	Intrusive	Post-hoc	Post-hoc	Post-hoc	Post-hoc	Post-hoc
MNIST → FashionMNIST								
ID NLL ↑	0.09 ± 0.01	0.09 ± 0.00	1.04 ± 0.01	0.10 ± 0.01	0.95 ± 0.02	0.94 ± 0.02	0.97 ± 0.03	1.13 ± 0.01
ΔConf. (ID-OOD) ↑	0.14 ± 0.05	0.13 ± 0.03	0.53 ± 0.05	0.14 ± 0.04	0.54 ± 0.20	0.45 ± 0.22	0.45 ± 0.10	0.82 ± 0.02
ΔConf. (ID-Adv) ↑	0.11 ± 0.02	0.10 ± 0.03	0.64 ± 0.03	0.12 ± 0.01	0.65 ± 0.30	0.11 ± 0.23	0.19 ± 0.10	0.87 ± 0.01
MNIST → KMNIST								
ID NLL ↑	0.09 ± 0.01	0.09 ± 0.01	1.04 ± 0.01	0.09 ± 0.01	0.95 ± 0.02	0.96 ± 0.02	0.97 ± 0.04	1.13 ± 0.01
ΔConf. (ID-OOD) ↑	0.17 ± 0.01	0.18 ± 0.01	0.62 ± 0.02	0.17 ± 0.00	0.69 ± 0.08	0.75 ± 0.07	0.61 ± 0.03	0.82 ± 0.02
ΔConf. (ID-Adv) ↑	0.08 ± 0.01	0.06 ± 0.01	0.63 ± 0.03	0.10 ± 0.01	0.17 ± 0.45	0.15 ± 0.32	0.15 ± 0.08	0.81 ± 0.02
MNIST → EMNIST*								
ID NLL ↑	0.09 ± 0.01	0.10 ± 0.01	1.04 ± 0.01	0.10 ± 0.01	0.95 ± 0.03	0.95 ± 0.01	0.97 ± 0.01	1.13 ± 0.00
ΔConf. (ID-OOD) ↑	0.16 ± 0.01	0.17 ± 0.02	0.59 ± 0.01	0.16 ± 0.01	0.49 ± 0.06	0.58 ± 0.06	0.47 ± 0.05	0.67 ± 0.01
ΔConf. (ID-Adv) ↑	0.10 ± 0.01	0.08 ± 0.01	0.63 ± 0.02	0.11 ± 0.01	0.20 ± 0.32	0.28 ± 0.20	0.19 ± 0.08	0.75 ± 0.02
CIFAR10 → SVHN								
ID NLL ↑	1.83 ± 0.20	1.63 ± 0.00	1.16 ± 0.02	1.80 ± 0.20	1.00 ± 0.22	0.65 ± 0.02	0.62 ± 0.02	1.17 ± 0.02
ΔConf. (ID-OOD) ↑	0.09 ± 0.04	0.10 ± 0.01	0.60 ± 0.08	0.12 ± 0.02	0.52 ± 0.15	0.33 ± 0.10	0.41 ± 0.10	0.72 ± 0.11
ΔConf. (ID-Adv) ↑	-0.02 ± 0.01	-0.03 ± 0.00	0.58 ± 0.09	-0.03 ± 0.01	0.37 ± 0.04	-0.01 ± 0.06	0.07 ± 0.11	0.48 ± 0.13
CIFAR10 → CIFAR100*								
ID NLL ↑	2.41 ± 0.23	1.69 ± 0.01	1.30 ± 0.02	1.12 ± 0.14	1.01 ± 0.19	0.92 ± 0.02	1.30 ± 0.56	1.34 ± 0.01
ΔConf. (ID-OOD) ↑	0.09 ± 0.01	0.10 ± 0.01	0.49 ± 0.01	0.23 ± 0.04	0.28 ± 0.05	0.36 ± 0.02	0.11 ± 0.42	0.55 ± 0.01
ΔConf. (ID-Adv) ↑	0.03 ± 0.01	-0.00 ± 0.01	0.58 ± 0.02	0.34 ± 0.04	0.24 ± 0.09	0.33 ± 0.03	0.26 ± 0.12	0.64 ± 0.01
Oxford Flowers (low shot) → Deep Weeds								
ID NLL ↑	1.31 ± 0.19	3.75 ± 0.01	3.00 ± 0.01	1.40 ± 0.43	3.87 ± 0.25	3.55 ± 0.26	3.91 ± 0.01	3.29 ± 0.01
ΔConf. (ID-OOD) ↑	0.03 ± 0.02	0.13 ± 0.08	0.33 ± 0.34	0.03 ± 0.08	0.10 ± 0.05	0.13 ± 0.06	0.18 ± 0.08	0.10 ± 0.22
ΔConf. (ID-Adv) ↑	0.04 ± 0.02	0.14 ± 0.06	0.43 ± 0.29	0.07 ± 0.09	0.12 ± 0.04	0.17 ± 0.05	0.22 ± 0.08	0.19 ± 0.25

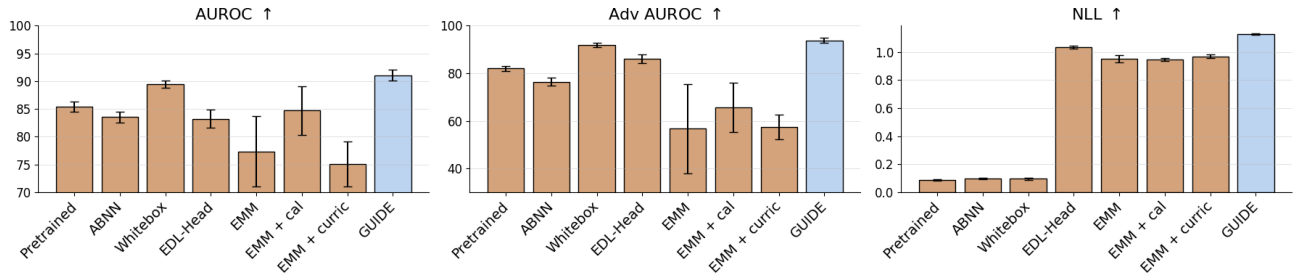
The extended results for AUROC, adversarial AUROC, and NLL across multiple dataset pairs are reported in Figure 9 and Table 3. Consistent with the main results, intrusive baselines such as ABNN and EDL-Head achieve modest gains in AUROC but generally fail to reduce overconfidence across all settings. Post-hoc baselines like Whitebox and EMM often attain higher NLL and competitive AUROC, yet remain unstable: EMM in particular shows large variance across runs, with AUROC values sometimes below 80% and adversarial AUROC highly inconsistent. In contrast, GUIDE achieves the highest AUROC on both near-OOD (e.g., MNIST → KMNIST, CIFAR-10 → CIFAR-100) and far-OOD settings (e.g., CIFAR-10 → SVHN, Oxford Flowers → DeepWeeds), while consistently producing the high NLL, indicating well-calibrated confidence. Importantly, GUIDE also delivers the largest drop in confidence between ID and both OOD and adversarial inputs across all dataset pairs, confirming its ability to reject harmful inputs while preserving in-distribution predictions. These trends hold across both simple (LeNet, MNIST) and deeper architectures (ResNet-18, DenseNet), demonstrating that GUIDE scales effectively and remains robust under diverse distribution shifts.

The confidence drop analysis in Figure 12 provides evidence of GUIDE's effectiveness across all evaluated dataset pairs, as breifly discussed within the main text. For MNIST → FMNIST, MNIST → KMNIST, and MNIST → EMNIST (Figures 12b and 12c), the pretrained model, ABNN, and Whitebox exhibit almost no drop in confidence ($\leq 15\%$), confirming their tendency to remain overconfident under shift. EDL-Head and EMM achieve larger drops ($40 \rightarrow 60\%$), but in some cases adversarial drops remain significantly smaller than OOD drops, indicating incomplete robustness. GUIDE clearly outperforms all baselines across the majority of dataset pairs, producing stable confidence drops above 80% in some cases for both OOD and adversarial inputs, whereas competing methods rarely exceed 60%. Taken together, these results confirm that GUIDE systematically learns to maintain high confidence on ID samples while sharply reducing confidence on both OOD and adversarial data, a behaviour that neither intrusive nor post-hoc baselines reliably achieve.

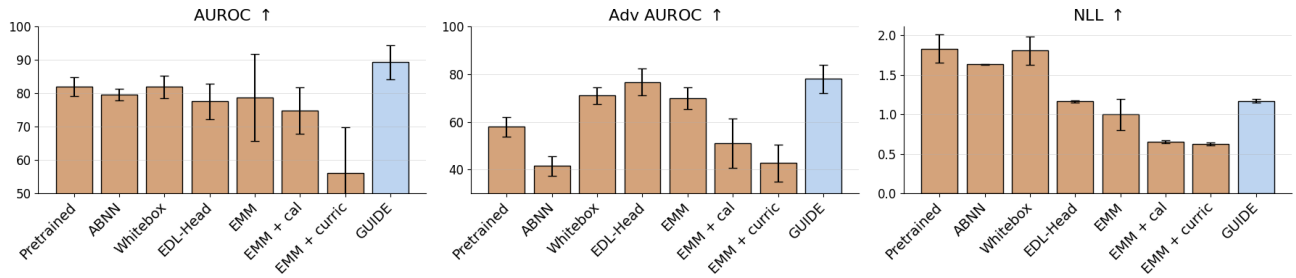
Figure 13 illustrates how the GUIDE meta-model transforms the base model's confidence c_t and the estimated noise level s_t into adjusted training targets $\tilde{y}_t$ for both the correct and incorrect classes. Across the entire domain, $\tilde{y}_t \in [0, 1]$, ensuring that the constructed target vector is a valid probability distribution compatible with the cross-entropy objective. For the correct class ($k = y$), $\tilde{y}_t$ increases smoothly with confidence while being attenuated by the noise level. In low-noise, high-confidence regions, the target approaches 1, reinforcing learning of the correct label. Conversely, when confidence is low or noise is high, the assigned probability decreases, allowing GUIDE to down-weight uncertain or corrupted inputs. Importantly, in the



(a) MNIST (ID) → KMNIST (OOD).



(b) MNIST (ID) → EMNIST (OOD).



(c) CIFAR-10 (ID) → SVHN (OOD).

Figure 9: AUROC, Adversarial AUROC (Adv AUROC), and Negative Log-Likelihood (NLL) for each evaluated ID→OOD dataset pair.

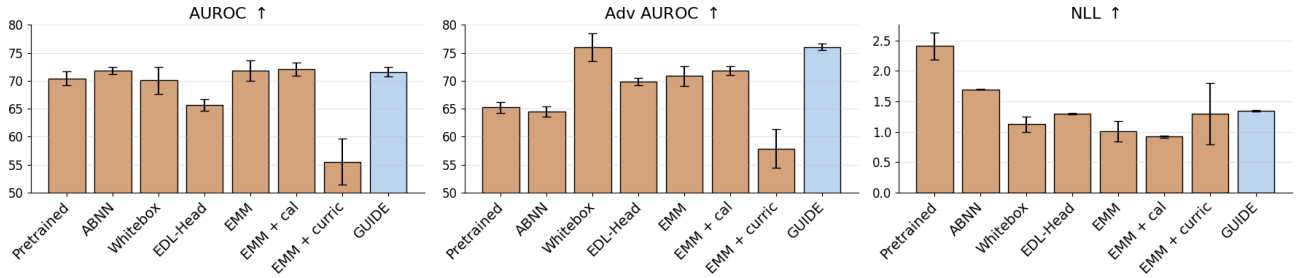
downstream evidential Dirichlet formulation, such cases still correspond to low total evidence, preserving uncertainty even when mean confidence is high.

