POST-HOC LATENT CONFLICT-AWARE EVIDENCE ADJUSTMENT FOR ROBUST EVIDENTIAL DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty quantification is essential for deploying deep learning models in high-stakes settings, where out-of-distribution and adversarial inputs can induce confident but unreliable predictions. Evidential Deep Learning provides efficient uncertainty estimates in a single forward pass, but can still assign high evidential strength to inputs that are poorly supported by the learned representation, such as adversarial inputs. We introduce CLEAR, a lightweight, task-agnostic post-hoc method that improves evidential robustness without retraining or altering the base prediction. Using held-out calibration data, CLEAR characterises the group-conditioned geometry of the model's latent space. At inference, it efficiently generates perturbation views directly in the latent space and measures their conflict relative to the calibrated geometry of the predicted group. High latent conflict indicates unsupported evidence, which CLEAR uses to selectively reduce evidential strength while retaining evidence for latent-consistent inputs. On ImageNet→CUB, CLEAR improves OOD and adversarial AUROC by $+8.29$ and $+5.01$ while running $17.4\times$ faster than competing post-hoc methods while preserving predictive performance across classification, regression, and object detection benchmarks.

1 INTRODUCTION

Deep learning has enabled increasingly capable predictive systems across domains such as computer vision and natural language processing (Dosovitskiy et al., 2020; Achiam et al., 2023). However, reliable deployment requires more than predictive accuracy. In safety-critical settings, models may encounter out-of-distribution (OOD) inputs that differ from the data observed during training, or adversarial inputs deliberately perturbed to induce erroneous predictions. In both cases, a model may remain highly confident despite operating beyond the conditions or environment under which it has been trained. Identifying such unreliable predictions is therefore fundamental to the safe deployment of deep learning systems.

Uncertainty Quantification (UQ) provides a means of expressing how strongly a model's prediction should be trusted (Abdar et al., 2021). Approaches such as Bayesian neural networks (Goan & Fookes, 2020), Monte Carlo Dropout (Gal & Ghahramani, 2016), and conformal methods (Angelopoulos et al., 2023) can provide reliable uncertainty estimates; however, they often require additional training or repeated inference. Evidential methods offer an efficient alternative by estimating uncertainty in a single deterministic forward pass. Evidential Deep Learning (EDL) models class probabilities using a Dirichlet distribution (Sensoy et al., 2018), while Deep Evidential Regression (DER) uses a Normal-Inverse-Gamma distribution for continuous predictions (Amini et al., 2020). Despite their appeal, evidential models can still assign strong evidence to OOD and adversarial inputs for which the prediction is poorly supported (Kopetzki et al., 2021).

Post-hoc methods have emerged as a particularly effective approach for improving uncertainty robustness in evidential models, as they enhance uncertainty estimation without retraining the predictor and have demonstrated strong robustness to OOD and adversarial inputs (Kopetzki et al., 2021; Barker et al., 2026). However, existing approaches may require repeated model evaluations or transformed inputs, reducing the computational advantages of evidential learning. They may also suppress evidence for valid in-distribution (ID) inputs, reducing retention despite correct predictions.

We introduce **C**onflict-aware **L**atent **E**vidence **A**djustment for adversarial **R**obustness (**CLEAR**), a task-agnostic post-hoc approach that evaluates whether the evidence assigned by a pretrained

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

evidential model is supported by its expected latent representation. Rather than repeatedly evaluating transformed inputs through the full model, CLEAR first uses held-out ID data to calibrate the latent geometry associated with different regions of the output space. At inference, CLEAR then generates inexpensive perturbation views directly in latent space and measures their conflict relative to the calibrated geometry of the predicted output group. Predictions exhibiting latent behaviour consistent with the calibrated ID structure retain their evidential strength, whereas high latent conflict indicates unsupported evidence and results in increased uncertainty. CLEAR therefore improves uncertainty under OOD and adversarial inputs without retraining the predictor or replacing its base prediction.

2 Related Work

Uncertainty Quantification. UQ characterises uncertainty arising from incomplete model knowledge and variability in observations (He & Jiang, 2023). Input-dependent distributions capture aleatoric uncertainty (Kendall & Gal, 2017), while Bayesian inference (Goan & Fookes, 2020; Blei et al., 2017), Monte Carlo Dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), and distance-aware models (Liu et al., 2020) address epistemic uncertainty. Ensembles with prediction intervals can account for both sources (Pearce et al., 2018). Test-time augmentation measures predictive variability without explicitly separating them (Ayhan & Berens, 2018). Conformal methods provide calibrated sets or intervals whose cardinality or width can serve as uncertainty indicators (Angelopoulos et al., 2023). These approaches differ in their computational and calibration requirements, motivating efficient estimation of both uncertainty types.

Evidential Deep Learning. Evidential deep learning (EDL) applies Dempster-Shafer theory (DST) (Dempster, 1968) to model class probabilities with a Dirichlet distribution, enabling epistemic and aleatoric uncertainty estimation in a single forward pass. Its full formulation is given in Section 3. EDL has been widely used for OOD detection by assigning greater uncertainty to unexpected inputs. However, its deterministic inference lacks the multiple predictive perspectives provided by methods such as Monte Carlo Dropout or deep ensembles, making overconfident errors difficult to recover from. This limitation is particularly evident under adversarial perturbations (Kopetzki et al., 2021).

Post-hoc EDL methods can improve the robustness of evidential models to adversarial and distribution-shifted inputs by refining uncertainty after the base model has been trained (Barker et al., 2026; Kopetzki et al., 2021). However, many approaches rely on multiple forward passes or stochastic sampling, increasing computational cost and sacrificing one of EDL's key advantages: single-pass uncertainty estimation. Others are tailored to specific tasks, e.g., classification (Barker et al., 2026), limiting their generality. These limitations motivate a lightweight post-hoc approach that strengthens evidential uncertainty under challenging inputs while preserving the efficiency and task flexibility of the underlying task.

3 Preliminaries

We consider a supervised learning setting where $\mathcal{X}$ and $\mathcal{Y}$ denote the input and task-dependent output spaces, respectively. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from an underlying distribution $p(x, y)$, the objective is to learn a predictive model with a latent representation $h(x) \in \mathbb{R}^d$ that captures the relationship between x and y and generalizes to unseen samples. CLEAR is a task-agnostic, post-hoc method that operates on pretrained evidential models. We begin by introducing the two evidential formulations considered throughout this work.

Classification. Each target y_i denotes one of K classes, such that $y_i \in \{1, \dots, K\}$. Evidential Deep Learning (EDL) replaces the softmax output of the model with a Dirichlet distribution over class probabilities. The model predicts non-negative evidence to define the Dirichlet parameters:

$$e(x) \in \mathbb{R}_+^K, \qquad \alpha(x) = e(x) + 1 \tag{1}$$

An EDL model is trained by matching the Dirichlet mean to the one-hot target while discouraging unsupported evidence. For a one-hot target y_i, the standard objective combines a squared-error term, the Dirichlet predictive variance, and a KL penalty to the uniform Dirichlet prior:

$$\mathcal{L}_{\mathrm{EDL}} = \sum_{k=1}^{K} (y_{ik} - \hat{p}_k(x_i))^2 + \sum_{k=1}^{K} \frac{\hat{p}_k(x_i)(1 - \hat{p}_k(x_i))}{S(x_i) + 1} + \lambda_{\mathrm{KL}} \mathrm{KL}(\mathrm{Dir}(\alpha(x_i)) \,\|\, \mathrm{Dir}(\mathbf{1})) \tag{2}$$

Given an input x_i, the predictive categorical mean, total evidential strength, and scalar uncertainty are:

$$\hat{p}_k(x_i) = \frac{\alpha_k(x_i)}{S(x_i)}, \qquad S(x_i) = \sum_{k=1}^{K} \alpha_k(x_i), \qquad u(x_i) = \frac{K}{S(x_i)} \tag{3}$$

High $S(x_i)$ signifies concentrated evidence around the predicted class, while low $S(x_i)$ indicates uncertainty.

Regression. Each target y_i denotes a continuous value, i.e., $y_i \in \mathbb{R}$. Deep Evidential Regression (DER) models uncertainty by placing a Normal-Inverse-Gamma (NIG) distribution over the parameters of a Gaussian likelihood:

$$y \sim \mathcal{N}(\mu, \sigma^2), \qquad (\mu, \sigma^2) \sim \mathrm{NIG}(\gamma, \nu, \alpha, \beta) \tag{4}$$

where γ represents the predictive mean and $\nu > 0$, $\alpha > 1$, and $\beta > 0$ determine the evidential strength and uncertainty. A DER model is trained using the NIG negative log-likelihood together with an evidence regulariser that penalises confident but inaccurate predictions:

$$\begin{aligned} \mathcal{L}_{\mathrm{DER}} = \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha \log \Omega + \left(\alpha + \frac{1}{2}\right) \log\left((y-\gamma)^2\nu + \Omega\right) \\ + \log\Gamma(\alpha) - \log\Gamma\left(\alpha + \frac{1}{2}\right) + \lambda_{\mathrm{reg}}|y-\gamma|(2\nu + \alpha) \end{aligned} \tag{5}$$

where $\Omega = 2\beta(1+\nu)$. For an input, the predictive mean and combined aleatoric and epistemic variance are defined as:

$$\mathbb{E}[y] = \gamma, \qquad \mathrm{Var}(y) = \frac{\beta(1+\nu)}{\nu(\alpha-1)} \tag{6}$$

The aleatoric and epistemic uncertainties are defined as:

$$u_{\mathrm{alea}}(x) = \frac{\beta(x)}{\alpha(x)-1}, \qquad u_{\mathrm{epi}}(x) = \frac{\beta(x)}{\nu(x)(\alpha(x)-1)} \tag{7}$$

Both EDL and DER encode prediction confidence through evidential strength: Dirichlet concentration in classification and NIG strength in regression. However, high evidence can still be assigned to inputs whose latent representation is inconsistent with the learned structure of the predicted class or output region. CLEAR addresses this by recalibrating evidential strength post-hoc using a latent consistency test, while leaving the base prediction unchanged.

4 Conflict-Aware Latent Evidence Adjustment for Adversarial Robustness

Our **C**onflict-aware **L**atent **E**vidence **A**djustment for adversarial **R**obustness (**CLEAR**) approach, whose high-level workflow is shown in Figure 1, boosts uncertainty under OOD and adversarially attacked inputs by determining whether a model's evidential strength is supported by familiar ID latent structure. During calibration, CLEAR characterises the latent structure of ID data from a pretrained evidential model without modifying or retraining the predictor. At inference, the model produces a latent representation $h(x)$ and evidential output, from which CLEAR samples multiple latent perturbation views and forms all unordered pairs between them. Their compatibility with the calibrated latent geometry of the predicted group is then measured, and evidential strength is down-weighted when conflict is high. Consequently, inputs inconsistent with the learned ID structure receive greater uncertainty, while evidence for compatible ID inputs is preserved.

4.1 Latent Calibration

CLEAR first estimates the approximate latent geometry associated with each output group using a held-out calibration set $\mathcal{D}_{\mathrm{cal}}$. The evidential model is frozen, and CLEAR extracts the latent representation of each calibration example $h_i = h(x_i)$. Each example is assigned to an output group g_i. For classification, the group corresponds directly to the class label $g_i = y_i$. For regression, where the target space is continuous, targets are partitioned into B bins and $g_i = b(y_i)$ where $b(\cdot)$

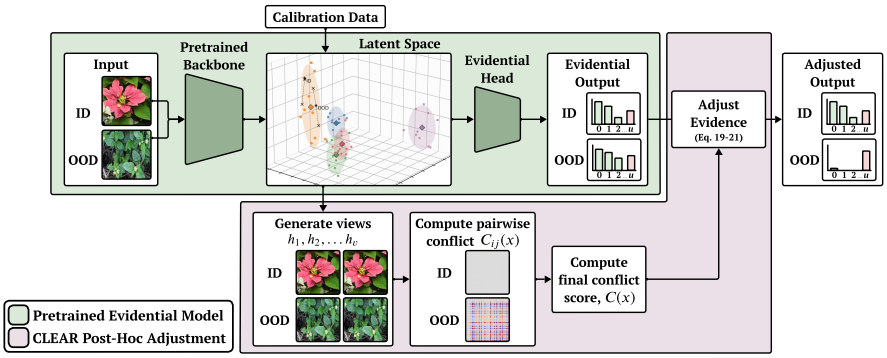


Figure 1: CLEAR architecture overview. A pretrained evidential model produces a latent representation and evidential output. CLEAR uses calibration-derived latent geometry to generate perturbed views, computes pairwise latent conflict, and aggregates this into the final conflict score $C(x)$. This score is used to down-weight unsupported evidential strength post-hoc while preserving the original prediction. The illustrative ID and OOD examples are drawn from Oxford Flowers and DeepWeeds.

maps a target value to its corresponding output bin. These bins are formed using calibration-set quantiles, giving approximately equal group cardinalities N_g.