For incorrect classes ($k \neq y$), the surface remains flat and close to zero, preventing spurious reinforcement of misclassified outputs. In extreme high-noise, high-confidence settings, the assignment approaches uniform mass $1/K$ due to the GUIDE meta-model's uniform-mixing term. While this may appear as an undesirable "bump," it is in fact deliberate: it enforces maximum uncertainty under severe corruption, maintaining probabilistic validity and discouraging overconfident misclassifications. Although this corner case rarely arises in practice, it is crucial for ensuring GUIDE's theoretical guarantees.

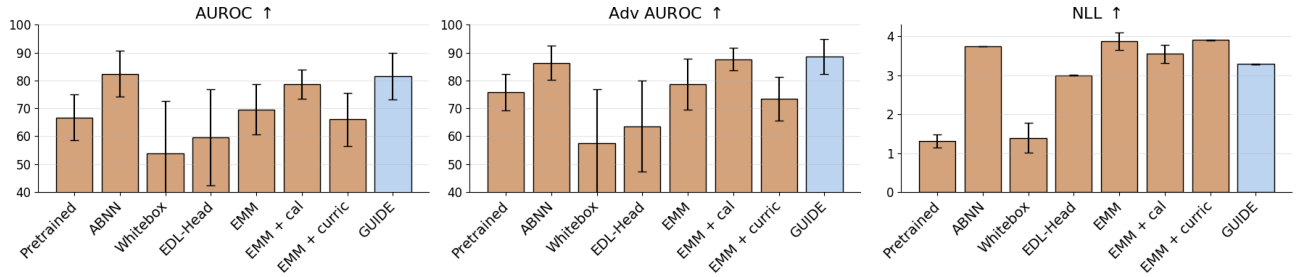
Taken together, these surfaces demonstrate that GUIDE adaptively balances reinforcement of reliable predictions with suppression of noisy or misleading signals, yielding soft targets that are both probabilistically sound and robust to distributional corruption.

B.2 ADVERSARIAL ATTACK ANALYSIS

Across all attack types and perturbation strengths evaluated in Table 4, GUIDE consistently achieves the highest adversarial AUROC, with margins that are both statistically significant and robust across the standard deviation. Under the L2PGD attack, GUIDE outperforms all baselines at every perturbation level, reaching 96.07% at $\epsilon = 0.1$ and retaining a remarkably high 94.67% even at $\epsilon = 1.0$. By contrast, competing methods such as EMM and its calibrated or curriculum variants exhibit



(a) CIFAR-10 (ID) → CIFAR-100 (OOD).



(b) Oxford Flowers (ID) → Deep Weeds (OOD).

Figure 10: AUROC, Adversarial AUROC (Adv AUROC), and Negative Log-Likelihood (NLL) for each evaluated ID→OOD dataset pair.

substantial degradation in performance, often falling below 60% AUROC under stronger perturbations.

A similar trend emerges under FGSM perturbations. While Whitebox attains strong performance at moderate to high perturbations ($\epsilon = 0.5, 1.0$), GUIDE maintains consistently competitive results across the entire range, outperforming all alternatives at lower perturbation strengths where robustness is most challenging. Importantly, GUIDE's performance remains within small standard deviation, highlighting both its stability and reliability in adversarial settings.

For the non-gradient-based Salt & Pepper noise attack, GUIDE again dominates across all perturbation magnitudes, exceeding 96% AUROC throughout. Competing baselines such as ABNN and Whitebox achieve high values in isolated cases, but their performance is less stable and fails to consistently match GUIDE's superior robustness. The fact that GUIDE maintains high AUROC scores across gradient-based and non-gradient-based attacks alike highlights its versatility and generalizability.

The AutoAttack results provide a stricter robustness check because they combine multiple complementary attacks within a standardised evaluation [Croce and Hein, 2020]. GUIDE remains strongest under both AutoAttack-L_2 and AutoAttack-L_∞, maintaining high AUROC as ϵ increases. In comparison, the base EMM baseline and the intermediate ablations (EMM+cal and EMM+curric) degrade more sharply under the stronger budgets, supporting the conclusion that robust detection relies on the full GUIDE pipeline rather than any single component in isolation. Taken together, these results show GUIDE is robust across both single-method attacks and the stronger AutoAttack suites among the evaluated approaches.

B.3 THRESHOLD ANALYSIS

Table 5 reports the performance of comparative approaches across a variety of uncertainty metrics used for the scoring rule. The results demonstrate that GUIDE achieves state-of-the-art performance consistently across all metrics, while alternative methods show clear weaknesses in either calibration, coverage, or robustness.

Under the Mutual Information metric, all methods achieve high in-distribution accuracy (above 99%), with Whitebox reaching the highest overall accuracy at 99.92%. However, accuracy differences at this scale are marginal and do not reflect robustness under distributional shift. For example, while EMM(+cal) attains the best in-distribution coverage (92.32%), it fails dramatically on adversarial coverage, yielding 55.47% compared to just 4.60% for GUIDE. Similarly, Whitebox reduces adversarial coverage to 9.04%, but GUIDE achieves the best overall balance by combining strong ID coverage (87.05%) with the lowest OOD (7.34%) and adversarial coverage (4.60%). These values translate into AUROC scores of

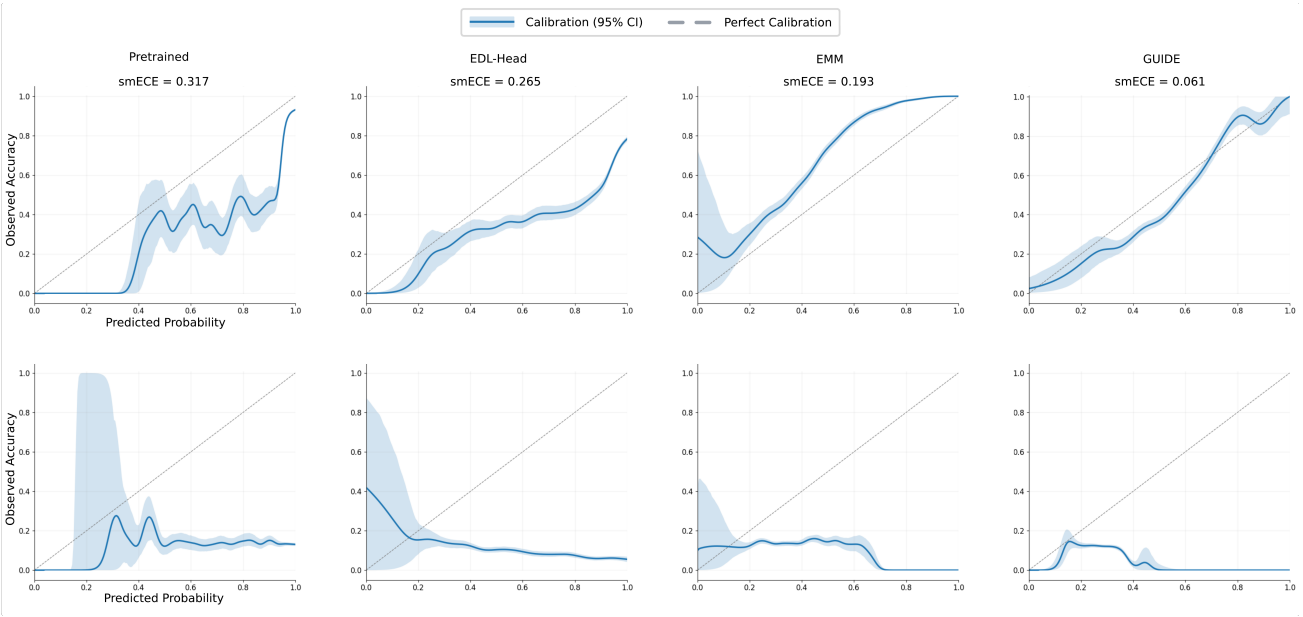


Figure 11: Calibration plots and smoothed expected calibration error (smECE) for ID (MNIST, top) and OOD (FashionMNIST, bottom) for comparative methods. The pretrained model is overconfident, EMM is underconfident, while GUIDE shows good calibration.

94.85% for OOD detection and 95.72% for adversarial detection, both substantially ahead of the next-best method.

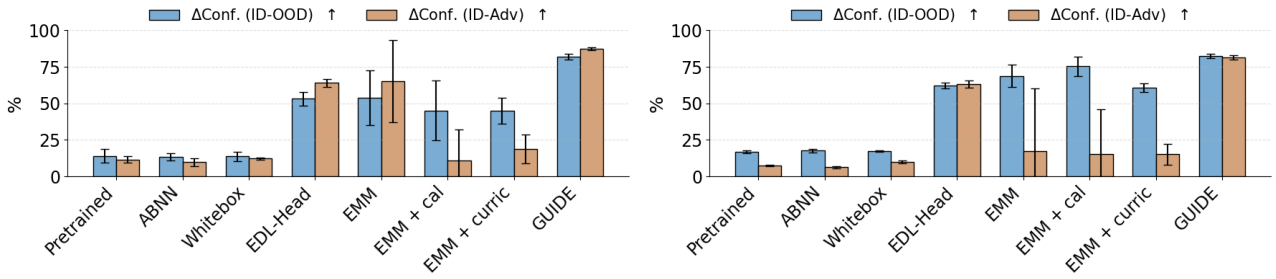
With the Maximum Probability metric, GUIDE again delivers the strongest results across nearly all evaluation criteria. It matches Whitebox in in-distribution accuracy at 99.92% but provides markedly improved robustness. For instance, OOD coverage is reduced to 6.77% and adversarial coverage to only 3.34%, whereas the best competing approaches (ABNN and Whitebox) still remain above 15%. This robustness translates into AUROC scores of 96.59% for OOD detection and 97.94% for adversarial detection, again setting the clear benchmark.

GUIDE's robust performance is most evident under Differential Entropy, where it achieves near-perfect separation of ID and shifted samples. Specifically, GUIDE attains an ID coverage of 91.55%, OOD coverage as low as 7.83%, and adversarial coverage of just 3.30%. The corresponding AUROC scores are 96.89% for OOD detection and 98.08% for adversarial detection, surpassing all baselines by wide margins. In contrast, alternative methods such as EDL-Head or EMM variants struggle, with adversarial AUROC values often below 75%. Taken together, these results show that although some baselines excel in isolated metrics (e.g., Whitebox in raw ID accuracy or EMM(+cal) in ID coverage), only GUIDE achieves consistently strong and balanced performance across ID, OOD, and adversarial evaluations. These results can also be visualised in Figure 14 for further comparative analysis.

B.4 ABLATION ANALYSIS

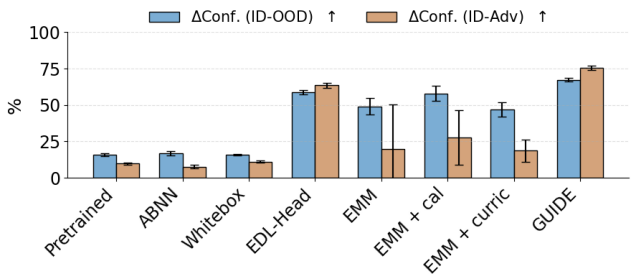
GUIDE's main hyperparameters have direct operational roles. The curriculum length T controls the number of corruption stages, γ controls how quickly corruption increases across those stages, η determines the amount of saliency mass retained when selecting layers, and λ_{KL} controls the strength of Dirichlet regularisation. The following ablations examine the sensitivity of GUIDE to these choices and provide practical guidance for selecting stable default values.

Table 6 presents the effect of varying the hyperparameter T on GUIDE's performance. Across all values, in-distribution accuracy remains very high (above 99.7%), confirming that predictive performance is stable with respect to T. The most notable differences arise in coverage and adversarial robustness. At $T = 2$, adversarial coverage is relatively higher (9.42%), while at $T = 5$ it is minimised to 4.60%, with a corresponding adversarial AUROC of 95.72%, the strongest result observed. Larger values of T (10 and 20) slightly improve AUROC on OOD detection (up to 95.24%) but come with modest increases in adversarial coverage (around 7-9%). These results suggest that $T = 5$ provides a strong default, offering the best adversarial rejection without increasing curriculum resolution unnecessarily. Larger T values may be useful when finer-grained corruption stages are required, but they should be used only when the added training cost is justified.

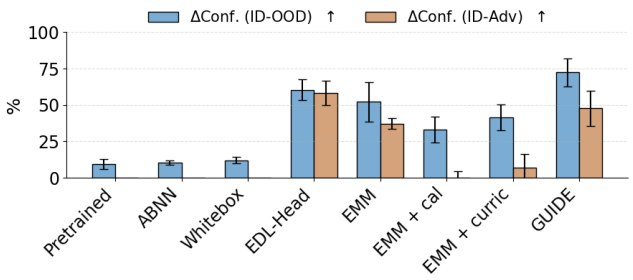


(a) MNIST (ID) → FMNIST (OOD).

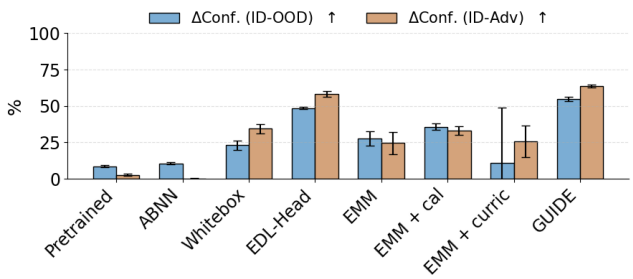
(b) MNIST (ID) → KMNIST (OOD).



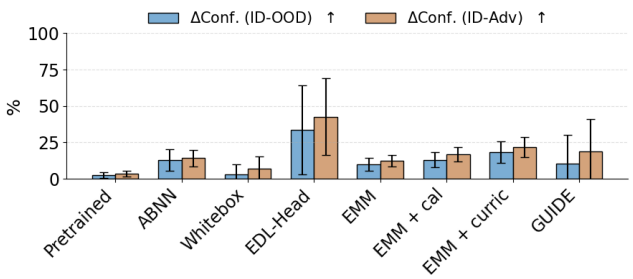
(c) MNIST (ID) → EMNIST (OOD).



(d) CIFAR-10 (ID) → SVHN (OOD).



(e) CIFAR-10 (ID) → CIFAR-100 (OOD).



(f) Oxford Flowers (ID) → Deep Weeds (OOD).

Figure 12: Mean drop in confidence between ID → OOD data and ID → Adv data for all evaluated upon dataset pairs.

Table 7 examines the sensitivity of GUIDE to the weighting parameter γ. In-distribution accuracy is uniformly high across settings ($\approx 99.8\%$), indicating that predictive accuracy is largely insensitive to γ. The primary effects manifest in coverage and robustness: a moderate value, $\gamma = 0.25$, offers the most balanced performance, attaining the lowest adversarial coverage (4.60%) with strong OOD coverage (7.34%) and the highest AUROC/Adv-AUROC pair (94.85%/95.72%). Lowering γ to 0.10 slightly increases both OOD and adversarial coverage (9.87% and 8.73%) and reduces OOD AUROC to 93.78%. Increasing γ beyond 0.25 raises ID coverage (e.g., 91.69% at $\gamma = 1.00$) but at the cost of markedly higher spurious coverage on OOD and adversarial inputs (12.14% and 13.37%, respectively) and weaker robustness (Adv-AUROC 93.59%). The degradation is most apparent at $\gamma = 0.75$, where OOD coverage peaks at 13.34% and OOD AUROC dips to 91.52%. Overall, the results suggest that creating a monotonic slope that is too sharp could degrade performance, but very slightly. These results support $\gamma = 0.25$ as a practical default, as it provides the best balance between ID coverage, OOD rejection, and adversarial robustness. Smaller γ values may be preferable when a smoother corruption schedule is desired, while larger values should be used cautiously because they reach high corruption levels more quickly and can mildly degrade robustness.

Finally, Table 8 analyses the impact of varying the η hyperparameter on GUIDE. Across all values, in-distribution accuracy remains consistently high (≈99.8–99.9%), indicating that predictive performance is largely unaffected by η. The main differences emerge in coverage and computational cost. Moderate settings ($\eta = 0.9$) achieve the strongest balance, with adversarial coverage reduced to 4.60% and corresponding adversarial AUROC of 95.72%, while also maintaining competitive OOD coverage (7.34%). Lower η values (e.g., 0.5-0.7) yield similar AUROC (around 95%) but involve fewer selected layers and slightly lower inference times ($\sim$1.65s). In contrast, higher η values (1.0) expand the set of selected layers ({pool1, conv1, pool2, conv2}), which increases inference time to 2.38s and slightly worsens adversarial coverage (7.71%). This suggests that very high η values can introduce weakly informative layers that increase the meta-model's sensitivity and

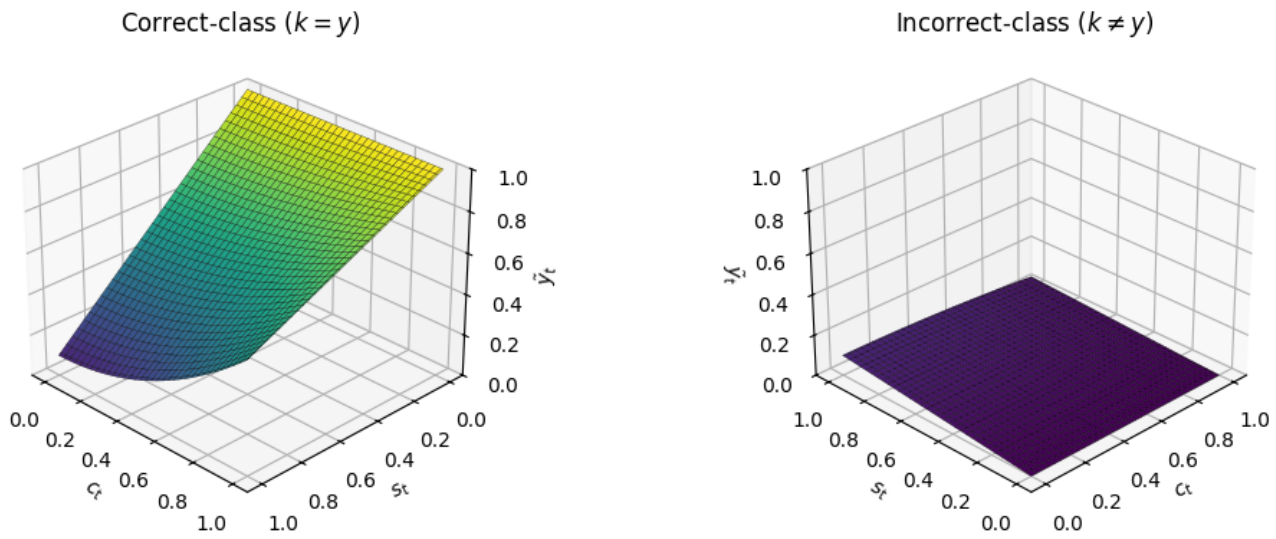


Figure 13: Visualization of the soft-target assignment $\tilde{y}_k$ under our weighting scheme as a function of model confidence c_t and noise s_t.

computational cost without providing additional robustness. Overall, η acts primarily as a coverage-efficiency parameter: moderate values retain sufficient decision-relevant information, whereas overly small or large values may respectively omit useful features or add redundant ones.

B.5 CALIBRATION COMPARISONS

GUIDE uses LRP-ϵ because it redistributes prediction relevance backwards through the frozen network in proportion to each neuron's contribution, while the ϵ stabilisation term reduces numerical instability when the redistribution denominator is small. This makes LRP well suited to GUIDE: a single relevance pass provides both layer-wise relevance scores for selecting salient intermediate layers and input-level relevance maps for constructing the saliency-weighted corruption masks used in the noise-driven curriculum.

We next examine whether GUIDE's saliency calibration mechanism can be instantiated using alternative attribution methods in place of Layer-wise Relevance Propagation (LRP). To this end, we repeat the relevance experiment using Grad-CAM Selvaraju et al. [2017] and Integrated Gradients (IG) Qi et al. [2020] to rank layers prior to coverage-controlled selection. The resulting calibration curves are shown in Figure 16, which reports the achieved relevance coverage as a function of the desired target η.

This experiment does not assess the interpretability or visual quality of the resulting explanations, as all three methods produce effective heatmaps. Rather, it probes a structural question central to GUIDE: whether attribution scores produced by a given method can be meaningfully treated as a relevance signal, such that selecting layers to cover a fraction η of relevance results in the retention of approximately η of the model's decision signal.

Across all target values, GUIDE closely tracks the desired coverage, exhibiting near-identity behaviour with low variance. In contrast, both Grad-CAM and IG display pronounced and systematic miscalibration. Grad-CAM consistently underestimates achieved coverage, with the gap widening as η increases, while IG degrades more severely, collapsing at higher target values and contributing little additional relevance beyond a small subset of layers. Both methods also exhibit increasing variability at larger η, indicating unstable layer rankings.

These failures stem from the fact that LRP is the only attribution method that satisfies a global relevance conservation principle, which makes relevance values comparable across layers and renders cumulative coverage meaningful. Consequently, as η increases, GUIDE (LRP) expands the selected set in a controlled manner by adding genuinely decision-relevant layers, rather than injecting spurious layers that contribute little to the underlying signal. In contrast, Grad-CAM and IG do not yield a conserved, additive notion of relevance across depth, so meeting higher target η forces the inclusion of many low-utility layers as an artefact of the ranking procedure. This manifests as degraded achieved coverage at large η and substantially

Table 4: The mean adversarial AUROC with standard deviation of the comparative approaches for a variety of adversarial attacks and maximum perturbations ϵ where the ID dataset is MNIST and the OOD dataset is FashionMNIST. Highlighted cells denote the best performance for each metric.