CLEAR uses these groups to characterise the expected latent structure across different regions of the output space. For each group g, the group-conditioned latent mean and covariance are estimated as:

$$\mu_g = \frac{1}{N_g} \sum_{i:g_i=g} h_i, \qquad \Sigma_g = \mathrm{Cov}(h_i | g_i = g) \tag{8}$$

Because the covariance is subsequently used to normalise latent representations, it is regularised to ensure stable inversion:

$$\widetilde{\Sigma}_g = \Sigma_g + \lambda_g I \tag{9}$$

where $\lambda_g > 0$ is a small stabilisation term. CLEAR then computes its Cholesky factor as $L_g L_g^\top = \widetilde{\Sigma}_g$. The factor L_g provides an efficient transformation for expressing latent representations relative to the covariance structure of group g, which is used during inference to assess whether a prediction behaves consistently with that group. CLEAR also derives a characteristic latent scale for each group:

$$\rho_g = \sqrt{\frac{d}{\left\| L_g^{-1} \right\|_F^2}} \tag{10}$$

where d is the dimensionality of the latent representation and $\| \cdot \|_F$ denotes the Frobenius norm. This scale adapts the magnitude of the latent perturbations used at inference to each group's calibrated geometry. Together, μ_g, L_g, and ρ_g describe the latent geometry against which future predictions are evaluated. If an input is assigned strong evidence for group g, its local latent behaviour is expected to remain compatible with this calibrated structure.

4.2 Latent Perturbation Views

At inference time, the frozen evidential model produces its original evidential output and a latent representation $h = h(x)$ from the penultimate layer. As the final feature representation used by the evidential head, this layer captures information relevant to the model's prediction while retaining the feature structure needed to assess whether that prediction is supported by familiar ID data. CLEAR then determines the output group $\hat{g}$ associated with the base prediction.

CLEAR assesses whether this prediction is supported by examining the local behaviour around its latent representation. Using the characteristic scale of the predicted group, it samples V perturbed latent views:

$$h_v = h + \tau \rho_{\hat{g}} \eta_v, \qquad \eta_v \sim \mathcal{N}(0, I), \qquad v = 1, \ldots, V \tag{11}$$

where $\tau > 0$ controls the overall perturbation magnitude and $\rho_{\hat{g}}$ adapts this magnitude to the calibrated latent scale of the predicted group.

Rather than evaluating individual perturbations independently, CLEAR compares their behaviour with one another. It therefore forms all unordered pairs of latent views:

$$\mathcal{P}_V = (i, j) : 1 \leq i < j \leq V \tag{12}$$

For a well-supported prediction, these neighbouring views should remain mutually compatible when interpreted relative to the calibrated latent geometry of the predicted group. A poorly supported prediction is expected to exhibit greater disagreement under the same perturbations. To measure this compatibility relative to the predicted group rather than in the original latent coordinate system, each perturbation view is centred and whitened (Ermolov et al., 2021) using its calibrated statistics:

$$q_v = L_{\hat{g}}^{-1} \left(h_v - \mu_{\hat{g}} \right), \qquad v = 1, \ldots, V \tag{13}$$

The whitened views are then stacked $Q = [q_1, \ldots, q_V]^\top \in \mathbb{R}^{V \times d}$. CLEAR then measures the off-diagonal interactions between distinct latent dimensions across each unordered pair of views. For $(i, j) \in \mathcal{P}_V$, the pairwise conflict is:

$$C_{ij}(x) = \frac{1}{d(d-1)} \sum_{\substack{a,b \in \{1,\ldots,d\} \\ a \neq b}} q_{i,a}^2 q_{j,b}^2. \tag{14}$$

Averaging across all view pairs gives the raw conflict score:

$$C_{\text{raw}}(x) = \frac{1}{|\mathcal{P}_V|} \sum_{(i,j) \in \mathcal{P}_V} C_{ij}(x) \tag{15}$$

Large values indicate stronger cross-dimensional interactions after whitening with respect to the predicted group's latent geometry. CLEAR interprets this as latent conflict: the evidential head associates the input with group $\hat{g}$, but its local latent behaviour is inconsistent with the structure calibrated for that group.

A single global reference conflict level r is estimated using the held-out calibration set. For each calibration example, CLEAR repeatedly samples the latent perturbation views and computes the corresponding raw conflict score. The resulting scores are averaged across the repeated draws to obtain a stable conflict estimate for each example:

$$\overline{C}_{\text{raw}}(x_i) = \frac{1}{T} \sum_{t=1}^{T} C_{\text{raw}}^{(t)}(x_i) \tag{16}$$

where T denotes the number of repeated perturbation draws. Repeating the stochastic perturbations reduces variation in the reference estimate for each calibration example. The overall reference conflict is then defined as the median across the calibration set:

$$r = \text{median}_{x_i \in \mathcal{D}_{\text{cal}}} \overline{C}_{\text{raw}}(x_i) \tag{17}$$

The raw conflict score is converted into a bounded value:

$$C(x) = \max \left(0, \frac{C_{\text{raw}}(x) - r}{C_{\text{raw}}(x) + r} \right) \tag{18}$$

where $r > 0$ is the median reference conflict estimated from the calibration set. This mapping leaves low-conflict inputs unchanged and smoothly increases towards one as the raw conflict grows. The corresponding evidence scaling factor is:

$$s(x) = \exp(-\delta C(x)) \tag{19}$$

where $\delta > 0$ controls the strength of evidence down-weighting. Thus, $s(x) \approx 1$ for latent-consistent inputs, while high-conflict inputs receive reduced evidential strength.

CLEAR uses the same latent conflict mechanism for classification and regression; only the final adjustment to the evidential output differs. In both cases, the objective is to reduce unsupported evidential strength without replacing the prediction produced by the frozen base model.

For classification, the original EDL model outputs evidence $e(x)$ and Dirichlet parameters $\alpha(x) = e(x) + \mathbf{1}$. CLEAR rescales the evidence directly $e'(x) = s(x)e(x)$ and forms the adjusted Dirichlet parameters $\alpha'(x) = e'(x) + \mathbf{1}$. The resulting predictive probabilities are then:

$$\hat{p}'_k(x) = \frac{\alpha'_k(x)}{\sum_j \alpha'_j(x)}. \tag{20}$$

When the latent conflict is low, $s(x)$ is close to one, and the original evidential prediction is preserved. When conflict is high, the Dirichlet concentration is reduced, increasing uncertainty without requiring any change to the classifier.

For regression, the base DER model outputs NIG parameters $(\gamma(x), \nu(x), \alpha(x), \beta(x))$. CLEAR preserves the predictive mean $\gamma(x)$ and scale parameter $\beta(x)$, and down-weights the evidence-related parameters:

$$\nu'(x) = s(x)\nu(x), \qquad \alpha'(x) = 1 + s(x)(\alpha(x) - 1), \tag{21}$$

The transformation keeps the NIG distribution valid because $\nu' > 0$, $\alpha' > 1$, and $\beta > 0$. It also ensures that CLEAR does not move the point prediction. Instead, it reduces evidential strength when the predicted output region is not supported by the latent representation.

5 Results

We evaluate CLEAR in a comprehensive series of experiments comparing it against state-of-the-art EDL-based and other competitive UQ approaches over 10 independent runs. Our evaluation focuses on both performance and uncertainty estimates produced per approach for OOD and adversarially attacked data. The CLEAR code and replication package are available at our open-source repository [1].

Comparative Approaches. We compare CLEAR against Posterior Networks (Charpentier et al., 2020), Evidential Deep Learning (EDL) (Sensoy et al., 2018), Fisher Information-based EDL ($\mathcal{I}$-EDL) (Deng et al., 2023), Hyper-Opinion EDL (H-EDL) (Qu et al., 2024), Relaxed EDL (R-EDL) (Chen et al., 2024), Regularized Evidential Model (RED) (Pandey & Yu, 2023), Smoothed EDL (S-EDL) (Kopetzki et al., 2021), and Conflict-aware Evidential Deep Learning (C-EDL) (Barker et al., 2026) to represent a range of approaches in EDL and UQ that allow for a fair comparison.

Datasets. Adopting the procedure on EDL-based evaluation from recent research (Deng et al., 2023; Chen et al., 2024), 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 (Nilsback & Zisserman, 2008), Deep Weeds (Olsen et al., 2019), ImageNet (Deng et al., 2009), and CUB (Welinder et al., 2010) 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).

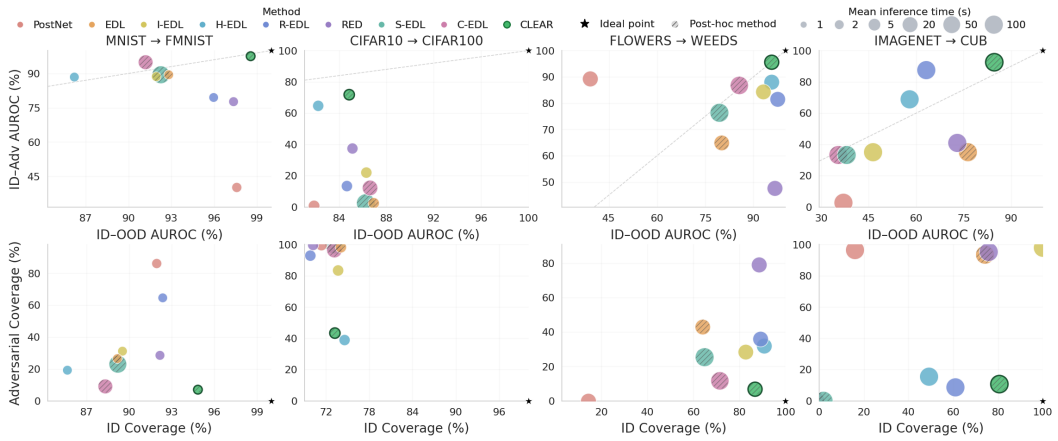


Figure 2: Comparison of nine uncertainty-estimation methods across MNIST→FashionMNIST, CIFAR10→CIFAR100, Oxford Flowers→Deep Weeds, and ImageNet→CUB. Rows show ID–OOD vs ID–Adv. AUROC (top) and ID vs Adv. coverage (bottom). Marker size denotes mean inference time, and the black star marks the ideal operating point.

[1] Our open-source repository is available at: `https://anonymous.4open.science/r/CLEARrepo/README.md`

Table 1: Mean accuracy, OOD detection, and adversarial attack detection performance with standard deviation (sd) of the comparative approaches with a variety of ID and OOD datasets in order of dataset difficulty. The adversarial attack is an L_2-PGD attack. For compactness, we denote adversarial AUROC as AAUROC. Highlighted cells denote the best performance for each metric. * indicates datasets classed as Near-OOD. Isolated best values arising from degenerate behaviour, such as near-universal acceptance/rejection or severely degraded ID accuracy, are not highlighted.