ϵ	**Pretrained**	**ABNN**	**EDL-Head**	**Whitebox**	**EMM**	**EMM(+cal)**	**EMM(+curric)**	**GUIDE**
Type	-	Intrusive	Intrusive	Post-hoc	Post-hoc	Post-hoc	Post-hoc	Post-hoc
					L2PGD			
0.1	91.50 ± 1.40	89.54 ± 1.70	61.06 ± 13.52	90.70 ± 4.22	88.66 ± 4.31	80.70 ± 12.33	51.09 ± 13.19	96.07 ± 1.39
0.5	89.68 ± 1.13	87.18 ± 2.62	60.15 ± 14.11	90.93 ± 3.70	64.02 ± 8.21	76.04 ± 12.81	51.44 ± 8.61	96.72 ± 1.08
1.0	83.67 ± 3.37	81.81 ± 3.95	88.73 ± 2.64	90.26 ± 0.59	83.43 ± 17.95	53.56 ± 13.23	53.62 ± 9.93	95.72 ± 0.91
					FGSM			
0.1	87.27 ± 2.77	86.62 ± 2.81	61.04 ± 8.02	92.66 ± 2.31	79.45 ± 14.09	76.11 ± 23.54	47.81 ± 7.19	95.67 ± 1.03
0.5	88.40 ± 1.39	88.52 ± 2.51	46.48 ± 9.00	95.56 ± 2.23	91.72 ± 5.11	72.20 ± 38.06	49.13 ± 4.50	94.35 ± 1.55
1.0	88.23 ± 0.91	89.06 ± 2.14	34.72 ± 9.19	96.92 ± 1.29	95.36 ± 1.69	72.33 ± 38.78	48.61 ± 5.97	94.19 ± 2.12
					Salt & Pepper			
0.1	83.60 ± 3.53	86.27 ± 4.44	61.27 ± 4.53	90.42 ± 3.61	65.32 ± 16.42	76.00 ± 11.77	63.62 ± 8.07	96.05 ± 1.49
0.5	86.88 ± 3.36	89.16 ± 4.07	61.30 ± 4.52	91.66 ± 3.28	$66.45 \pm 15.75\%$	78.27 ± 12.06	64.33 ± 7.45	96.34 ± 1.35
1.0	89.63 ± 3.07	91.00 ± 3.87	61.47 ± 4.43	92.75 ± 2.84	$67.15 \pm 15.13\%$	80.46 ± 12.52	65.09 ± 6.92	96.60 ± 1.20
					AutoAttack L_2			
0.1	83.22 ± 4.37	87.18 ± 3.41	82.36 ± 6.42	90.02 ± 3.53	72.66 ± 12.98	59.79 ± 14.49	55.99 ± 8.86	94.76 ± 1.31
0.5	72.62 ± 3.63	87.22 ± 2.52	80.13 ± 3.61	89.77 ± 1.42	44.02 ± 14.40	57.40 ± 14.39	39.74 ± 8.68	95.70 ± 1.11
1.0	64.35 ± 3.60	88.50 ± 1.84	74.67 ± 3.06	88.94 ± 1.72	23.78 ± 17.83	29.84 ± 12.76	14.63 ± 7.43	91.31 ± 2.40
					AutoAttack L_∞			
$\frac{2}{255}$	82.54 ± 3.07	88.66 ± 1.75	82.08 ± 7.01	89.97 ± 3.19	70.88 ± 12.55	79.00 ± 12.76	54.41 ± 7.53	94.37 ± 1.78
$\frac{4}{255}$	78.65 ± 3.27	89.06 ± 1.69	81.95 ± 6.51	89.75 ± 2.93	61.74 ± 9.66	76.13 ± 8.99	52.87 ± 7.99	94.71 ± 1.63
$\frac{8}{255}$	71.74 ± 4.44	89.60 ± 1.64	80.52 ± 5.78	89.19 ± 2.80	36.69 ± 11.70	60.52 ± 10.69	36.73 ± 8.20	94.97 ± 1.32

higher variance, indicating that the resulting layer ordering is unstable and can change markedly across random seeds.

B.6 CORRUPTION BENCHMARKS

The corrupted benchmark in Table 9 presents a more challenging setting than the standard ID→OOD pairs, since the OOD data is based off corruptions of the original ID dataset rather than drawing from a completely different distribution. As a result, many corrupted samples still preserve the global structure, semantics, and low-level statistics of the ID data, especially at lower corruption strengths. This creates substantial overlap between the ID and OOD score distributions, making separation inherently difficult.

Despite this difficulty, GUIDE consistently achieves one of or the best AUROC and adversarial AUROC across both corrupted settings. This suggests that the saliency-driven curriculum encourages the meta-model to recognise gradual degradation in salient features, rather than relying solely on coarse distributional differences. In turn, this enables GUIDE to maintain stronger separation between clean and corrupted inputs, even when the OOD data remains visually and statistically close to the ID distribution.

B.7 GRAPHICAL MODEL

Figure 17 summarises GUIDE as a Dirichlet meta-model with a curriculum-driven training signal. A curriculum index t determines a corruption level s_t, which is used to generate a corrupted view $\tilde{x}_t$. The meta-model maps $\tilde{x}_t$ to evidential parameters $\alpha(\tilde{x}_t)$, inducing a Dirichlet distribution over class probabilities, $\pi \sim \mathrm{Dir}(\alpha(\tilde{x}_t))$, and a categorical label variable $y \sim \mathrm{Cat}(\pi)$. In parallel, the frozen base model produces softmax outputs from which we compute the confidence score $c_t = \langle \sigma(f_\theta(\tilde{x}_t)), y \rangle$. The curriculum soft target $\tilde{y}_t$ is a deterministic function of (s_t, c_t); the dashed edge $\tilde{y}_t \dashrightarrow \alpha(\tilde{x}_t)$ indicates supervision through the training objective rather than a generative dependency.

Table 5: The mean accuracy, OOD detection, and adversarial attack detection performance with standard deviation of the comparative approaches with a variety of uncertainty metrics. The adversarial attack is an L2PGD attack. For compactness, we denote adversarial AUROC as AAUROC. Highlighted cells denote the best performance for each metric.

	Pretrained	**ABNN**	**EDL-Head**	**Whitebox**	**EMM**	**EMM(+cal)**	**EMM(+curric)**	**GUIDE**
Type	-	Intrusive	Intrusive	Post-hoc	Post-hoc	Post-hoc	Post-hoc	Post-hoc
Mutual Information								
ID Acc (%) ↑	99.80 ± 0.07	99.78 ± 0.06	99.43 ± 0.17	99.92 ± 0.08	99.55 ± 0.08	99.69 ± 0.09	99.21 ± 0.25	99.87 ± 0.04
ID Cov (%) ↑	79.26 ± 4.36	74.88 ± 7.98	77.44 ± 3.24	76.38 ± 5.96	90.58 ± 4.02	92.32 ± 1.57	79.46 ± 10.25	87.05 ± 1.63
OOD Cov (%) ↓	24.89 ± 9.71	18.23 ± 8.60	22.51 ± 5.50	13.50 ± 2.72	31.87 ± 26.23	35.69 ± 17.35	40.18 ± 14.28	7.34 ± 3.42
Adv Cov (%) ↓	25.84 ± 7.84	24.35 ± 9.67	13.86 ± 2.87	9.04 ± 4.21	22.18 ± 22.49	55.47 ± 13.92	62.11 ± 15.18	4.60 ± 2.69
AUROC (%) ↑	84.18 ± 5.93	85.93 ± 2.11	83.23 ± 3.88	88.38 ± 1.95	77.68 ± 20.60	74.54 ± 14.51	72.85 ± 8.03	94.85 ± 1.27
AAUROC (%) ↑	83.67 ± 3.37	81.81 ± 3.95	88.73 ± 2.64	90.26 ± 0.59	83.43 ± 17.95	53.56 ± 13.23	53.62 ± 9.93	95.72 ± 0.91
Maximum Probability								
ID Acc (%) ↑	99.76 ± 0.06	99.81 ± 0.06	99.09 ± 1.18	99.92 ± 0.03	99.80 ± 0.11	99.70 ± 0.14	99.62 ± 0.09	99.92 ± 0.05
ID Cov (%) ↑	81.59 ± 2.49	77.70 ± 8.83	80.47 ± 4.82	78.55 ± 3.96	87.21 ± 4.44	87.82 ± 4.33	81.34 ± 5.10	90.47 ± 1.10
OOD Cov (%) ↓	20.78 ± 6.98	16.77 ± 5.23	23.01 ± 14.86	20.67 ± 7.07	36.60 ± 20.78	29.76 ± 13.94	38.39 ± 16.91	6.77 ± 2.13
Adv Cov (%) ↓	26.39 ± 7.99	21.49 ± 3.67	14.46 ± 16.46	15.23 ± 7.47	67.96 ± 10.13	49.46 ± 10.22	52.31 ± 11.96	3.34 ± 1.70
AUROC (%) ↑	87.07 ± 3.74	87.59 ± 1.95	85.17 ± 6.64	85.18 ± 3.69	75.54 ± 16.45	82.09 ± 9.33	75.28 ± 11.23	96.59 ± 1.04
AAUROC (%) ↑	84.49 ± 3.48	83.66 ± 3.34	89.42 ± 7.37	87.99 ± 2.99	50.24 ± 9.33	61.29 ± 10.24	63.06 ± 9.96	97.94 ± 0.78
Differential Entropy								
ID Acc (%) ↑	-	-	96.86 ± 1.64	-	99.69 ± 0.11	99.67 ± 0.11	99.39 ± 0.11	99.82 ± 0.07
ID Cov (%) ↑	-	-	73.62 ± 3.57	-	86.82 ± 5.48	88.13 ± 2.98	79.31 ± 4.53	91.55 ± 1.28
OOD Cov (%) ↓	-	-	34.44 ± 9.01	-	34.86 ± 20.02	39.13 ± 26.98	39.91 ± 6.52	7.83 ± 1.98
Adv Cov (%) ↓	-	-	31.24 ± 7.94	-	60.45 ± 21.57	65.68 ± 7.74	58.22 ± 15.19	3.30 ± 1.57
AUROC (%) ↑	-	-	73.40 ± 6.07	-	77.03 ± 13.23	73.09 ± 22.74	74.25 ± 1.90	96.89 ± 0.77
AAUROC (%) ↑	-	-	74.82 ± 6.66	-	52.98 ± 18.64	52.85 ± 6.76	59.30 ± 13.57	98.08 ± 0.29

Table 6: The accuracy, OOD detection, and adversarial attack detection, AUROC, and adversarial AUROC performance of GUIDE with standard deviation under different values of the T hyperparameter. The adversarial attack is L2PGD (1.0 maximum perturbation), and the ID and OOD datasets are MNIST and FashionMNIST.

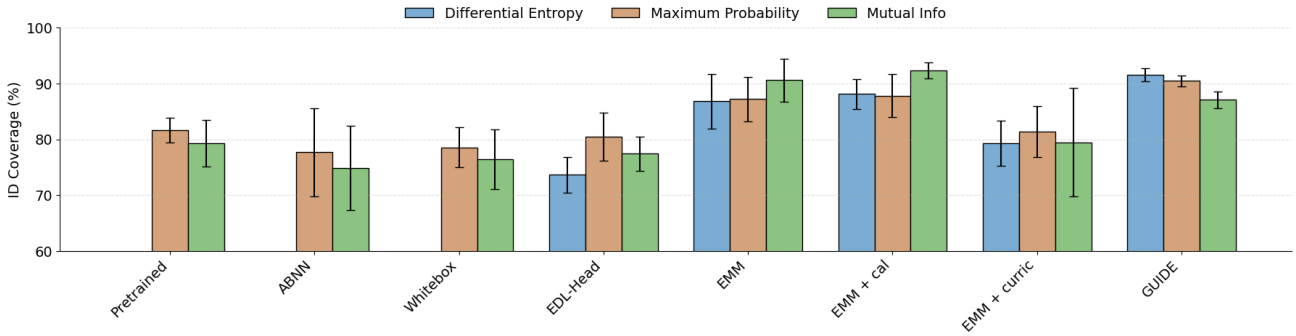
	2.0	5.0	10.0	20.0
ID Acc (%) ↑	99.87 ± 0.06	99.87 ± 0.04	99.70 ± 0.10	99.75 ± 0.11
ID Cov (%) ↑	88.21 ± 2.77	87.05 ± 1.63	89.62 ± 2.53	87.78 ± 3.84
OOD Cov (%) ↓	9.12 ± 2.25	7.34 ± 3.42	7.84 ± 2.49	7.73 ± 3.99
Adv Cov (%) ↓	9.42 ± 3.38	4.60 ± 2.69	8.98 ± 3.25	6.67 ± 1.58
AUROC (%) ↑	94.50 ± 0.32	94.85 ± 1.27	95.24 ± 0.93	94.51 ± 1.50
Adv AUROC (%) ↑	94.60 ± 0.79	95.72 ± 0.91	95.04 ± 0.54	94.58 ± 1.75

C EXPERIMENT DETAILS

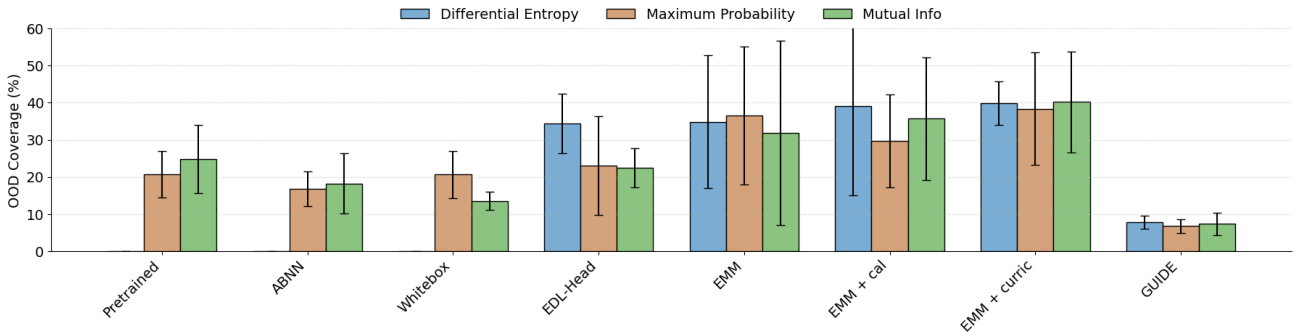
This section outlines our experimental setup in detail to ensure reproducibility and to make the evaluation protocol transparent. We first present the datasets employed for both in-distribution (ID) and out-of-distribution (OOD) evaluation, after which we describe the threshold-based scoring metrics used for uncertainty-driven rejection. Next, we summarise the comparative methods considered in our study and provide a thorough account of the model training process. Lastly, we report task-specific configurations and implementation details pertinent to each experimental setting.