	PostNet	EDL	I-EDL	H-EDL	R-EDL	RED	S-EDL	C-EDL	CLEAR
Type	In-Training						Post-Hoc		
MNIST → FashionMNIST									
ID Acc (%) ↑	98.96 ± 0.11	99.26 ± 0.07	99.20 ± 0.07	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	99.26 ± 0.07	**99.28 ± 0.04**	99.26 ± 0.07
ID Cov (%) ↑	91.93 ± 1.89	89.16 ± 2.73	89.49 ± 3.06	85.63 ± 5.84	92.33 ± 2.10	92.15 ± 1.48	89.19 ± 2.67	88.29 ± 3.54	**94.83 ± 0.85**
OOD Cov (%) ↓	5.14 ± 2.63	12.57 ± 8.80	13.20 ± 10.46	21.84 ± 16.94	7.83 ± 5.99	3.64 ± 1.03	12.68 ± 8.93	16.60 ± 9.68	**1.84 ± 0.91**
Adv Cov (%) ↓	86.16 ± 10.82	26.73 ± 13.70	31.36 ± 16.33	19.44 ± 14.92	64.58 ± 12.73	28.79 ± 9.32	23.05 ± 13.12	9.20 ± 7.30	**7.21 ± 3.90**
AUROC (%) ↑	97.55 ± 1.12	92.54 ± 5.70	91.94 ± 7.22	86.18 ± 12.94	95.95 ± 3.22	97.34 ± 0.85	92.52 ± 5.73	91.20 ± 6.14	**98.54 ± 0.36**
AAUROC (%) ↑	40.19 ± 12.45	88.78 ± 5.29	88.75 ± 4.11	88.59 ± 10.79	79.67 ± 3.17	77.72 ± 8.60	90.23 ± 4.71	94.91 ± 2.13	**97.66 ± 0.53**
MNIST → KMNIST									
ID Acc (%) ↑	98.96 ± 0.11	99.26 ± 0.07	99.20 ± 0.07	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	99.26 ± 0.07	**99.28 ± 0.04**	99.26 ± 0.07
ID Cov (%) ↑	93.38 ± 0.89	93.93 ± 1.16	93.70 ± 0.95	91.40 ± 1.94	92.63 ± 0.83	94.50 ± 1.04	94.16 ± 0.94	93.55 ± 0.83	**95.29 ± 0.39**
OOD Cov (%) ↓	3.97 ± 0.68	3.58 ± 1.11	3.06 ± 0.90	8.00 ± 2.55	3.15 ± 0.87	7.26 ± 1.28	3.76 ± 1.06	4.12 ± 1.10	**2.07 ± 0.66**
Adv Cov (%) ↓	83.34 ± 6.22	30.08 ± 6.78	30.82 ± 4.52	**4.69 ± 2.12**	56.63 ± 7.00	51.92 ± 15.06	26.41 ± 5.86	6.71 ± 2.47	13.76 ± 3.47
AUROC (%) ↑	98.41 ± 0.26	98.73 ± 0.25	98.79 ± 0.23	97.18 ± 0.98	98.40 ± 0.20	97.16 ± 0.54	98.73 ± 0.25	98.58 ± 0.33	**99.16 ± 0.10**
AAUROC (%) ↑	54.48 ± 7.98	93.63 ± 1.57	93.37 ± 1.42	**98.04 ± 0.73**	87.54 ± 2.05	76.14 ± 5.72	94.53 ± 1.33	97.51 ± 0.59	97.36 ± 0.39
MNIST → EMNIST*									
ID Acc (%) ↑	98.96 ± 0.11	**99.26 ± 0.07**	99.20 ± 0.07	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	**99.26 ± 0.07**	99.25 ± 0.06	**99.26 ± 0.07**
ID Cov (%) ↑	91.32 ± 0.47	90.95 ± 0.99	91.26 ± 0.95	87.01 ± 2.88	91.04 ± 0.94	90.51 ± 1.59	91.21 ± 0.83	91.18 ± 0.75	**93.01 ± 1.07**
OOD Cov (%) ↓	11.44 ± 0.72	8.52 ± 0.93	9.68 ± 1.18	13.22 ± 3.26	9.52 ± 0.92	14.81 ± 1.55	8.70 ± 0.85	10.09 ± 0.99	**7.60 ± 0.87**
Adv Cov (%) ↓	67.98 ± 7.06	9.58 ± 4.35	12.17 ± 5.25	7.23 ± 3.14	35.93 ± 4.13	45.63 ± 11.36	7.86 ± 3.70	**3.81 ± 0.86**	3.96 ± 1.50
AUROC (%) ↑	95.25 ± 0.30	**96.47 ± 0.26**	96.12 ± 0.51	94.25 ± 1.32	96.22 ± 0.26	91.95 ± 1.10	**96.47 ± 0.26**	95.98 ± 0.30	97.37 ± 0.15
AAUROC (%) ↑	65.18 ± 7.24	95.53 ± 0.81	95.30 ± 0.97	96.37 ± 1.18	90.61 ± 1.87	71.20 ± 6.79	96.04 ± 0.67	97.55 ± 0.30	**97.63 ± 0.22**
CIFAR10 → SVHN									
ID Acc (%) ↑	89.09 ± 0.26	**89.76 ± 0.19**	88.71 ± 0.24	85.08 ± 0.37	89.28 ± 0.23	88.88 ± 0.30	89.66 ± 0.11	90.11 ± 0.19	**89.76 ± 0.19**
ID Cov (%) ↑	79.25 ± 1.85	82.45 ± 1.09	84.77 ± 1.13	**86.72 ± 0.97**	81.66 ± 1.48	79.82 ± 1.16	75.60 ± 1.55	82.80 ± 1.51	76.55 ± 1.69
OOD Cov (%) ↓	12.99 ± 2.24	9.23 ± 1.30	**6.56 ± 1.06**	9.11 ± 1.93	11.81 ± 1.91	8.18 ± 1.47	10.21 ± 1.51	10.62 ± 1.71	13.80 ± 1.96
Adv Cov (%) ↓	99.92 ± 0.04	99.39 ± 0.11	99.25 ± 0.45	50.04 ± 12.59	95.84 ± 0.71	99.99 ± 0.01	99.33 ± 0.12	99.28 ± 0.11	**36.38 ± 8.39**
AUROC (%) ↑	90.55 ± 1.02	93.18 ± 0.38	94.44 ± 0.31	**95.21 ± 1.05**	92.39 ± 0.59	92.44 ± 0.84	88.94 ± 0.63	93.10 ± 0.44	89.59 ± 0.65
AAUROC (%) ↑	0.14 ± 0.06	1.15 ± 0.16	23.65 ± 1.90	61.62 ± 11.10	15.14 ± 1.60	33.84 ± 1.10	3.72 ± 0.50	10.97 ± 0.68	**77.45 ± 0.42**
CIFAR10 → CIFAR100*									
ID Acc (%) ↑	89.09 ± 0.26	89.76 ± 0.19	88.71 ± 0.24	85.08 ± 0.37	89.28 ± 0.23	88.88 ± 0.30	89.61 ± 0.16	**90.12 ± 0.31**	89.76 ± 0.19
ID Cov (%) ↑	70.78 ± 3.35	73.39 ± 1.42	73.66 ± 1.75	**74.58 ± 2.80**	69.85 ± 3.47	70.80 ± 1.64	73.45 ± 2.65	73.55 ± 1.34	73.22 ± 1.71
OOD Cov (%) ↓	21.38 ± 3.01	**15.58 ± 1.42**	16.50 ± 1.49	24.57 ± 3.00	16.89 ± 3.79	15.66 ± 1.69	15.87 ± 2.66	15.68 ± 1.56	20.08 ± 1.66
Adv Cov (%) ↓	99.47 ± 0.14	98.60 ± 0.15	83.56 ± 3.28	**39.02 ± 8.94**	92.97 ± 0.60	99.95 ± 0.03	98.54 ± 0.15	97.96 ± 0.13	43.43 ± 5.59
AUROC (%) ↑	81.84 ± 0.52	86.55 ± 0.23	86.29 ± 0.28	82.20 ± 0.53	84.63 ± 0.20	85.10 ± 0.24	86.58 ± 0.23	**86.59 ± 0.24**	84.83 ± 0.20
AAUROC (%) ↑	0.64 ± 0.12	2.01 ± 0.16	22.04 ± 1.24	64.68 ± 7.91	13.54 ± 0.84	37.52 ± 0.70	2.83 ± 0.19	12.22 ± 0.30	**71.93 ± 0.45**
Oxford Flowers (low-shot) → Deep Weeds									
ID Acc (%) ↑	5.92 ± 1.49	80.54 ± 0.33	81.09 ± 0.86	74.31 ± 1.22	**85.18 ± 0.85**	83.27 ± 0.33	79.41 ± 0.31	80.63 ± 0.31	80.54 ± 0.32
ID Cov (%) ↑	14.30 ± 9.90	64.00 ± 5.55	82.62 ± 4.11	**90.58 ± 1.89**	89.11 ± 2.28	88.43 ± 2.84	64.63 ± 4.24	71.41 ± 4.69	86.59 ± 1.52
OOD Cov (%) ↓	8.47 ± 7.22	18.78 ± 4.68	8.79 ± 3.91	13.14 ± 2.04	6.24 ± 1.35	7.41 ± 2.65	20.57 ± 4.75	15.97 ± 3.41	**5.09 ± 2.12**
Adv Cov (%) ↓	0.00 ± 0.01	43.15 ± 7.80	28.53 ± 6.32	32.09 ± 5.39	36.13 ± 3.48	79.03 ± 3.29	25.55 ± 6.52	11.80 ± 3.16	**6.77 ± 2.88**
AUROC (%) ↑	39.16 ± 7.07	80.11 ± 4.12	93.07 ± 2.12	95.60 ± 0.93	**97.48 ± 0.79**	96.66 ± 0.89	79.38 ± 4.21	85.47 ± 2.95	95.57 ± 0.62
AAUROC (%) ↑	89.35 ± 2.71	64.91 ± 2.65	84.43 ± 2.34	87.99 ± 2.43	81.49 ± 2.85	47.68 ± 1.76	76.34 ± 2.40	86.90 ± 1.43	**95.60 ± 0.84**
ImageNet → CUB									
ID Acc (%) ↑	37.69 ± 0.25	**79.99 ± 0.03**	79.53 ± 0.01	78.71 ± 0.00	77.14 ± 0.03	79.92 ± 0.03	75.60 ± 0.02	76.09 ± 0.05	75.86 ± 0.04
ID Cov (%) ↑	15.97 ± 1.31	74.02 ± 1.98	**99.97 ± 0.04**	49.05 ± 0.00	60.93 ± 0.12	75.68 ± 1.48	0.03 ± 0.02	0.03 ± 0.02	80.54 ± 0.39
OOD Cov (%) ↓	11.55 ± 2.13	33.68 ± 2.42	99.93 ± 0.03	31.69 ± 0.00	41.51 ± 0.11	36.49 ± 2.01	0.17 ± 0.11	0.12 ± 0.06	**23.67 ± 0.55**
Adv Cov (%) ↓	96.61 ± 0.37	93.50 ± 0.44	98.02 ± 0.88	15.48 ± 0.22	**8.84 ± 0.19**	95.42 ± 1.82	0.34 ± 0.18	0.14 ± 0.07	10.84 ± 0.36
AUROC (%) ↑	36.97 ± 1.41	76.34 ± 0.45	46.26 ± 0.05	57.87 ± 0.00	63.05 ± 0.06	72.87 ± 0.29	36.73 ± 0.21	36.52 ± 0.21	**84.63 ± 0.02**
AAUROC (%) ↑	2.74 ± 0.30	35.28 ± 0.26	35.08 ± 0.10	68.98 ± 0.05	87.60 ± 0.08	41.04 ± 0.13	32.89 ± 0.21	33.82 ± 0.26	**92.61 ± 0.03**

5.1 Core Results

Table 1 shows that CLEAR provides consistently strong uncertainty performance across the core classification benchmarks while preserving the predictive performance of the underlying evidential model. This behaviour is also reflected in Figure 2, where CLEAR generally occupies the desirable operating region across both OOD and adversarial objectives. On MNIST→FashionMNIST, CLEAR achieves the highest ID coverage (94.83%), lowest OOD and adversarial coverage (1.84% and 7.21%), and highest AUROC and AAUROC (98.54% and 97.66%). This performance largely persists across the remaining MNIST settings: on MNIST→KMNIST, CLEAR achieves the highest ID coverage, lowest OOD coverage, and highest AUROC, while retaining an AAUROC of 97.36%; on the near-OOD MNIST→EMNIST setting, it again achieves the strongest ID/OOD coverage together with the highest AUROC and AAUROC.