C.1 DATASETS

We conduct evaluations using a broad suite of well-established computer vision benchmarks that cover different domains and levels of task complexity. This design allows for a thorough assessment of robustness and uncertainty estimation across diverse conditions. To examine the capacity of models to discriminate between in-distribution (ID) and out-of-distribution (OOD) samples, we consider both near-OOD and far-OOD scenarios. The near-OOD case involves datasets with partial class overlap relative to the ID data, whereas far-OOD datasets originate from entirely distinct visual domains. Detailed information on each dataset, including sample counts, input resolution, class composition, and data splits, is provided below. Representative examples are displayed in Figure 18.



(a) ID coverage.



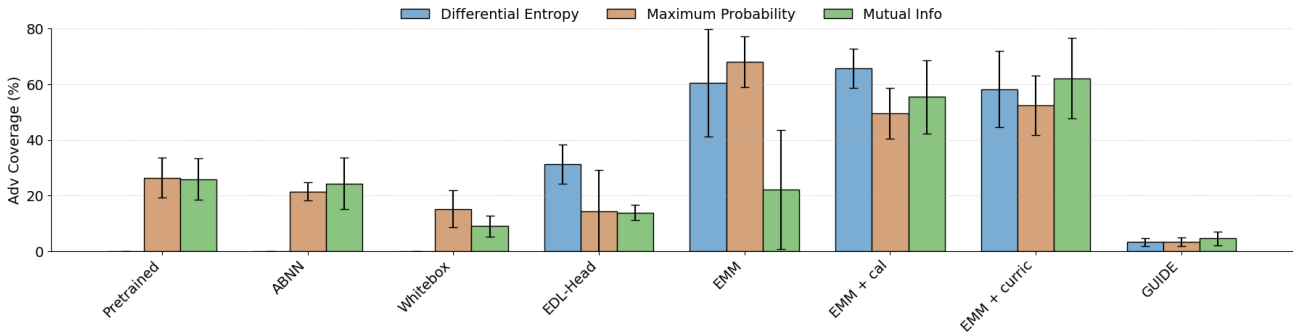
(b) OOD coverage.

Figure 14: Evaluated metrics for different ID-OOD threshold metrics (differential entropy, maximum probability, mutual information) where the ID dataset is MNIST, and the OOD dataset is FashionMNIST.

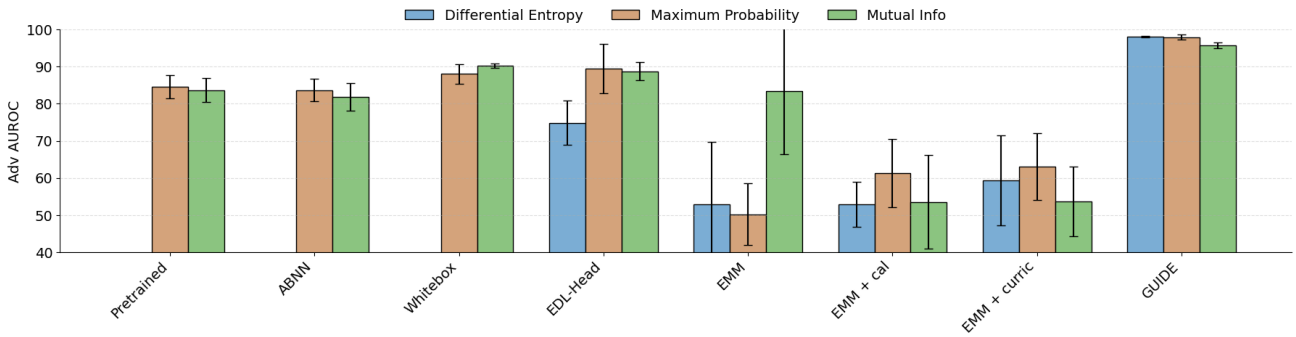
- **MNIST [LeCun et al., 1998]:** consists of 28x28 greyscale images of handwritten digits (0-9) spanning 10 classes. We utilise the widely used standard split of 60,000 training samples, 8000 test samples, and 2000 validation samples. Pixel normalisation was applied, bounded $[0, 1]$. In our experimental evaluation, MNIST serves as one of the ID datasets. MNIST was paired with FashionMNIST and KMNIST as far-OOD datasets due to them sharing no overlapping classes, but exhibit similar visual characteristics, including greyscale appearance, low resolution, and hand-drawn style. MNIST was also paired with EMNIST as a near-OOD dataset due to it sharing the same visual characteristics and sharing overlapping classes with EMNIST containing handwritten digits in addition to handwritten letters.
- **FashionMNIST [Xiao et al., 2017]:** consists of 28x28 greyscale images of clothing items (e.g., coats, bags, t-shirts, etc.) spanning 10 classes. We utilise the widely used standard split of 60,000 training sample, 8000 test samples, and 2000 validation samples. Pixel normalisation was applied, bounded $[0, 1]$. In our experimental evaluation, we use this dataset as a far-OOD pairing with the ID MNIST dataset.
- **KMNIST [Clanuwat et al., 2018]:** consists of 28x28 greyscale images of handwritten Japanese characters (specifically Kuzushiji characters from classical Japanese literature) spanning 10 classes. We utilise the widely used standard split

Table 7: The accuracy, OOD detection, and adversarial attack detection, AUROC, and adversarial AUROC performance of GUIDE with standard deviation under different values of the γ hyperparameter. The adversarial attack is L2PGD (1.0 maximum perturbation), and the ID and OOD datasets are MNIST and FashionMNIST.

	0.10	0.25	0.50	0.75	1.00
ID Acc (%) ↑	99.88 ± 0.07	99.87 ± 0.04	99.82 ± 0.07	99.84 ± 0.10	99.76 ± 0.05
ID Cov (%) ↑	87.82 ± 3.94	87.05 ± 1.63	88.31 ± 3.61	87.13 ± 3.47	91.69 ± 1.07
OOD Cov (%) ↓	9.87 ± 3.42	7.34 ± 3.42	8.92 ± 1.93	13.34 ± 1.84	12.14 ± 3.29
Adv Cov (%) ↓	8.73 ± 3.63	4.60 ± 2.69	9.79 ± 4.98	6.49 ± 3.07	13.37 ± 4.59
AUROC (%) ↑	93.78 ± 2.87	94.85 ± 1.27	94.41 ± 2.03	91.52 ± 1.71	94.28 ± 2.11
Adv AUROC (%) ↑	94.51 ± 1.66	95.72 ± 0.91	93.54 ± 2.14	94.48 ± 1.49	93.59 ± 2.55



(a) Adversarial coverage.



(b) Adversarial AUROC.

Figure 15: Evaluated metrics for different ID-OOD threshold metrics (differential entropy, maximum probability, mutual information) where the ID dataset is MNIST, and the OOD dataset is FashionMNIST.

of 60,000 training samples, 8000 test samples, and 2000 validation samples. Pixel normalisation was applied, bounded $[0, 1]$. In our experimental evaluation, we use this dataset as a far-OOD pairing with the ID MNIST dataset.

- **EMNIST [Cohen et al., 2017]:** consists of 28x28 greyscale images of both handwritten letters (from the Latin alphabet) and digits (0-9) spanning 47 classes. We use a split of 112,800 training samples, 15,040 test samples, and 3760 validation samples. Pixel normalisation was applied, bounded $[0, 1]$. In our experimental evaluation, we use this dataset as a near-OOD pairing with the ID MNIST dataset due to it sharing the same visual characteristics and overlapping classes with MNIST (handwritten digits).
- **CIFAR10 [Krizhevsky et al., 2009]:** consists of 32x32 RGB images of real-world objects (e.g, birds, trucks, airplanes, frogs, etc) spanning 10 classes. We utilise the widely used standard split of 50,000 training samples, 8000 test samples, and 2000 validation samples. Pixel normalisation was applied, bounded $[0, 1]$ across all three RGB channels. In our experimental evaluation, CIFAR10 serves as one of the ID datasets. CIFAR10 was paired with FashionMNIST and SVHN as a far-OOD dataset due to them sharing no overlapping classes, but exhibiting similar visual characteristics, including coloured appearance, real-world pictures. CIFAR-10 was also paired with CIFAR-100 as a near-OOD dataset, due to the datasets sharing the same visual characteristics and overlapping classes. Specifically, CIFAR-10 contains broad object categories (e.g., cat, dog, truck, ship), which correspond to finer-grained classes like (e.g., house cat, beagle, pickup truck, cruise ship) in CIFAR-100.
- **CIFAR100 [Krizhevsky et al., 2009]:** consists of 32x32 RGB images of real-world objects spanning of 100 classes that are fine-grained counterparts of the classes from CIFAR10. For example, CIFAR10 has the class *truck*, whilst CIFAR100 has the classes *pickup truck* and *train*. We utilise the widely used standard split of 50,000 training samples, 8000 test samples, and 2000 validation samples. Pixel normalisation was applied, bounded $[0, 1]$ across all three RGB channels. In our experimental evaluation, we use this dataset as a near-OOD pairing with the ID CIFAR10 dataset due to it sharing the same visual characteristics and overlapping classes.
- **SVHN [Netzer et al., 2011]:** consists of 32x32 RGB images of house numbers collected from Google Street View spanning 10 classes (0-9). We use a split of 73,257 training samples, 10,832 test samples, and 5200 validation samples. Pixel normalisation was applied, bounded $[0, 1]$ across all three RGB channels. In our experimental evaluation, we use this dataset as a far-OOD pairing with the ID CIFAR10 dataset due to the substantial domain shift between street-level

Table 8: The accuracy, OOD detection, and adversarial attack detection, AUROC, and adversarial AUROC performance with standard deviation of GUIDE under different values of the η hyperparameter. Also included are the selected layers for each η, and the training and inference time, in seconds, to train and infer upon the whole dataset. The adversarial attack is L2PGD (1.0 maximum perturbation), and the ID and OOD datasets are MNIST and FashionMNIST. For compactness, we denote adversarial AUROC as AAUROC, Training Time as Train. Time, and Inference Time as Inf. Time.