The advantage of CLEAR becomes particularly apparent on the more challenging CIFAR10 experiments. On CIFAR10→SVHN, CLEAR achieves the lowest adversarial coverage (36.38%) and highest AAUROC (77.45%), while on CIFAR10→CIFAR100 it again achieves the highest AAUROC (71.93%) while remaining competitive in ID retention and OOD detection. Figure 2 shows that CLEAR therefore maintains strong performance across both uncertainty objectives rather than favouring one form of distribution shift. Cases where CLEAR is weaker on an individual metric, including adversarial coverage on MNIST→KMNIST and OOD separation on the CIFAR10 benchmarks, are examined in more detail in Appendix D.

This pattern continues on the larger and more difficult benchmarks. On Oxford Flowers→DeepWeeds, CLEAR achieves the lowest non-degenerate OOD and adversarial coverage (5.09% and 6.77%) and

the highest AAUROC (95.60%), while retaining an AUROC of 95.57%. On ImageNet→CUB, a substantially harder setting due to ImageNet's large-scale 1000-class output space and highly diverse visual representation, CLEAR achieves the highest AUROC and AAUROC, at 84.63% and 92.61%, while retaining 80.54% ID coverage. The degenerate operating points observed for several comparative methods on these harder benchmarks are also discussed in Appendix D. Together, these results show that CLEAR remains effective as dataset scale and difficulty increase.

Table 2: Mean accuracy, OOD and Adv. detection performance with standard deviation on ID and OOD regression datasets. Attacks use L_∞-PGD; AAUROC denotes adversarial AUROC. Highlighted cells indicate the best result per metric, and * marks Near-OOD datasets.

	DER	UR-ERN	CLEAR
California Housing → California Housing (Price Extremes)*			
ID MAE ↓	**$25.26k ± 0.61**	$25.33k ± 0.56	**$25.26k ± 0.61**
ID RMSE ↓	**$32.84k ± 0.60**	$32.95k ± 0.66	**$32.84k ± 0.60**
ID Cov (%) ↑	**93.01 ± 5.53**	85.05 ± 28.54	57.18 ± 2.98
OOD Cov (%) ↓	91.60 ± 5.57	85.05 ± 28.54	**26.34 ± 4.11**
Adv Cov (%) ↓	99.90 ± 0.19	89.89 ± 29.96	**1.05 ± 1.26**
AUROC (%) ↑	41.62 ± 3.15	40.36 ± 3.32	**71.01 ± 1.99**
AAUROC (%) ↑	2.48 ± 1.77	2.32 ± 1.63	**89.75 ± 2.42**
Wine Quality (White Wine) → Wine Quality (Red Wine)*			
ID MAE ↓	**0.56 ± 0.01**	**0.56 ± 0.01**	**0.56 ± 0.01**
ID RMSE ↓	**0.72 ± 0.01**	**0.72 ± 0.01**	**0.72 ± 0.01**
ID Cov (%) ↑	82.30 ± 10.69	**85.37 ± 10.76**	77.54 ± 4.95
OOD Cov (%) ↓	68.17 ± 17.13	70.67 ± 19.92	**14.75 ± 3.76**
Adv Cov (%) ↓	68.17 ± 17.13	80.79 ± 15.13	**12.25 ± 3.64**
AUROC (%) ↑	54.55 ± 10.86	54.76 ± 10.80	**89.10 ± 1.45**
AAUROC (%) ↑	41.40 ± 10.13	42.01 ± 10.57	**89.99 ± 1.31**
Airfoil Self-Noise → Airfoil Self-Noise (Covariant Extremes- Frequency)*			
ID MAE ↓	4.12 ± 0.80	**3.92 ± 0.98**	4.12 ± 0.80
ID RMSE ↓	4.88 ± 0.78	**4.74 ± 0.94**	4.88 ± 0.78
ID Cov (%) ↑	0.00 ± 0.00	0.00 ± 0.00	**72.87 ± 5.79**
OOD Cov (%) ↓	0.00 ± 0.00	0.00 ± 0.00	**20.46 ± 5.84**
Adv Cov (%) ↓	0.00 ± 0.00	0.00 ± 0.00	**0.62 ± 1.85**
AUROC (%) ↑	14.91 ± 3.35	15.15 ± 3.46	**82.81 ± 3.78**
AAUROC (%) ↑	0.54 ± 0.38	0.56 ± 0.41	**94.79 ± 2.21**
UTKFace → Oxford Flowers			
ID MAE ↓	7.91 ± 0.14	**7.89 ± 0.14**	7.99 ± 0.18
ID RMSE ↓	10.41 ± 0.23	**10.38 ± 0.19**	10.48 ± 0.25
ID Cov (%) ↑	**81.82 ± 12.89**	81.08 ± 10.14	76.81 ± 3.05
OOD Cov (%) ↓	42.42 ± 12.72	41.55 ± 9.49	**14.98 ± 3.23**
Adv Cov (%) ↓	49.82 ± 12.68	49.31 ± 10.53	**18.09 ± 3.33**
AUROC (%) ↑	73.99 ± 9.06	74.57 ± 8.15	**87.53 ± 3.15**
AAUROC (%) ↑	64.38 ± 11.62	64.86 ± 10.35	**85.38 ± 3.53**
UTKFace → UTKFace (Age Extremes)*			
ID MAE ↓	**7.98 ± 0.13**	8.00 ± 0.18	**7.98 ± 0.13**
ID RMSE ↓	**10.47 ± 0.17**	10.50 ± 0.25	**10.47 ± 0.17**
ID Cov (%) ↑	56.81 ± 10.16	**57.66 ± 7.94**	53.15 ± 4.52
OOD Cov (%) ↓	39.74 ± 8.91	41.01 ± 8.47	**37.31 ± 3.20**
Adv Cov (%) ↓	61.32 ± 5.88	62.59 ± 5.28	**44.01 ± 2.91**
AUROC (%) ↑	**61.63 ± 1.95**	61.31 ± 1.91	60.77 ± 1.88
AAUROC (%) ↑	45.64 ± 2.88	45.01 ± 2.99	**53.32 ± 2.41**

Overall, CLEAR either achieves the strongest result or remains competitive across ID retention, OOD detection, and adversarial detection, without consistently sacrificing one objective for another.

5.2 Regression Results

Across the regression benchmarks, CLEAR consistently improves separation between ID predictions and both OOD and adversarial inputs while preserving comparable predictive error. On California Housing, CLEAR reduces OOD coverage from 91.60% with DER to 26.34% and adversarial coverage from 99.90% to 1.05%, while increasing AUROC and AAUROC to 71.01% and 89.75%, respectively. This is accompanied by lower ID coverage (57.18%), a trade-off also observed more mildly on several other regression benchmarks and discussed further in Appendix D. Importantly, the underlying regression prediction is unchanged, so the stronger rejection does not correspond to degraded predictive accuracy. Similar improvements are observed on Wine Quality, where CLEAR achieves an AUROC of 89.10% and AAUROC of 89.99% without changing MAE or RMSE.

The Airfoil experiment provides an especially clear example under covariate shift, where OOD samples are defined by extreme values of the input frequency variable rather than by extremes of the regression target. This makes the shift difficult for uncertainty estimates that do not adequately capture changes in the input representation: DER and UR-ERN retain 0% of both ID and OOD samples under the selected threshold, indicating little useful separation between the two distributions. In contrast, CLEAR retains 72.87% ID coverage while reducing OOD and adversarial coverage to 20.46% and 0.62%, respectively, with an AAUROC of 94.79%. This suggests that CLEAR's latent-space conflict is sensitive to changes in the covariate distribution even when the prediction target itself is not the variable used to define the shift.

This behaviour remains consistent on the UTKFace experiments. For UTKFace→Oxford Flowers, CLEAR reduces OOD and adversarial coverage to 14.98% and 18.09%, while increasing AUROC and AAUROC to 87.53% and 85.38%. The subtler UTKFace age-extremes shift is more competitive, because the OOD samples remain within the same face domain and differ only in their target distribution, resulting in less distinct separation between their latent representations and those of ID samples. Consequently, CLEAR achieves an AUROC of 60.77% remaining close to the best value of 61.63%, while still improving adversarial coverage to 44.01% and AAUROC to 53.32%. Overall, these results show that CLEAR generalises its uncertainty improvements beyond classification, maintaining strong predictive performance while improving robustness to both distribution shift and adversarial perturbation.

5.3 Object Detection Results

CLEAR also generalises effectively to object detection, where uncertainty must reflect not only whether a detection is reliable but also the quality of its localisation. As shown in Table 7, Appendix C.6, CLEAR improves every evaluated uncertainty metric over EMM on clean CelebA detections, increasing AUROC from 50.3% to 58.5% and AUPRC from 12.5% to 21.1%, while reducing FPR95 from 90.7% to 82.4% and ECE from 84.4% to 75.0%. Importantly, the correlation between uncertainty and localisation error increases from approximately zero ($r = 0.003$) for EMM to $r = 0.168$ for CLEAR, indicating that poorly localised detections are more consistently assigned greater uncertainty. This is important for object detection, where a confident detection does not necessarily imply that its predicted bounding box is accurately localised.

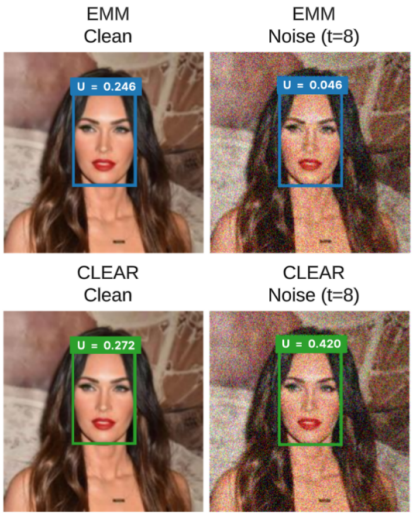


Figure 3: Qualitative object detection comparison of EMM and CLEAR uncertainty under a covariate Gaussian noise on CelebA.

This localisation-aware behaviour persists under covariate corruption. Table 8, Appendix C.6, shows that across brightness, blur, noise, compression, and resize/crop shifts, CLEAR produces a stronger positive association between predictive uncertainty and localisation error than EMM in every case, increasing the overall Pearson correlation from 0.010 to 0.049. Although the absolute correlations remain modest, the consistent improvement across all five shifts indicates that CLEAR better reflects degradation in bounding-box quality as the input distribution changes. This trend is further illustrated in Figure 9, where CLEAR maintains a stronger uncertainty-localisation relationship across the evaluated corruptions. Together, these results demonstrate that CLEAR's latent-conflict mechanism transfers beyond classification and regression to structured localisation outputs while producing uncertainty that is more informative of detection quality. Full training details for this setup are described in Appendix E.2.

5.4 Further Experiments and Ablation

An extensive set of additional experiments is provided in the Appendix, including corrupted benchmarks (Appendix C.1) and additional uncertainty metrics (Appendix C.2). We then examine CLEAR through hyperparameter ablations (Appendix C.3), uncertainty-threshold analysis (Appendix C.4), adversarial attack analysis (Appendix C.5), and extended object-detection results (Appendix C.6). Finally, we investigate the latent mechanism through held-out validation of the calibrated geometry (Appendix C.7), sensitivity to latent representation depth (Appendix C.8), and qualitative visualisation of the generated latent perturbation views (Appendix C.9). Across these analyses, CLEAR maintains strong uncertainty performance under corruption and increasing attack strength, remains broadly stable across its hyperparameters and thresholding method, and exhibits consistent behaviour across different latent representations.