	0.50	0.60	0.70	0.80	0.90	1.00
ID Acc (%) ↑	99.78 ± 0.08	99.84 ± 0.06	99.78 ± 0.10	99.83 ± 0.07	99.87 ± 0.04	99.90 ± 0.03
ID Cov (%) ↑	88.21 ± 2.23	85.04 ± 1.54	87.67 ± 3.61	86.65 ± 3.56	87.05 ± 1.63	87.73 ± 3.65
OOD Cov (%) ↓	7.40 ± 2.89	7.86 ± 2.64	6.21 ± 3.07	7.41 ± 1.13	7.34 ± 3.42	6.64 ± 1.25
Adv Cov (%) ↓	5.81 ± 2.49	5.86 ± 1.26	5.95 ± 1.51	5.37 ± 1.32	4.60 ± 2.69	7.71 ± 2.85
AUROC (%) ↑	94.82 ± 2.01	93.89 ± 1.17	95.32 ± 1.27	94.35 ± 1.35	94.85 ± 1.27	95.11 ± 1.43
AAUROC (%) ↑	95.45 ± 1.28	94.29 ± 0.62	95.14 ± 1.14	94.43 ± 2.12	95.72 ± 0.91	94.64 ± 1.39
Selected Layers	{pool1}	{pool1, conv1}	{pool1, conv1}	{pool1, conv1}	{pool1, conv1, pool2}	{pool1, conv1, pool2, conv2}
Train. Time (s)	287.87 ± 1.91	285.83 ± 5.64	289.87 ± 3.47	287.56 ± 4.55	288.45 ± 4.08	295.06 ± 4.01
Inf. Time (s)	1.66 ± 0.48	1.77 ± 0.45	1.65 ± 0.50	1.68 ± 0.47	1.94 ± 0.56	2.38 ± 0.67

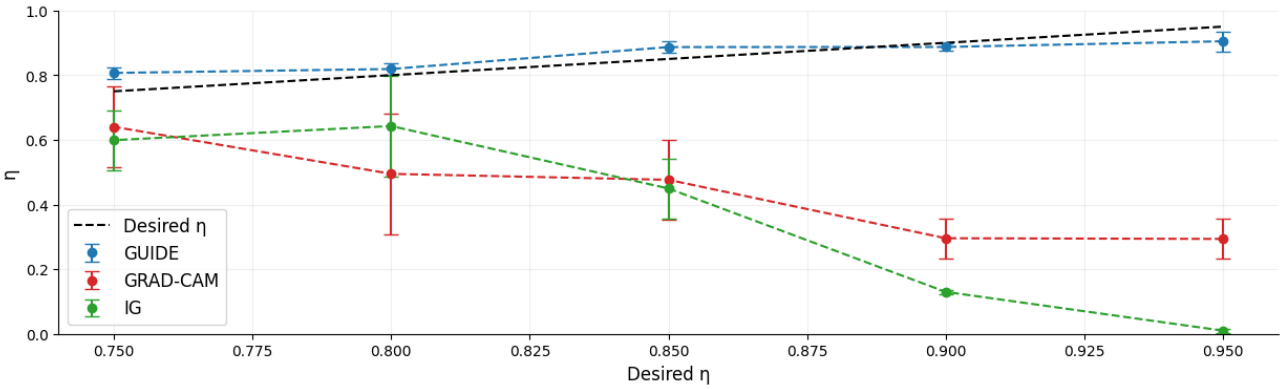


Figure 16: Visual approximation of Theorem 1: achieved cumulative relevance coverage (η) under saliency calibration versus the desired target for GUIDE's LRP-based calibration, as well as GRAD-CAM and Integrated Gradient alternatives. The empirical estimates, obtained via Jacobian-based Fisher information, closely track or exceed the theoretical bound.

digit photographs and object-centric natural images.

- **Oxford Flowers [Nilsback and Zisserman, 2008]:** consists of 64x64 RGB images of flowers commonly found in the United Kingdom, spanning 102 (e.g., Water Lily, Wild Pansy, etc.) classes. We utilise a split of 9826 training samples, 1638 test samples, and 1638 validation samples. This is double the size of the original dataset, as each image has been duplicated with a random augmentation to aid generalisation within training. Pixel normalisation was applied, bounded $[0, 1]$ across all three RGB channels. In our experimental evaluation, we use this dataset as an ID dataset. Oxford Flowers was paired with Deep Weeds because they share no overlapping classes but exhibit similar visual characteristics, including features found in nature (leaves, flowers, etc.).
- **Deep Weeds [Olsen et al., 2019]:** consists of 64x64 RGB images of species of weeds found in Australia, spanning over 8 classes (e.g., Snake weed, Rubber vine, etc). We use the standard split of 10,505 training samples, 3502 test samples, and 3502 validation samples. Pixel normalisation was applied, bounded $[0, 1]$ across all three RGB channels. In our experimental evaluation, Deep Weeds serves as a far-OOD pairing with the ID Oxford Flowers dataset due to the substantial domain shift between photographs of flowers and photographs of weeds.
- **MNIST-C [Mu and Gilmer, 2019]:** consists of the same data points as the traditional MNIST dataset, but where the images have been corrupted with 15 different distributional shifts (e.g., shot noise, glass blur, zigzag artefacts, etc.). In our experimental evaluation, we use this dataset as a near-OOD pairing with the ID MNIST dataset due to it sharing the same dataset but with some distributional shift.
- **CIFAR-10-C [Hendrycks and Dietterich, 2019]:** consists of the same data points as the traditional CIFAR-10 dataset, but where the images have been corrupted with 15 different distributional shifts (e.g., shot noise, JPEG-artefacts, zigzag

Table 9: Mean accuracy, OOD detection, and adversarial attack detection performance with standard deviation of the comparative approaches with a variety of ID and corrupted OOD datasets in order of dataset difficulty. The adversarial attack is an L2PGD attack. Highlighted cells denote the best performance for each metric.

	Pretrained	**ABNN**	**EDL-Head**	**Whitebox**	**EMM**	**EMM(+cal)**	**EMM(+curric)**	**GUIDE**
Type	-	Intrusive	Intrusive	Post-hoc	Post-hoc	Post-hoc	Post-hoc	Post-hoc
MNIST → MNIST-C								
ID Acc ↑	99.84 ± 0.07	99.77 ± 0.06	99.63 ± 0.14	99.84 ± 0.05	99.55 ± 0.11	99.48 ± 0.20	98.58 ± 0.61	99.96 ± 0.03
ID Cov ↑	72.45 ± 5.42	75.63 ± 7.48	71.78 ± 5.37	74.75 ± 4.02	82.97 ± 5.28	83.93 ± 2.07	63.17 ± 33.83	71.56 ± 3.76
OOD Cov ↓	39.75 ± 5.08	44.04 ± 6.57	37.72 ± 4.32	36.85 ± 4.08	47.26 ± 9.15	52.37 ± 6.62	61.21 ± 32.91	30.13 ± 1.51
Adv Cov ↓	20.84 ± 4.18	26.56 ± 6.47	16.26 ± 3.92	18.51 ± 3.46	52.51 ± 11.06	57.37 ± 9.54	56.59 ± 25.79	4.40 ± 1.41
AUROC ↑	71.58 ± 1.27	69.80 ± 1.11	71.58 ± 1.05	73.81 ± 1.23	69.40 ± 6.65	65.17 ± 6.13	48.62 ± 2.34	75.47 ± 1.76
Adv AUROC ↑	83.16 ± 1.04	80.35 ± 1.53	85.17 ± 0.81	84.72 ± 1.05	54.92 ± 9.90	50.30 ± 9.97	49.04 ± 2.71	91.19 ± 1.32
CIFAR-10 → CIFAR-10-C (Severity 1)								
ID Acc ↑	96.80 ± 1.95	96.32 ± 0.38	89.90 ± 1.76	97.46 ± 1.35	95.82 ± 0.96	93.70 ± 2.31	53.90 ± 34.75	96.34 ± 1.02
ID Cov ↑	49.35 ± 8.40	63.42 ± 1.00	56.68 ± 8.88	43.59 ± 7.37	46.26 ± 5.95	56.15 ± 8.52	61.90 ± 31.37	39.97 ± 6.49
OOD Cov ↓	39.93 ± 8.72	54.04 ± 1.35	44.89 ± 8.68	34.03 ± 7.24	37.34 ± 5.44	47.65 ± 8.87	60.28 ± 31.80	32.21 ± 6.46
Adv Cov ↓	38.78 ± 15.49	66.33 ± 1.73	27.39 ± 7.69	27.85 ± 9.97	25.29 ± 5.68	38.35 ± 12.53	52.98 ± 34.86	14.78 ± 5.30
AUROC ↑	56.54 ± 0.28	55.19 ± 0.61	58.49 ± 0.60	56.66 ± 0.38	56.15 ± 0.65	56.21 ± 0.90	49.37 ± 4.18	55.16 ± 0.67
Adv AUROC ↑	55.52 ± 1.69	47.34 ± 1.09	70.45 ± 0.90	56.48 ± 0.87	62.00 ± 1.40	62.25 ± 1.57	58.31 ± 10.07	66.37 ± 0.86
CIFAR-10 → CIFAR-10-C (Severity 5)								
ID Acc ↑	95.79 ± 0.69	96.27 ± 0.25	94.40 ± 0.72	94.73 ± 1.83	93.12 ± 1.73	93.28 ± 3.81	81.49 ± 26.89	93.33 ± 0.55
ID Cov ↑	56.04 ± 3.45	62.78 ± 1.76	51.66 ± 2.37	58.75 ± 8.03	59.39 ± 4.67	55.29 ± 5.94	38.79 ± 24.72	63.36 ± 2.82
OOD Cov ↓	25.35 ± 1.63	30.71 ± 1.53	22.84 ± 2.36	27.99 ± 6.99	31.73 ± 6.07	27.66 ± 4.22	27.45 ± 22.50	26.67 ± 1.98
Adv Cov ↓	46.85 ± 5.39	61.89 ± 4.97	62.85 ± 2.78	42.24 ± 13.67	37.42 ± 10.24	33.00 ± 6.33	26.11 ± 21.18	31.72 ± 4.50
AUROC ↑	70.13 ± 1.71	67.99 ± 0.79	67.95 ± 0.44	70.28 ± 1.46	68.63 ± 2.10	68.35 ± 3.38	54.26 ± 8.00	74.17 ± 1.01
Adv AUROC ↑	58.42 ± 1.36	48.16 ± 1.62	41.73 ± 2.41	60.98 ± 1.13	64.90 ± 3.81	64.52 ± 2.76	56.12 ± 9.39	71.87 ± 1.83

artefacts, etc.). In our experimental evaluation, we use this dataset as a near-OOD pairing with the ID CIFAR-10 dataset due to it sharing the same dataset but with some distributional shift.

For datasets lacking predefined partitions, we created manual splits using stratification to maintain consistent class distributions across subsets. In all experiments, the validation portion of each dataset was employed to calibrate the ID–OOD threshold for uncertainty-based rejection. This calibration enabled evaluation of both ID and OOD coverage on the corresponding test sets.

C.2 ID-OOD THRESHOLDS

To support abstention from uncertain predictions, we determine the optimal ID–OOD threshold using the validation set corresponding to each experimental configuration. Our evaluation considers several ID–OOD scoring metrics, enabling a rigorous assessment of GUIDE alongside the baseline methods.