6 Conclusions

We introduced CLEAR, a lightweight post-hoc method for improving evidential uncertainty without retraining or altering the underlying prediction. Across a broad set of datasets spanning classification, regression, and object detection, CLEAR consistently improves uncertainty quality while preserving predictive performance. A thorough evaluation across diverse OOD settings, covariate shifts, adversarial attacks, and uncertainty thresholds demonstrates that CLEAR remains effective under substantially different forms of distributional and adversarial challenge. Importantly, CLEAR improves complementary uncertainty objectives jointly, avoiding the pronounced trade-offs exhibited by several comparative methods, while introducing only modest additional inference overhead. These results demonstrate that latent conflict provides a practical and broadly applicable signal for identifying unsupported evidential confidence. Future work will investigate evaluating CLEAR on denser structured prediction tasks, such as semantic segmentation, where uncertainty must be estimated over spatially distributed outputs.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

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.

Charmaine Barker, Daniel Bethell, and Simos Gerasimou. Robust adversarial quantification via conflict-aware evidential deep learning. In *International conference on learning representations*, volume 2026, pp. 128886–128919, 2026.

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.

Thomas F Brooks, D Stuart Pope, and Michael A Marcolini. Airfoil self-noise and prediction. Technical report, 1989.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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)*, pp. 2921–2926. IEEE, 2017.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

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

Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *International Conference on Machine Learning*, pp. 7596–7616. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pp. 3015–3024. PMLR, 2021.

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

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*, pp. 45–87, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

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.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pp. 5707–5718. PMLR, 2021.

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.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

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, pp. 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.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: theory and practice. In *International Conference on Machine Learning*, pp. 26963–26989. PMLR, 2023.

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*, pp. 4075–4084. PMLR, 2018.

Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37: 84645–84668, 2024.

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.

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, pp. 9772–9781, 2023.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

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.

Kai Ye, Tiejin Chen, Hua Wei, and Liang Zhan. Uncertainty regularized evidential regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16460–16468, 2024.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

APPENDIX CONTENTS

A Implementation Details **14**

B Theoretical Analysis **15**

B.1 Proposition 1 15

C Additional Experiments **17**

C.1 Corrupted Benchmarks 17

C.2 Extra Metrics Analysis 18

C.3 Ablation Analysis 18

C.4 Threshold Analysis 21

C.5 Adversarial Attack Analysis 22

C.6 Full Object Detection Results 22

C.7 Held-Out Analysis of the Latent Whitening Geometry 24

C.8 Sensitivity to Latent Representation Depth 26

C.9 Qualitative Analysis of Latent Perturbation Views 27

D Additional Discussion **27**

E Experimental Setup **28**

E.1 Datasets 28

E.2 Models 31

E.3 ID-OOD Thresholds 32

E.4 Comparative Approaches 33

E.5 Adversarial Attacks 34

E.6 Training Details 35

E.7 Runtime and Complexity 35

A IMPLEMENTATION DETAILS

Algorithms 1 and 2 summarise the calibration and inference stages of CLEAR. The pretrained evidential model remains frozen and in evaluation mode throughout both stages. Calibration is performed once, after which the group statistics and global reference conflict are reused for subsequent inputs.

During calibration, class labels or target-quantile bins determine the groups used to estimate μ_g, L_g, and ρ_g. Reference conflict is then evaluated relative to the group $\hat{g}$ selected by each example's original prediction, matching the inference procedure. Each calibration example receives T independent perturbation draws, each containing V latent views. The raw conflict scores are averaged per example before taking the median across $\mathcal{D}_{\text{cal}}$ to obtain r.

At inference, a single forward pass produces the latent representation and original evidential output. Each input then receives one set of V latent views, with no additional evaluations of the backbone or evidential head. Whitening is implemented using triangular solves with the stored Cholesky factor $L_{\hat{g}}$. Pairwise conflict is computed using vector reductions, avoiding explicit construction of cross-dimensional interaction matrices. The resulting $C_{\text{raw}}(x)$ is mapped to $C(x)$ and used to obtain the evidence scaling factor $s(x)$. The final adjustment preserves the predicted class in classification and the predictive mean $\gamma(x)$ in regression.

Algorithm 1 CLEAR calibration

Require: Frozen evidential model; calibration set $\mathcal{D}_{\text{cal}}$
Require: Latent views V, perturbation magnitude τ, reference draws T, stabilisation terms λ_g, regression bins B
Ensure: Group statistics $\{\mu_g, L_g, \rho_g\}_g$, reference r, and bin map $b(\cdot)$ for regression
1: **if** regression **then**
2: Construct $b(\cdot)$ using B target-quantile bins from $\mathcal{D}_{\text{cal}}$.
3: **end if**
4: **for all** $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ **do**
5: Obtain $h_i = h(x_i)$ and the original evidential output.
6: $g_i \leftarrow y_i$ for classification, or $g_i \leftarrow b(y_i)$ for regression.
7: **end for**
8: **for all** output groups g **do**
9: $N_g \leftarrow |\{i : g_i = g\}|$.
10: $\mu_g \leftarrow \frac{1}{N_g} \sum_{i:g_i=g} h_i$; $\Sigma_g \leftarrow \text{Cov}(h_i \mid g_i = g)$.
11: $\widetilde{\Sigma}_g \leftarrow \Sigma_g + \lambda_g I$.
12: Compute L_g such that $L_g L_g^\top = \widetilde{\Sigma}_g$.
13: $\rho_g \leftarrow \sqrt{\frac{d}{\|L_g^{-1}\|_F^2}}$.
14: **end for**
15: **for all** $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ **do**
16: $\hat{g} \leftarrow \arg\max_k \hat{p}_k(x_i)$ for classification, or $\hat{g} \leftarrow b(\gamma(x_i))$ for regression.
17: **for** $t = 1, \ldots, T$ **do**
18: Compute $C_{\text{raw}}^{(t)}(x_i)$ using lines 3–13 of Algorithm 2 for x_i, with $h = h_i$ and fresh perturbations.
19: **end for**
20: $\overline{C}_{\text{raw}}(x_i) \leftarrow \frac{1}{T} \sum_{t=1}^{T} C_{\text{raw}}^{(t)}(x_i)$.
21: **end for**
22: $r \leftarrow \text{median}_{x_i \in \mathcal{D}_{\text{cal}}} \overline{C}_{\text{raw}}(x_i)$.
23: **return** $\{\mu_g, L_g, \rho_g\}_g$, r, and $b(\cdot)$ for regression.

B Theoretical Analysis

This appendix provides the theoretical analysis to complement CLEAR presented in Section 4.

B.1 Proposition 1

Proposition 1. CLEAR preserves the base prediction from the underlying model while reducing evidential strength in proportion to the latent conflict. In classification:

$$\arg\max_k \hat{p}'_k(x) = \arg\max_k \hat{p}_k(x), \qquad u'(x) \geq u(x) \tag{22}$$

while in regression:

$$\mathbb{E}'[y] = \mathbb{E}[y] = \gamma(x) u'_{\text{alea}}(x) = \exp(\delta C(x)) u_{\text{alea}}(x), \qquad u'_{\text{epi}}(x) = \exp(2\delta C(x)) u_{\text{epi}}(x) \tag{23}$$

Proof. We first consider the classification setting, where CLEAR rescales the original evidence:

$$e'(x) = s(x) e(x) \tag{24}$$

such that the Dirichlet parameters are also adjusted:

$$\alpha'_k(x) = 1 + s(x) e_k(x) \tag{25}$$

For any two classes j and k:

$$\alpha'_j(x) - \alpha'_k(x) = s(x)(e_j(x) - e_k(x)) \tag{26}$$

and since $s(x) > 0$, multiplication by $s(x)$preserves the ordering of the original evidence terms. Moreover, because:

$$\alpha_j(x) - \alpha_k(x) = e_j(x) - e_k(x) \tag{27}$$

the ordering of the original and adjusted Dirichlet parameters is identical. The predictive probabilities in both cases are obtained by normalising these parameters by a positive quantity. Therefore:

$$\arg\max_k \hat{p}'_k(x) = \arg\max_k \hat{p}_k(x) \tag{28}$$

Algorithm 2 CLEAR inference

Require: Frozen evidential model; input x; calibrated $\{\mu_g, L_g, \rho_g\}_g$ and r
Require: Latent views V, perturbation magnitude τ, adjustment strength δ; bin map $b(\cdot)$ for regression
Ensure: Adjusted evidential output

Obtain $h = h(x)$ and the original evidential output in one forward pass.
$\hat{g} \leftarrow \arg\max_k \hat{p}_k(x)$ for classification, or $\hat{g} \leftarrow b(\gamma(x))$ for regression.
for $v = 1, \ldots, V$ **do**
    Sample $\eta_v \sim \mathcal{N}(0, I)$ independently.
    $h_v \leftarrow h + \tau \rho_{\hat{g}} \eta_v$.
    $q_v \leftarrow L_{\hat{g}}^{-1}(h_v - \mu_{\hat{g}})$.
end for
$Q \leftarrow [q_1, \ldots, q_V]^\top \in \mathbb{R}^{V \times d}$.
$\mathcal{P}_V \leftarrow \{(i, j) : 1 \le i < j \le V\}$.
for all $(i, j) \in \mathcal{P}_V$ **do**
    $C_{ij}(x) \leftarrow \frac{1}{d(d-1)} \sum_{a \neq b} q_{i,a}^2 q_{j,b}^2$.
end for
$C_{\mathrm{raw}}(x) \leftarrow \frac{1}{|\mathcal{P}_V|} \sum_{(i,j) \in \mathcal{P}_V} C_{ij}(x)$.
$C(x) \leftarrow \max\left(0, \frac{C_{\mathrm{raw}}(x) - r}{C_{\mathrm{raw}}(x) + r}\right)$.
$s(x) \leftarrow \exp(-\delta C(x))$.
if classification **then**
    $e'(x) \leftarrow s(x) e(x)$.
    $\alpha'(x) \leftarrow e'(x) + 1$.
    return $\alpha'(x)$.
else
    $\nu'(x) \leftarrow s(x) \nu(x)$.
    $\alpha'(x) \leftarrow 1 + s(x)(\alpha(x) - 1)$.
    return $(\gamma(x), \nu'(x), \alpha'(x), \beta(x))$.
end if

Thus, the CLEAR adjustment cannot change the predicted class. Given the original total evidential strength is:

$$S(x) = \sum_{k=1}^{K} \alpha_k(x) = K + \sum_{k=1}^{K} e_k(x) \tag{29}$$

and the adjusted total strength is:

$$\begin{aligned} S'(x) &= \sum_{k=1}^{K} \alpha'_k(x) \\ &= K + s(x) \sum_{k=1}^{K} e_k(x) \\ &= K + s(x)(S(x) - K) \end{aligned} \tag{30}$$

Because $0 < s(x) \le 1$ and $S(x) - K \ge 0$:

$$S'(x) \le S(x) \tag{31}$$

Using the scalar uncertainty definition from Equation 3:

$$u'(x) = \frac{K}{S'(x)} = \frac{K}{K + s(x)(S(x) - K)} \tag{32}$$

It follows directly that:

$$u'(x) \ge \frac{K}{S(x)} = u(x) \tag{33}$$

Hence, CLEAR preserves the predicted class while reducing the total evidential strength, therefore increasing, or leaving unchanged, the corresponding scalar uncertainty. For completeness, this exact adjustment can also be expressed as a contraction of the original predictive distribution towards the uniform distribution. Let:

$$\lambda(x) = \frac{s(x) S(x)}{K + s(x)(S(x) - K)} \tag{34}$$

Since, $0 < s(x) \leq 1$, it follows that $0 < \lambda(x) \leq 1$, and the adjusted predictive probability satisfies:

$$\hat{p}'_k(x) = \lambda(x)\hat{p}_k(x) + (1 - \lambda(x))\frac{1}{K} \tag{35}$$

Thus, increasing conflict contracts the predictive probabilities towards the uniform distribution without changing their ordering.