- **Differential Entropy:** measures the dispersion or uncertainty inherent in the Dirichlet distribution. It is given by:
$$\sum_{k=1}^{K} \ln \Gamma(\alpha_k) - \ln \Gamma(S) - \sum_{k=1}^{K} (\alpha_k - 1)(\Psi(\alpha_k) - \Psi(S))$$
where $B(\boldsymbol{\alpha})$ represents the multivariate Beta function, S is the overall concentration parameter, and $\Psi(\cdot)$ is the digamma function. Larger entropy values indicate greater uncertainty, which is typically observed for OOD inputs. This score applies only to Dirichlet-based uncertainty models (e.g., posterior networks, evidential networks).
- **Mutual Information:** captures epistemic uncertainty by quantifying the amount of information that the model parameters contribute to the predictive distribution. It is expressed as the gap between the predictive entropy and the expected conditional entropy under the posterior over parameters:
$$-H\big(\mathbb{E}_{q(\omega)}[\,p(y \mid x, \omega)\,]\big) \;-\; \mathbb{E}_{q(\omega)}[\,H(p(y \mid x, \omega))]$$
where $H(p) = -\sum_{k=1}^{K} p_k \log p_k$ and $q(\omega)$ is the (explicit or implicit) posterior over parameters.
- **Maximum Probability:** is a simple confidence score that measures the model's most likely prediction. It is defined as the maximum softmax probability over all classes:
$$\max_{k \in \{1,\ldots,K\}} \; p(y = k \mid x)$$
Higher values indicate that the model is more confident in its predicted class, while lower values suggest greater uncertainty.

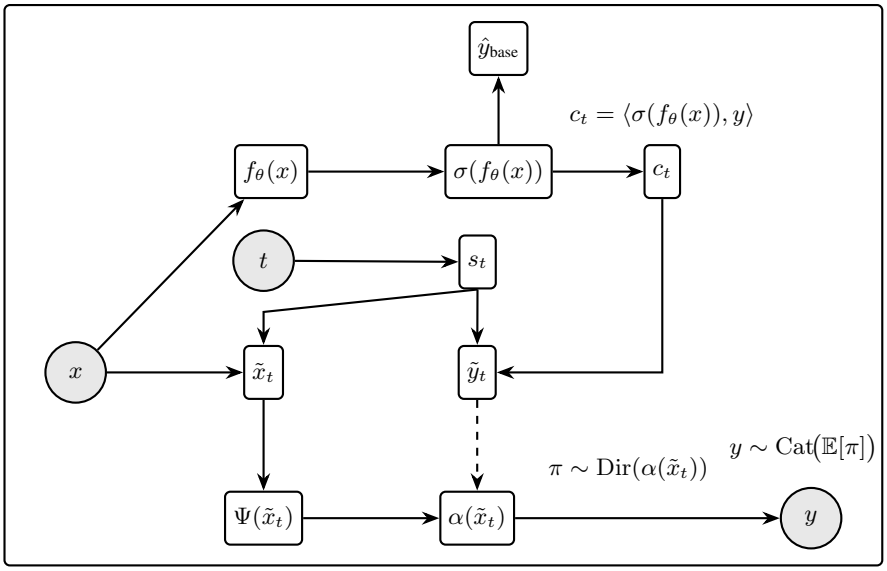


Figure 17: Graphical model view of GUIDE.

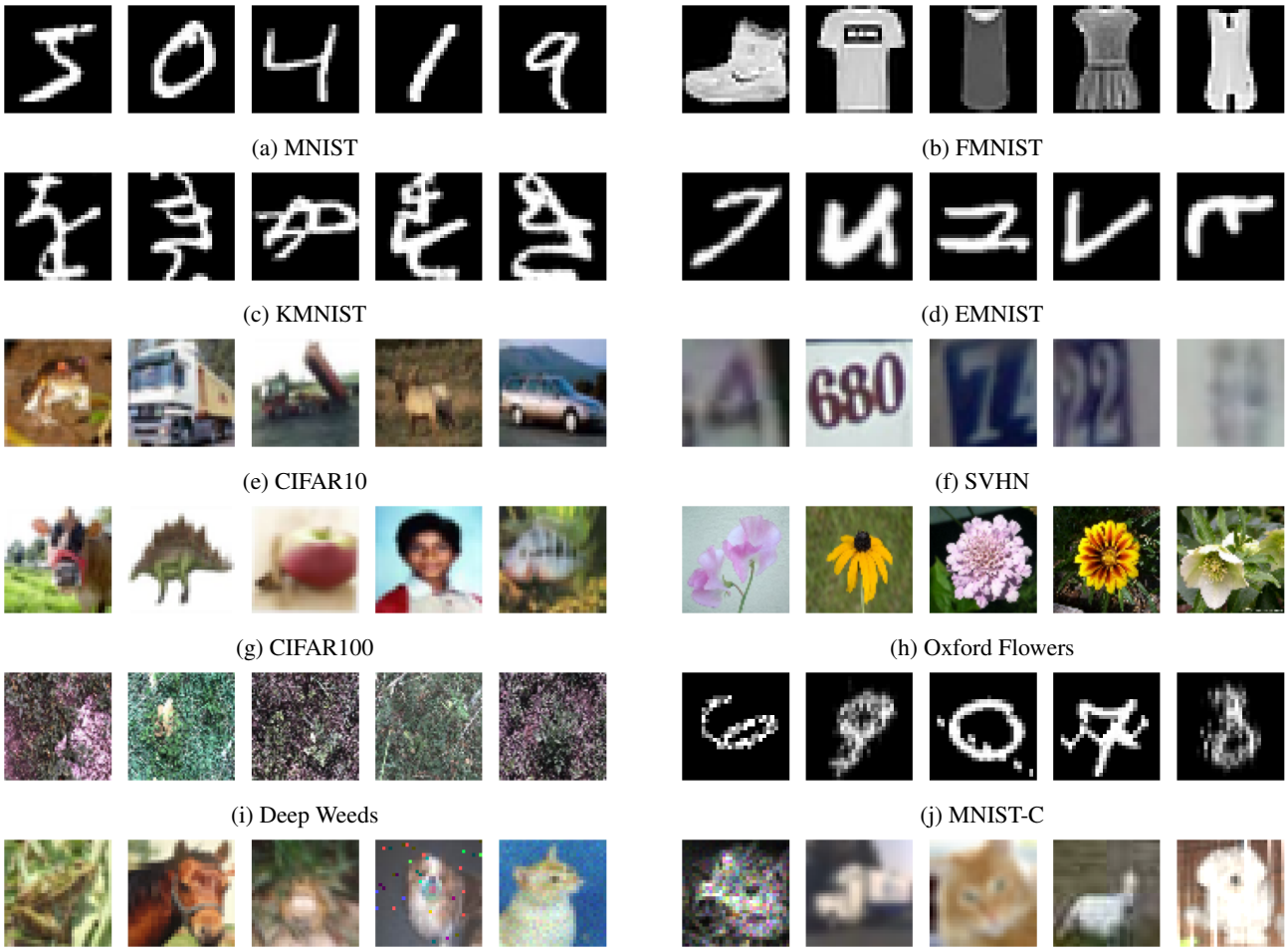
In line with prior works [Shen et al., 2023], the optimal ID–OOD threshold is determined using the validation datasets. For a selected scoring metric (chosen from the list above), we first compute scores on both ID and OOD validation samples. These scores are then used to construct a receiver operating characteristic (ROC) curve, where ID samples are regarded as positive and OOD samples as negative. The ROC curve provides a visualisation of how well the chosen metric separates ID from OOD data. To identify the decision boundary, we calculate the threshold that maximises TPR $-$ FPR, yielding the point of optimal separation. This criterion achieves a principled balance between retaining ID inputs and filtering out OOD inputs, avoiding the need for ad hoc tuning, and enabling clear reporting of ID and OOD coverage under deployment-like conditions.

After this calibration stage, the threshold is fixed and subsequently applied to the test sets. During evaluation, any test input whose score (under the chosen metric) crosses this threshold is rejected. This provides a consistent basis for comparing ID retention and OOD rejection across all models and datasets.

C.3 COMPARATIVE APPROACHES

To assess the performance of GUIDE, we benchmark it against a set of recent post-hoc uncertainty estimation methods that capture the state of the art in this area. Below, we provide a concise description of each method together with the implementation details and hyperparameter settings adopted in our experiments:

- **Pretrained classifier:** refers to the frozen base neural network on which all post-hoc methods, including GUIDE, are applied. This model is trained on the in-distribution training data using conventional cross-entropy loss and outputs softmax probabilities over the class set. It serves as the deterministic baseline without an explicit uncertainty modelling component.
- **Adaptable Bayesian Neural Networks (ABNN) [Franchi et al., 2024]:** provide a lightweight post-hoc strategy for equipping pretrained deterministic networks with Bayesian uncertainty estimation capabilities. Rather than retraining the full model or relying on costly ensembles, ABNN introduces Bayesian Normalization Layers (BNLs), which inject Gaussian perturbations into existing normalization layers. Only the parameters of these BNLs are fine-tuned for a few epochs, enabling the deterministic DNN to be transformed into a scalable approximation of a BNN with minimal overhead. During inference, multiple stochastic forward passes are performed by sampling perturbations, producing diverse predictions that capture epistemic uncertainty.
- **Whitebox [Chen et al., 2019]:** is a post-hoc meta-model based confidence scoring approach that observes the internal representations of a base classifier. It introduces linear classifier probes at multiple intermediate layers of the base model, and trains an auxiliary meta-model on the probe outputs to predict whether the base model's prediction is correct.
- **EDL-Head [Sensoy et al., 2018]:** represents a simple evidential deep learning baseline built directly on top of a



(a) MNIST (b) FMNIST

(c) KMNIST (d) EMNIST

(e) CIFAR10 (f) SVHN

(g) CIFAR100 (h) Oxford Flowers

(i) Deep Weeds (j) MNIST-C

(k) CIFAR-10-C (severity 1) (l) CIFAR-10-C (severity 5)

Figure 18: Example images from the datasets used in our experimental evaluation.

pretrained classifier. The softmax output layer of the pretrained model is removed and replaced with an evidential head that predicts the Dirichlet concentration parameters α for each class. The evidential head is then trained while keeping the backbone fixed, allowing the model to encode both aleatoric and epistemic uncertainty through the resulting Dirichlet distribution. This setup provides a direct comparison to standard softmax classifiers by isolating the effect of evidential parameterisation on uncertainty estimation.

- **Dirichlet Meta-Model (EMM) [Shen et al., 2023]:** is a post-hoc uncertainty learning method that introduces a meta-model that leverages intermediate feature representations from the base model and parameterises a Dirichlet distribution over class probabilities. By training only this lightweight meta-model, EMM captures both aleatoric and epistemic uncertainty while leaving the predictive performance of the base model unchanged. Training is performed using an ELBO loss that balances classification accuracy with uncertainty calibration, regularised by a KL-divergence term to avoid overconfidence. During inference, the Dirichlet meta-model produces concentration parameters whose dispersion reflects uncertainty, enabling effective applications to OOD detection, misclassification detection, and trustworthy transfer learning.
- **EMM(+cal):** extends the Dirichlet Meta-Model (EMM) [Shen et al., 2023] with the saliency-based calibration stage from GUIDE. Instead of arbitrarily selecting intermediate layers for the meta-model, this variant uses relevance propagation to identify the most salient layers of the pretrained backbone, ensuring that the evidential head is trained on semantically meaningful features.
- **EMM(+curric):** modifies the Dirichlet Meta-Model (EMM) [Shen et al., 2023] by incorporating only the uncertainty-guided curriculum component from GUIDE, while omitting the saliency-based calibration stage and the new loss formulation. In this setup, the evidential head is trained on arbitrarily chosen intermediate features, but training proceeds

under a monotonic noise-driven curriculum that gradually corrupts inputs from low to high noise levels.

C.4 ADVERSARIAL ATTACKS

This section describes the adversarial attack strategies employed to evaluate the robustness of uncertainty-aware models. Adversarial attacks consist of carefully designed, imperceptible perturbations that can cause models to misclassify while still producing highly confident predictions. Such manipulations pose a particular challenge for uncertainty-based systems, as they can distort both aleatoric and epistemic signals, leading adversarial inputs to resemble in-distribution samples or bypass detection mechanisms.