We now consider the regression setting. From Equation 21, CLEAR leaves $\gamma(x)$ and $\beta(x)$ unchanged and applies:

$$\nu'(x) = s(x)\nu(x), \qquad \alpha'(x) = 1 + s(x)(\alpha(x) - 1) \tag{36}$$

Since the predictive mean in Equation 6 is:

$$\mathbb{E}[y] = \gamma(x) \tag{37}$$

and CLEAR does not modify $\gamma(x)$:

$$\mathbb{E}'[y] = \mathbb{E}[y] = \gamma(x) \tag{38}$$

For the aleatoric uncertainty defined in Equation 7:

$$\begin{aligned} u'_{alea}(x) &= \frac{\beta(x)}{\alpha(x) - 1} \\ &= \frac{\beta(x)}{s(x)(\alpha(x) - 1)} \\ &= \frac{1}{s(x)} u_{alea}(x) \end{aligned} \tag{39}$$

Since $s(x) = \exp(-\delta C(x))$, then:

$$u'_{alea}(x) = \exp(\delta C(x)) u_{alea}(x) \tag{40}$$

Similarly, the adjusted epistemic uncertainty is:

$$\begin{aligned} u'_{epi}(x) &= \frac{\beta(x)}{\nu'(x)(\alpha(x) - 1)} \\ &= \frac{\beta(x)}{s(x)\nu(x)s(x)(\alpha(x) - 1)} \\ &= \frac{1}{s(x)^2} u_{epi}(x) \end{aligned} \tag{41}$$

and therefore:

$$u'_{epi}(x) = \exp(2\delta C(x)) u_{epi}(x) \tag{42}$$

Because $C(x) \geq 0$ and $\delta > 0$, both multiplicative factors are at least one. Consequently, CLEAR leaves the regression mean unchanged while monotonically increasing aleatoric and epistemic uncertainty as latent conflict increases. □

C Additional Experiments

C.1 Corrupted Benchmarks

Table 3 evaluates performance under corrupted versions of the ID datasets. On MNIST→MNIST-C, CLEAR achieves the strongest OOD and adversarial detection, with OOD and adversarial coverage of 19.26% and 3.34%, and AUROC and ID-Adv. AUROC of 89.85% and 96.51%, respectively. On CIFAR-10-C, CLEAR remains competitive for corruption detection and provides the strongest adversarial separation at both severity levels, achieving ID-Adv. AUROC of 69.90% at severity 1 and 66.32% at severity 5. Although I-EDL attains the highest ID-OOD AUROC at severity 5, CLEAR consistently gives the lowest adversarial coverage across both severities. The increase in ID-OOD AUROC from severity 1 to severity 5 for most methods also suggests that stronger corruptions become easier to distinguish as they move further from the clean ID distribution, while CLEAR continues to provide further robust separation under adversarial shift.

Table 3: 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 L_2-PGD attack. For compactness, we denote adversarial AUROC as AAUROC. Highlighted cells denote the best performance for each metric. * indicates datasets classed as Near-OOD.

	PostNet	EDL	I-EDL	H-EDL	R-EDL	RED	S-EDL	C-EDL	CLEAR
Type	In-Training						Post-Hoc		
MNIST → MNIST-C*									
ID Acc (%) ↑	98.96 ± 0.11	99.26 ± 0.07	99.20 ± 0.07	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	99.26 ± 0.07	$\mathbf{99.28 \pm 0.04}$	99.26 ± 0.07
ID Cov (%) ↑	84.57 ± 2.09	83.08 ± 1.15	81.24 ± 2.21	77.86 ± 3.37	82.33 ± 2.21	$\mathbf{86.52 \pm 2.64}$	81.77 ± 2.61	81.38 ± 1.93	84.14 ± 1.46
OOD Cov (%) ↓	29.15 ± 2.49	27.46 ± 4.19	26.24 ± 6.49	35.50 ± 5.17	25.75 ± 3.15	29.34 ± 2.66	26.27 ± 4.47	30.35 ± 3.92	$\mathbf{19.26 \pm 1.97}$
Adv Cov (%) ↓	49.64 ± 6.84	12.85 ± 3.91	13.44 ± 5.42	19.44 ± 6.08	23.38 ± 4.62	47.47 ± 6.96	10.23 ± 4.10	5.88 ± 3.45	$\mathbf{3.34 \pm 0.92}$
AUROC (%) ↑	84.11 ± 1.12	83.89 ± 3.67	83.16 ± 4.51	77.01 ± 5.08	84.84 ± 1.43	81.52 ± 0.93	83.89 ± 3.68	82.02 ± 3.52	$\mathbf{89.85 \pm 0.34}$
AAUROC (%) ↑	61.89 ± 5.32	90.49 ± 2.59	90.02 ± 2.63	85.90 ± 5.12	85.47 ± 1.85	65.77 ± 4.61	91.60 ± 2.50	93.99 ± 1.94	$\mathbf{96.51 \pm 0.37}$
CIFAR-10 → CIFAR-10-C (Severity-1)*									
ID Acc (%) ↑	88.89 ± 0.22	89.76 ± 0.19	88.56 ± 0.13	85.05 ± 0.21	89.10 ± 0.20	88.61 ± 0.27	89.64 ± 0.12	$\mathbf{90.11 \pm 0.15}$	89.76 ± 0.19
ID Cov (%) ↑	59.54 ± 5.22	55.26 ± 5.10	55.42 ± 5.01	$\mathbf{63.84 \pm 6.26}$	55.64 ± 4.14	55.83 ± 4.79	58.59 ± 3.40	55.64 ± 7.81	55.69 ± 3.47
OOD Cov (%) ↓	50.35 ± 5.03	$\mathbf{43.85 \pm 4.94}$	43.97 ± 5.10	54.72 ± 6.51	44.63 ± 4.05	44.88 ± 4.79	46.83 ± 3.54	44.44 ± 7.76	44.22 ± 3.53
Adv Cov (%) ↓	90.95 ± 1.12	97.53 ± 0.13	82.10 ± 1.29	30.99 ± 8.19	89.92 ± 1.89	93.20 ± 17.66	97.39 ± 0.15	94.28 ± 1.80	$\mathbf{12.24 \pm 3.45}$
AUROC (%) ↑	56.23 ± 0.19	57.84 ± 0.13	$\mathbf{57.97 \pm 0.16}$	56.29 ± 0.39	57.40 ± 0.14	57.77 ± 0.16	57.93 ± 0.09	57.78 ± 0.14	57.94 ± 0.12
AAUROC (%) ↑	9.19 ± 1.07	2.72 ± 0.13	14.66 ± 0.71	67.91 ± 5.90	11.52 ± 0.51	37.20 ± 0.48	3.28 ± 0.15	13.23 ± 0.31	$\mathbf{69.90 \pm 0.42}$
CIFAR-10 → CIFAR-10-C (Severity-5)*									
ID Acc (%) ↑	88.82 ± 0.11	89.74 ± 0.18	88.59 ± 0.11	85.07 ± 0.23	89.08 ± 0.19	88.59 ± 0.26	89.64 ± 0.14	$\mathbf{90.14 \pm 0.09}$	89.77 ± 0.17
ID Cov (%) ↑	70.85 ± 1.32	71.72 ± 1.84	$\mathbf{75.55 \pm 2.03}$	75.48 ± 1.75	67.72 ± 3.89	70.68 ± 3.30	71.96 ± 2.91	73.02 ± 2.06	65.61 ± 1.34
OOD Cov (%) ↓	33.13 ± 1.14	29.27 ± 1.89	31.77 ± 1.82	35.52 ± 2.03	27.65 ± 3.84	29.73 ± 3.49	29.61 ± 3.00	30.98 ± 2.27	$\mathbf{26.84 \pm 1.58}$
Adv Cov (%) ↓	97.66 ± 0.10	98.53 ± 0.21	86.48 ± 3.30	40.30 ± 7.44	92.04 ± 0.78	99.75 ± 0.09	98.50 ± 0.22	97.61 ± 0.31	$\mathbf{31.63 \pm 3.58}$
AUROC (%) ↑	74.70 ± 0.38	77.85 ± 0.09	$\mathbf{79.05 \pm 0.25}$	76.04 ± 0.88	76.97 ± 0.30	77.18 ± 0.26	77.86 ± 0.08	77.59 ± 0.07	75.94 ± 0.17
AAUROC (%) ↑	2.52 ± 0.09	2.14 ± 0.23	21.18 ± 0.75	63.48 ± 6.86	14.82 ± 1.23	38.00 ± 0.58	3.04 ± 0.29	14.49 ± 0.43	$\mathbf{66.32 \pm 0.53}$

C.2 Extra Metrics Analysis

Table 4 provides additional insight into how CLEAR modifies the uncertainty produced by the underlying evidential model. Across the MNIST and CIFAR settings, CLEAR produces consistently larger confidence reductions for OOD and adversarial inputs, supporting the intended behaviour of the method: latent conflict primarily acts on predictions whose representations are inconsistent with the calibrated ID geometry, rather than uniformly suppressing evidence. This effect is strongest on settings where ID and shifted inputs are well separated in latent space, while the smaller confidence drops observed on the more challenging Oxford Flowers and ImageNet benchmarks suggest that the distinction becomes less pronounced when the underlying representation itself provides weaker separation. NLL and ECE show a different behaviour because these metrics evaluate calibration of the ID predictive distribution rather than separation between ID and unreliable inputs. CLEAR does not explicitly optimise either quantity and, by reducing evidential strength for latent-conflicting samples while preserving the base prediction, can produce more conservative confidence estimates without necessarily improving conventional calibration metrics. The results therefore highlight the distinction between predictive calibration and uncertainty separation: CLEAR is designed to expose unsupported evidence under distribution shift and attack, with its strongest gains appearing in the confidence separation between reliable and unreliable inputs. The additional latent conflict computation also introduces only a modest inference overhead relative to the underlying single-pass models; this computational behaviour is examined further in Section E.7.

C.3 Ablation Analysis

We assess the sensitivity of CLEAR to five quantities controlling its latent conflict mechanism: the number of latent views V, perturbation magnitude τ, evidence down-weighting strength δ, reference conflict quantile q, and, for regression, the number of output bins B. Unless otherwise stated, we use $V = 3$, $\tau = 0.05$, $\delta = 10$, and $q = 0.5$, with all non-ablated parameters held fixed. The first four parameters affect different stages of the same inference process: V determines how many local latent views are sampled, τ controls their displacement from the original representation, q determines the calibration reference against which raw conflict is compared, and δ controls how strongly the resulting conflict reduces evidential strength. The regression-specific parameter B instead determines the granularity of the output-conditioned latent geometry.

Figure 4 examines the perturbation magnitude τ used to generate latent views, $h_v = h + \tau\rho_{\hat{g}}\eta_v$. Performance is remarkably stable across the evaluated range. ID, OOD, and adversarial coverage and both AUROC measures change only marginally for most datasets, despite the substantial change in the absolute perturbation scale. This robustness is consistent with the calibration-relative construction of CLEAR: τ changes the neighbourhood explored around the latent representation, but conflict is subsequently interpreted relative to a reference estimated using the same perturbation mechanism.

Table 4: Mean calibration, confidence-drop, and inference-time performance with standard deviation of the comparative approaches across a variety of ID and OOD datasets in order of dataset difficulty. The adversarial attack is an L_2-PGD attack, and ΔConf. denotes the mean confidence drop from ID to OOD or adversarial inputs. Highlighted cells denote the best performance for each metric. * indicates datasets classed as Near-OOD. Isolated best values arising from degenerate behaviour or severely degraded ID accuracy are not highlighted.