To ensure a thorough robustness evaluation, we examine both gradient-driven and gradient-free attacks. Gradient-based methods, characteristic of white-box settings, exploit direct access to model gradients in order to craft adversarial perturbations. In contrast, gradient-free approaches inject random or structured noise into inputs and are more representative of black-box adversarial scenarios. All attacks in our study are implemented using Foolbox [Rauber et al., 2017], covering the following configurations:

- **L2 Projected Gradient Descent (L2PGD):** is an iterative, white-box adversarial attack that perturbs inputs within a bounded L_2-norm ball to maximise the model's loss. At each iteration, the input is updated in the direction of the gradient of the loss with respect to the input, followed by projection back onto the L_2-ball of radius ϵ. This results in smooth, high-precision perturbations that remain less perceptible to humans:
$$x'_{t+1} = \text{Proj}_{\epsilon}^{L_2}(x'_t + \alpha \cdot \nabla_x J(x'_t, y))$$
where α is the step size and $J(x, y)$ is the loss function.
- **Fast Gradient Sign Method (FGSM):** is a single-step white-box adversarial attack that perturbs the input in the direction of the sign of the gradient of the loss:
$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y))$$
where ϵ controls the perturbation magnitude. FGSM generates perceptible but targeted perturbations with minimal computational overhead.
- **Salt & Pepper Noise:** is a non-gradient-based black-box perturbation that randomly sets a proportion of input pixels to their minimum or maximum value. This form of structured noise simulates impulsive corruption and tests the model's resilience to sparse, high-intensity artefacts. It does not rely on model gradients and is agnostic to internal model parameters.
- **AutoAttack (L_2):** is a standardised, parameter-free evaluation suite that applies a sequence of complementary white-box and query-based components (e.g., APGD variants, FAB, and Square Attack) within the same threat model to reduce the risk of overestimating robustness due to a single attack choice [Croce and Hein, 2020].
- **AutoAttack (L_∞):** uses the same suite structure under an L_∞ constraint [Croce and Hein, 2020]. We evaluate $\epsilon \in \{2, 4, 8\}/255$, which is a standard range for pixel-bounded perturbations [Croce et al., 2020].

C.5 GUIDE ALGORITHM

Algorithm 1 summarises the full GUIDE procedure. The pretrained classifier remains frozen throughout; only the parameters of the evidential meta-model are updated.

C.6 TRAINING DETAILS

To ensure consistency and reproducibility across all experiments, we adopt a fixed training configuration and architectural setup for all models, including our proposed GUIDE variants and baseline comparators.

All models were trained using a custom convolutional neural network architectures. Table 10 shows the architecture of these models. LeNet-5 was used for MNIST dataset pairs, ResNet-20 was used for CIFAR-10 $\rightarrow$ SVHN, and SE-Net was used for CIFAR-10 $\rightarrow$ CIFAR-100 and Oxford Flowers dataset pairs. The output layer of every model consists of K units (one per class according to the respective ID dataset). All models are trained using the Adam optimiser [Kingma and Ba, 2014].

All experiments were implemented in TensorFlow 2.15 and executed on a large performance GPU cluster using a maximum of three Nvidia A40 GPUs, 32 CPU cores, 167GB of memory (per GPU). All models were trained from scratch, and no pretraining or external data was used. All runs used random seeds.

Algorithm 1 GUIDE: Gradual Uncertainty Refinement via Noise-Driven Curriculum

Require: Frozen pretrained classifier f_θ; calibration data $\mathcal{D}_{\text{cal}}$; relevance coverage threshold η; curriculum steps T; corruption rate γ; training epochs E

Ensure: Trained evidential meta-model g_ϕ and selected salient layers L_{sal}

Freeze f_θ and set it to inference mode
for each $(x_i, y_i) \in \mathcal{D}\text{cal}$ **do**
    Compute logits $z_i = f_\theta(x_i)$ and initialise output relevance using y_i
    Propagate relevance through f_θ using the LRP-ϵ rule
    Accumulate layer relevance scores M_ℓ using Eq. 2
    Store the input-level weight map $\mathcal{W}(x_i)$ using Eq. 4
end for
Select the smallest salient layer set L_{sal} satisfying Eq. 3
Construct g_ϕ using projection branches from layers in L_{sal} and a Dirichlet output head
for epoch $e = 0, \ldots, E-1$ **do**
    Set curriculum difficulty $\rho_e = \big(e/(E-1)\big)^2$
    for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{cal}}$ **do**
        for each $(x, y) \in \mathcal{B}$ **do**
            Sample a curriculum stage $t \in 0, \ldots, T$ using $\kappa_e(s) \propto (1-\rho_e)(1-s) + \rho_e s$
            Compute the corruption level $s_t = 1 - \exp(-\gamma t)$
            Compute saliency-weighted corruption probabilities $p_t(h, w)$ using Eq. 5
            Corrupt x with the fixed base mask rule in Eq. 6 to obtain $\tilde{x}_t$
            Compute base-model confidence $c_t = \langle \sigma(f_\theta(\tilde{x}_t)), y \rangle$
            Form the soft target $\tilde{y}_t = \frac{\tilde{s}_t}{K}\mathbf{1}K + (1 - \tilde{s}t)y$, where $\tilde{s}_t = s_t(1 - \frac{1}{2}c_t^2)$
        end for
        Extract frozen features $\Phi_\ell(\tilde{x}_t)_{\ell \in L_{\text{sal}}}$
        Predict Dirichlet parameters $\alpha(\tilde{x}_t) = g_\phi(\Phi_\ell(\tilde{x}_t)_{\ell \in L\text{sal}})$
        Update ϕ by minimising the GUIDE loss in Eq. 7
    end for
end for
return g_ϕ, L_{sal}

Under our parameterisation, the GUIDE loss is continuously differentiable with respect to the meta-model parameters ϕ: the map $\phi \mapsto \alpha(x)$ is smooth (affine layers composed with $\exp$ ensure $\alpha_k > 0$), the Dirichlet-ELBO terms involve C^∞ functions (gamma/digamma), and the SRE component is a smooth rational function of α. If non-smooth activations (e.g., ReLU) are used internally, the GUIDE loss is differentiable almost everywhere, and we rely on subgradients, as is standard in first-order optimisation.

C.7 EXPERIMENTAL SETUPS

This subsection outlines additional implementation details and experimental choices that are specific to certain experiments that may help with reproducibility and transparency.

For practical stability, we restrict candidate layers to semantically meaningful, deterministic representations (Conv/Linear, Norm in inference mode, Pool/Flatten, and post-activation block outputs). We exclude stochastic layers (Dropout/Noise), non-differentiable or discrete operators (e.g., Argmax/Top-k/Quantize implemented via Lambda), softmax activations, and element-wise merge/gate nodes (Add/Multiply/Min/Max/Average). All relevance and feature extraction are computed with the backbone in inference mode to avoid perturbing internal statistics.

In particular, for architectures with explicit residual or modular stage boundaries (e.g., ResNet, WideResNet), we consider only post-activation outputs at the end of each residual block or stage (after the skip connection addition and final ReLU), as these represent the actual semantic feature maps propagated forward in the backbone. For non-residual models (e.g., VGG, LeNet, SENet), we take stage-level post-activation features after pooling layers, global average pooling, or fully connected layers, as appropriate. This ensures that all tapped representations are stable, comparable in scale, and maximally informative for the meta-model.

Table 10: Architectures and hyperparameters of the base networks used within our experimental evaluation. Conv(C, k, s) = $k\times k$ conv, C channels, stride s; MP/AP = max/avg pool (2×2, s=2); BN = batch norm; GAP = global avg pool; SE(r) = squeeze–excitation (reduction r); BB(C, s) = basic ResNet block with two 3×3 convs, projection skip if s=2.

	LeNet-5	ResNet-20	SE-Net
Input	$H\times W\times D$	$H\times W\times D$	$H\times W\times D$
Stage 1	Conv(6, 5, 1), `tanh`, *pad=same* → AP [pool1]	Conv(16, 3, 1), *pad=same*, no bias → BN → ReLU	Conv(32, 3, 1) + ReLU → BN → SE(8) → MP [pool1]
Stage 2	Conv(16, 5, 1), `tanh`, *valid* → AP [pool2]	$3\times$ BB(16, 1)	Conv(64, 3, 1) + ReLU → BN → SE(8) → MP [pool2]
Stage 3	Conv(120, 5, 1), `tanh`, *valid* → Flatten	BB(32, **2**) + $2\times$ BB(32, 1)	Conv(128, 3, 1) + ReLU → BN → SE(8) → GAP
Head	FC(84), `tanh` → FC(K), softmax	BB(64, **2**) + $2\times$ BB(64, 1) → GAP → FC(K), softmax	FC(128) → Dropout(0.5) → FC(K), softmax
Epochs	50	200	200
KL-Weight (λ_{kl})	1e-1	1e-3	1e-3
Prior (β)	1.0	1.0	1.0
Learning rate	1e-2	1e-2	1e-4

In Table 1, the adversarial attack used is a gradient-based L2PGD attack and the uncertainty metric was mutual information. For MNIST to FashionMNIST, KMNIST, and EMNIST, a maximum perturbation of 1.0 was used. For CIFAR10 to SVHN, CIFAR10 to CIFAR100, and Oxford Flowers to Deep Weeds, a maximum perturbation of 0.1 was used. This choice reflects the fact that adversarial examples on natural image datasets such as CIFAR-10 require smaller perturbations to significantly degrade model performance, due to the increased complexity and lower inherent separability of the visual features. In contrast, MNIST-like datasets typically require larger perturbations to achieve a comparable adversarial effect.

For all experiments, unless stated otherwise, GUIDE and its variants utilises the hyperparameter combination of $T = 5$, $\gamma = 0.25$, and $\eta = 0.9$ for MNIST to FashionMNIST, KMNIST, and EMNIST. A combination of $T = 10$, $\gamma = 0.05$, and $\eta = 0.9$ for CIFAR10 to SVHN, CIFAR10 to CIFAR100, and Oxford Flowers to Deep Weeds.

C.8 RUNTIME AND COMPLEXITY

Table 11 reports calibration and training times for GUIDE and the comparative approaches on the MNIST→FashionMNIST setting. The additional calibration time for GUIDE corresponds to the LRP-based saliency stage used to select salient layers and construct the curriculum.

Table 11: Calibration and training times for GUIDE and comparative approaches on the MNIST→FashionMNIST setting. Calibration time refers to the additional saliency-calibration stage used by GUIDE and EMM(+cal). Results are reported as mean ± standard deviation.

	Pretrained	ABNN	EDL-Head	Whitebox	EMM	EMM(+cal)	EMM(+curric)	GUIDE
Calibration Time (s)	–	–	–	–	–	239.65 ± 6.75	–	243.31 ± 4.08
Train Time (s)	23.00 ± 0.55	92.96 ± 1.42	72.62 ± 0.97	161.02 ± 2.07	45.14 ± 7.00	48.59 ± 8.04	44.37 ± 7.79	46.24 ± 6.30

Asymptotically, GUIDE differs from EMM primarily through this offline saliency-calibration stage. Let C_f denote a frozen-model forward pass, C_b a relevance/backward pass, C_g the lightweight meta-model cost, E the number of epochs, and N the number of training samples. EMM training has complexity $O(E/cdotN(C_f + C_g))$, while GUIDE adds a one-time calibration cost of $O(N(C_f + C_b))$ and then trains with the same order when one curriculum level is sampled per example. This additional LRP cost is not incurred at deployment: GUIDE inference remains $O(C_f + C_g)$, since the selected layers are fixed and no relevance propagation or curriculum generation is performed at inference time.