	PostNet	EDL	I-EDL	H-EDL	R-EDL	RED	S-EDL	C-EDL	CLEAR
Type			In-Training					Post-Hoc	
MNIST → FashionMNIST									
NLL ↓	**0.04 ± 0.00**	0.11 ± 0.00	0.16 ± 0.00	1.15 ± 0.12	0.06 ± 0.00	0.06 ± 0.01	0.11 ± 0.00	0.15 ± 0.00	0.58 ± 0.01
ECE (%) ↓	**0.40 ± 0.06**	8.58 ± 0.13	12.80 ± 0.17	66.05 ± 3.34	3.27 ± 0.06	0.74 ± 0.11	8.59 ± 0.13	11.97 ± 0.12	30.38 ± 0.33
ΔConf. ID-OOD (%) ↑	26.31 ± 4.68	24.72 ± 4.72	25.87 ± 5.65	7.14 ± 2.11	24.94 ± 4.51	35.12 ± 10.01	24.72 ± 4.73	29.43 ± 5.73	**58.84 ± 0.33**
ΔConf. ID-Adv. (%) ↑	0.30 ± 1.36	18.13 ± 4.73	16.76 ± 2.78	8.64 ± 1.85	5.61 ± 2.38	7.20 ± 4.53	18.85 ± 4.73	33.27 ± 3.59	**58.76 ± 0.36**
Inf. Time (s) ↓	2.21 ± 0.06	1.96 ± 0.02	1.96 ± 0.02	1.96 ± 0.04	**1.94 ± 0.03**	2.02 ± 0.03	57.56 ± 0.09	13.15 ± 0.11	2.36 ± 0.15
MNIST → KMNIST									
NLL ↓	**0.04 ± 0.00**	0.11 ± 0.00	0.16 ± 0.00	1.15 ± 0.12	0.06 ± 0.00	0.06 ± 0.01	0.11 ± 0.00	0.15 ± 0.00	0.58 ± 0.01
ECE (%) ↓	**0.40 ± 0.06**	8.58 ± 0.13	12.80 ± 0.17	66.05 ± 3.34	3.27 ± 0.06	0.74 ± 0.11	8.59 ± 0.13	11.97 ± 0.12	30.38 ± 0.33
ΔConf. ID-OOD (%) ↑	35.33 ± 1.07	33.16 ± 1.08	37.10 ± 0.97	8.92 ± 1.18	33.36 ± 1.49	52.93 ± 4.82	33.17 ± 1.08	40.75 ± 1.27	**58.84 ± 0.33**
ΔConf. ID-Adv. (%) ↑	0.63 ± 0.47	17.38 ± 1.84	16.69 ± 2.31	9.83 ± 1.30	3.81 ± 0.38	4.13 ± 2.88	18.04 ± 1.82	34.47 ± 1.86	**58.59 ± 0.37**
Inf. Time (s) ↓	2.14 ± 0.04	1.94 ± 0.03	**1.92 ± 0.03**	1.98 ± 0.06	1.97 ± 0.03	1.99 ± 0.02	57.73 ± 0.62	12.94 ± 0.17	2.22 ± 0.02
MNIST → EMNIST*									
NLL ↓	**0.04 ± 0.00**	0.11 ± 0.00	0.16 ± 0.00	1.15 ± 0.12	0.06 ± 0.00	0.06 ± 0.01	0.11 ± 0.00	0.15 ± 0.00	0.58 ± 0.01
ECE (%) ↓	**0.40 ± 0.06**	8.58 ± 0.13	12.80 ± 0.17	66.05 ± 3.34	3.27 ± 0.06	0.74 ± 0.11	8.60 ± 0.12	11.93 ± 0.13	30.38 ± 0.32
ΔConf. ID-OOD (%) ↑	26.51 ± 0.63	26.94 ± 0.44	29.13 ± 0.63	7.80 ± 1.03	24.67 ± 0.62	37.29 ± 2.34	26.95 ± 0.44	33.86 ± 0.65	**57.36 ± 0.29**
ΔConf. ID-Adv. (%) ↑	2.33 ± 0.72	18.81 ± 1.63	18.98 ± 2.04	9.35 ± 1.15	6.46 ± 0.56	4.07 ± 1.73	19.47 ± 1.61	34.62 ± 1.45	**58.73 ± 0.32**
Inf. Time (s) ↓	2.95 ± 0.05	2.73 ± 0.18	2.51 ± 0.02	**2.50 ± 0.03**	2.55 ± 0.04	2.57 ± 0.04	83.96 ± 1.80	18.52 ± 0.45	2.96 ± 0.13
CIFAR10 → SVHN									
NLL ↓	**0.36 ± 0.01**	0.42 ± 0.00	0.47 ± 0.01	1.03 ± 0.11	0.40 ± 0.01	0.40 ± 0.01	0.42 ± 0.00	0.44 ± 0.00	0.99 ± 0.01
ECE (%) ↓	3.62 ± 0.22	5.66 ± 0.17	13.84 ± 0.38	45.51 ± 5.04	**3.29 ± 0.16**	4.71 ± 0.25	5.79 ± 0.12	11.19 ± 0.19	33.46 ± 0.37
ΔConf. ID-OOD (%) ↑	30.53 ± 2.54	29.80 ± 0.88	44.15 ± 0.66	11.73 ± 1.26	28.35 ± 1.14	32.16 ± 1.94	23.53 ± 0.84	33.88 ± 0.88	**44.17 ± 0.43**
ΔConf. ID-Adv. (%) ↑	−7.40 ± 0.27	−9.99 ± 0.30	0.82 ± 1.75	9.95 ± 1.65	2.17 ± 1.23	15.72 ± 0.71	−8.69 ± 0.37	−11.43 ± 0.40	**42.83 ± 0.33**
Inf. Time (s) ↓	6.07 ± 0.11	5.39 ± 0.15	6.52 ± 3.90	5.51 ± 0.18	**5.05 ± 0.09**	5.25 ± 0.24	181.17 ± 2.29	31.59 ± 0.22	6.96 ± 0.04
CIFAR10 → CIFAR100*									
NLL ↓	**0.36 ± 0.01**	0.42 ± 0.00	0.47 ± 0.01	1.03 ± 0.11	0.40 ± 0.01	0.40 ± 0.01	0.42 ± 0.00	0.43 ± 0.00	0.99 ± 0.01
ECE (%) ↓	3.62 ± 0.22	5.66 ± 0.17	13.84 ± 0.38	45.51 ± 5.04	**3.29 ± 0.16**	4.71 ± 0.25	5.77 ± 0.18	11.23 ± 0.29	33.46 ± 0.37
ΔConf. ID-OOD (%) ↑	20.35 ± 0.36	20.33 ± 0.23	30.44 ± 0.50	7.53 ± 0.69	19.95 ± 0.24	19.75 ± 0.18	20.58 ± 0.27	25.03 ± 0.28	**39.63 ± 0.39**
ΔConf. ID-Adv. (%) ↑	−7.02 ± 0.29	−9.95 ± 0.14	−0.57 ± 1.06	10.39 ± 1.43	−0.27 ± 0.57	8.07 ± 0.73	−9.60 ± 0.18	−11.42 ± 0.16	**36.78 ± 0.61**
Inf. Time (s) ↓	3.69 ± 0.19	3.29 ± 0.06	3.24 ± 0.29	3.16 ± 0.11	3.24 ± 0.09	**3.10 ± 0.09**	103.78 ± 2.37	18.39 ± 0.23	4.14 ± 0.10
Oxford Flowers (low-shot) → Deep Weeds									
NLL ↓	4.48 ± 0.04	2.72 ± 0.00	3.07 ± 0.01	4.32 ± 0.00	1.65 ± 0.02	**0.89 ± 0.01**	2.74 ± 0.01	3.09 ± 0.01	3.42 ± 0.01
ECE (%) ↓	4.03 ± 1.46	71.93 ± 0.36	75.90 ± 0.81	72.95 ± 1.22	56.97 ± 0.91	**22.84 ± 0.23**	70.92 ± 0.34	74.75 ± 0.34	75.14 ± 0.32
ΔConf. ID-OOD (%) ↑	−0.26 ± 0.20	4.53 ± 0.29	2.67 ± 0.14	0.12 ± 0.01	20.38 ± 0.56	**47.12 ± 0.90**	4.38 ± 0.28	3.36 ± 0.16	4.40 ± 0.03
ΔConf. ID-Adv. (%) ↑	0.87 ± 0.14	3.93 ± 0.20	2.58 ± 0.08	0.15 ± 0.01	**10.88 ± 0.83**	0.88 ± 1.83	4.34 ± 0.19	3.56 ± 0.09	4.40 ± 0.03
Inf. Time (s) ↓	20.62 ± 2.95	18.48 ± 0.72	**18.09 ± 1.22**	19.17 ± 4.19	18.93 ± 2.45	20.00 ± 0.49	660.19 ± 0.23	74.46 ± 0.31	18.69 ± 0.45
ImageNet → CUB									
NLL ↓	3.35 ± 0.02	0.99 ± 0.00	4.91 ± 0.00	6.47 ± 0.00	4.36 ± 0.00	**0.90 ± 0.00**	6.66 ± 0.00	6.72 ± 0.00	6.77 ± 0.00
ECE (%) ↓	10.20 ± 0.25	4.56 ± 0.05	78.69 ± 0.01	78.55 ± 0.00	75.46 ± 0.03	**4.37 ± 0.04**	75.47 ± 0.02	75.97 ± 0.05	75.74 ± 0.04
ΔConf. ID–OOD (%) ↑	5.24 ± 0.73	**21.41 ± 0.33**	0.15 ± 0.00	0.01 ± 0.00	0.25 ± 0.00	14.51 ± 0.13	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
ΔConf. ID–Adv (%) ↑	−67.34 ± 0.56	−11.81 ± 0.11	0.17 ± 0.00	0.01 ± 0.00	**0.68 ± 0.00**	−1.70 ± 0.13	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00
Inf. Time (s) ↓	453.67 ± 115.99	**397.02 ± 8.76**	451.55 ± 119.19	404.87 ± 19.59	411.16 ± 19.94	418.26 ± 43.74	7356.02 ± 105.32	1033.67 ± 55.50	423.66 ± 44.30

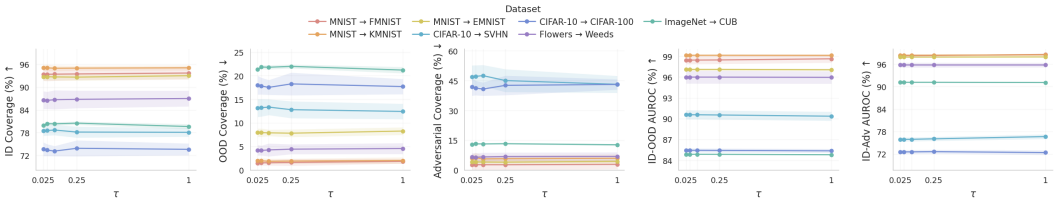


Figure 4: Hyperparameter sensitivity of CLEAR to the perturbation magnitude τ. Each line denotes a dataset. From left to right, panels report ID coverage, OOD coverage, adversarial coverage, ID–OOD AUROC, and ID–adversarial AUROC. All other CLEAR hyperparameters are held fixed at their default values.

Consequently, CLEAR depends more strongly on whether the local latent behaviour is unusual relative to calibrated ID structure than on the absolute size of the perturbation itself. We use $\tau = 0.05$, which provides a local perturbation while lying within the broad region of stable performance.

Figure 5 varies the number of sampled latent views V. With V views, CLEAR evaluates $|\mathcal{P}_V| = \binom{V}{2} = \frac{V(V-1)}{2}$ unordered view pairs before averaging their conflicts. Increasing V therefore increases the number of local comparisons quadratically, but does not directly increase the magnitude of the final score because the pairwise conflicts are averaged. Across the evaluated datasets, the resulting metrics remain almost unchanged as V increases, indicating that a small number of views is already sufficient to characterise the local latent inconsistency used by CLEAR. Additional views consequently provide little empirical benefit while increasing the number of pairwise comparisons. We therefore select $V = 3$, the smallest setting beyond the two-view case that provides multiple distinct pairwise comparisons, yielding three view pairs at low computational cost.

Figure 6 varies the evidence down-weighting strength δ in $s(x) = \exp(-\delta C(x))$. Increasing δ therefore strengthens the effect of any non-zero conflict while leaving samples with $C(x) = 0$ unchanged. The results exhibit the corresponding trade-off. Moving away from very small values

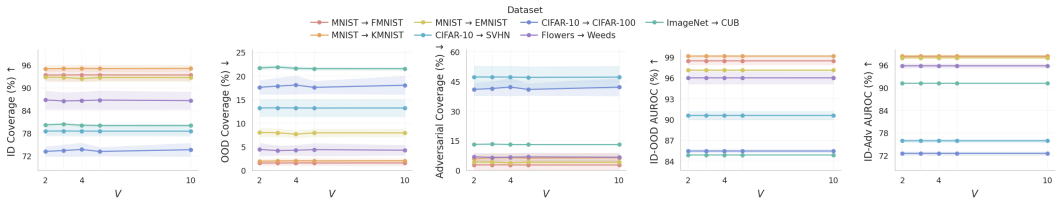


Figure 5: Hyperparameter sensitivity of CLEAR to the number of latent views V. Each line denotes a dataset. From left to right, panels report ID coverage, OOD coverage, adversarial coverage, ID–OOD AUROC, and ID–adversarial AUROC. All other CLEAR hyperparameters are held fixed at their default values.

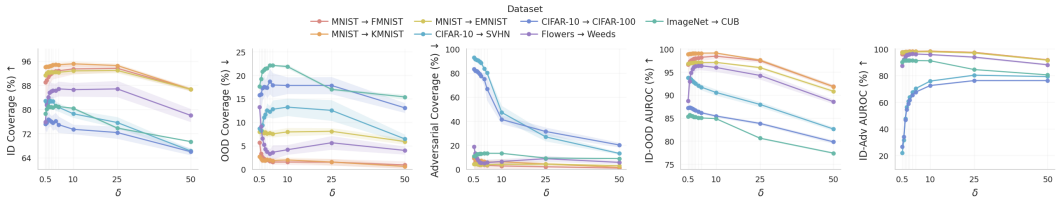


Figure 6: Hyperparameter sensitivity of CLEAR to the evidence down-weighting strength δ. Each line denotes a dataset. From left to right, panels report ID coverage, OOD coverage, adversarial coverage, ID–OOD AUROC, and ID–adversarial AUROC. All other CLEAR hyperparameters are held fixed at their default values.

substantially improves adversarial rejection and ID-Adv. separation on the more challenging datasets, indicating that weak down-weighting does not sufficiently penalise unsupported evidence. However, increasingly large values eventually reduce ID coverage and ID-OOD AUROC on several datasets, as even moderate latent conflict induces a strong reduction in evidence. In contrast, the easier dataset pairs remain comparatively insensitive across a broad range of δ. We select $\delta = 10$, which lies within the stable intermediate region and provides strong OOD and adversarial discrimination without the degradation in retained ID behaviour observed under more aggressive down-weighting.

CLEAR defines the reference conflict level as the median calibration conflict, $r = \mathrm{median}_{x_i \in \mathcal{D}_{\mathrm{cal}}} \overline{C}_{\mathrm{raw}}(x_i)$, which provides a robust central reference for typical ID latent behaviour and reduces sensitivity to unusually high-conflict calibration examples. Although the method is defined using the median, the underlying reference statistic can be varied by replacing the median with different calibration quantiles. We therefore perform this additional sensitivity analysis in Figure 7 to assess whether the intended median reference is well behaved, rather than treating the reference quantile as a parameter to tune for each dataset.

The results favour an intermediate reference level rather than either extreme. When a low calibration quantile is used, the resulting reference r becomes small and a broad range of positive raw conflict scores are mapped towards large values of $C(x)$. This can reduce discrimination by causing both

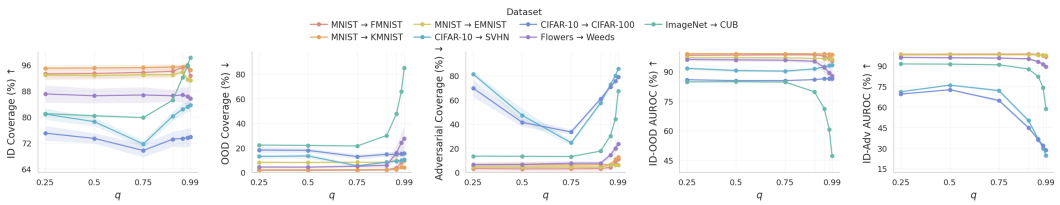


Figure 7: Hyperparameter sensitivity of CLEAR to the reference conflict quantile q. Each line denotes a dataset. From left to right, panels report ID coverage, OOD coverage, adversarial coverage, ID–OOD AUROC, and ID–adversarial AUROC. All other CLEAR hyperparameters are held fixed at their default values.



Figure 8: Hyperparameter sensitivity of CLEAR for regression to the number of bins B. Each line denotes a dataset. From left to right, panels report ID coverage, OOD coverage, adversarial coverage, ID–OOD AUROC, and ID–adversarial AUROC. All other CLEAR hyperparameters are held fixed at their default values.

moderately and strongly conflicting samples to receive similarly aggressive evidence down-weighting. Conversely, very high calibration quantiles produce a large reference level, making CLEAR increasingly permissive because more samples fall below or close to the calibrated reference and therefore receive little adjustment. The median lies within the stable intermediate region across the evaluated datasets, supporting its use as the fixed reference statistic throughout the main experiments.

Finally, Figure 8 examines the regression-specific number of output bins B. The number of regression bins B controls the granularity of the output-conditioned latent geometry. Smaller values produce broader groups with more calibration examples and therefore more stable estimates of μ_g and Σ_g, while larger values provide more locally specific geometry at the cost of fewer samples per group. The ablation therefore examines the trade-off between geometric specificity and reliable calibration.

The results indicate that relatively coarse output partitioning is sufficient, with little benefit from increasing B aggressively. ID coverage remains broadly stable across the evaluated range, but larger values of B tend to increase OOD and adversarial coverage for several datasets, indicating weaker rejection as the calibration examples are divided among increasingly small groups. This is reflected more clearly in the separation metrics: ID-OOD and ID-Adv. AUROC generally remain strongest for small-to-moderate B, while several datasets deteriorate slightly as B becomes large. In particular, the Airfoil and UTKFace regression settings show increasingly poorer separation at $B = 20$ and $B = 50$, consistent with the per-bin estimates of μ_g and Σ_g becoming less reliable when supported by fewer calibration examples. Very small B, however, produces comparatively coarse output-conditioned regions that may obscure local changes in latent geometry. We therefore use $B = 10$, which provides a practical balance between output specificity and sufficiently populated calibration groups, while lying within the region of strong performance across the regression datasets.

C.4 Threshold Analysis

Table 5 examines the effect of the uncertainty measure used for thresholding. CLEAR performs consistently across all four measures, maintaining strong OOD and adversarial separation despite changes in the uncertainty statistic. Differential entropy provides the strongest overall balance, retaining 94.83% of ID samples while reducing OOD and adversarial coverage to 1.84% and 7.21%, with ID-OOD AUROC of 98.54% and ID-Adv. AUROC of 97.66%.

The alternative measures remain competitive, particularly for adversarial detection. Predictive entropy achieves an AAUROC of 95.36% with adversarial coverage of 4.61%, while mutual information and total evidence retain AAUROC above 93%. Although these alternatives reduce ID coverage and OOD separation relative to differential entropy, CLEAR continues to distinguish unreliable inputs effectively across all settings. This indicates that the effectiveness of the latent conflict adjustment is not tied to a particular uncertainty statistic, while differential entropy offers the most favourable balance between retaining ID predictions and rejecting OOD and adversarial inputs.

The comparative approaches show greater variation across threshold types, with changes in the uncertainty statistic often producing larger shifts in coverage and AUROC. CLEAR therefore exhibits comparatively stable behaviour across the evaluated thresholding choices, while differential entropy provides a consistent default operating point for the main experiments.

Table 5: Mean accuracy, OOD detection, and adversarial attack detection performance with standard deviation of the comparative approaches across a variety of uncertainty thresholding types for the MNIST → FashionMNIST setting. For compactness, we denote adversarial AUROC as AAUROC. Highlighted cells denote the best performance for each metric.

	PostNet	EDL	I-EDL	H-EDL	R-EDL	RED	S-EDL	C-EDL	CLEAR
Type	In-Training						Post-Hoc		
Differential Entropy (Default)									
ID Acc (%) ↑	98.96 ± 0.11	99.26 ± 0.07	99.20 ± 0.07	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	99.26 ± 0.07	**99.28 ± 0.04**	99.26 ± 0.07
ID Cov (%) ↑	91.93 ± 1.89	89.16 ± 2.73	89.49 ± 3.06	85.63 ± 5.84	92.33 ± 2.10	92.15 ± 1.48	89.19 ± 2.67	88.29 ± 3.54	**94.83 ± 0.85**
OOD Cov (%) ↓	5.14 ± 2.63	12.57 ± 8.80	13.20 ± 10.46	21.84 ± 16.94	7.83 ± 5.99	3.64 ± 1.03	12.68 ± 8.93	16.60 ± 9.68	**1.84 ± 0.91**
Adv Cov (%) ↓	86.16 ± 10.82	26.73 ± 13.70	31.36 ± 16.33	19.44 ± 14.92	64.58 ± 12.73	28.79 ± 9.32	23.05 ± 13.12	9.20 ± 7.30	**7.21 ± 3.90**
AUROC (%) ↑	97.55 ± 1.12	92.54 ± 5.70	91.94 ± 7.22	86.18 ± 12.94	95.95 ± 3.22	97.34 ± 0.85	92.52 ± 5.73	91.20 ± 6.14	**98.54 ± 0.36**
AAUROC (%) ↑	40.19 ± 12.45	88.78 ± 5.29	88.75 ± 4.11	88.59 ± 10.79	79.67 ± 3.17	77.72 ± 8.60	90.23 ± 4.71	94.91 ± 2.13	**97.66 ± 0.53**
Predictive Entropy									
ID Acc (%) ↑	98.96 ± 0.11	98.69 ± 0.16	**99.20 ± 0.07**	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	98.69 ± 0.17	98.76 ± 0.16	98.69 ± 0.16
ID Cov (%) ↑	90.99 ± 2.12	90.20 ± 2.29	89.96 ± 2.88	88.31 ± 4.96	**92.37 ± 2.12**	92.16 ± 1.49	90.25 ± 2.55	89.91 ± 2.57	78.76 ± 7.87
OOD Cov (%) ↓	4.23 ± 2.48	8.72 ± 3.07	10.84 ± 8.24	4.55 ± 2.37	9.03 ± 6.20	**3.52 ± 1.05**	8.72 ± 2.90	6.07 ± 2.06	8.00 ± 2.71
Adv Cov (%) ↓	84.30 ± 11.35	10.58 ± 5.66	24.85 ± 12.99	9.80 ± 6.16	65.87 ± 12.69	17.61 ± 6.61	9.31 ± 4.99	**2.35 ± 0.94**	4.61 ± 2.39
AUROC (%) ↑	97.44 ± 1.06	96.72 ± 1.32	93.34 ± 6.00	95.56 ± 2.49	95.65 ± 3.40	**97.47 ± 0.87**	96.72 ± 1.32	97.15 ± 0.93	93.31 ± 2.92
AAUROC (%) ↑	42.20 ± 12.37	95.72 ± 1.67	90.56 ± 4.26	94.73 ± 2.94	79.35 ± 3.22	89.22 ± 3.68	96.11 ± 1.56	**98.23 ± 0.56**	95.36 ± 2.77
Mutual Information									
ID Acc (%) ↑	98.96 ± 0.11	98.69 ± 0.16	**99.20 ± 0.07**	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	98.69 ± 0.17	98.76 ± 0.16	98.69 ± 0.16
ID Cov (%) ↑	91.40 ± 1.47	84.61 ± 5.71	87.69 ± 2.97	86.08 ± 5.87	90.83 ± 2.54	**92.19 ± 1.49**	84.57 ± 5.51	86.43 ± 5.77	75.49 ± 12.68
OOD Cov (%) ↓	6.35 ± 3.65	16.15 ± 7.00	13.86 ± 12.13	28.25 ± 19.40	9.64 ± 6.51	**3.99 ± 1.06**	16.23 ± 7.80	11.23 ± 5.69	10.07 ± 6.50
Adv Cov (%) ↓	86.92 ± 10.22	71.58 ± 15.78	33.91 ± 15.05	22.51 ± 18.23	71.80 ± 18.28	38.11 ± 9.72	63.67 ± 16.56	22.90 ± 7.48	**8.38 ± 5.31**
AUROC (%) ↑	**97.19 ± 1.31**	91.68 ± 4.80	90.51 ± 8.67	80.93 ± 14.99	94.95 ± 3.79	96.98 ± 0.99	91.68 ± 4.78	94.15 ± 3.93	91.47 ± 3.76
AAUROC (%) ↑	50.04 ± 7.93	61.86 ± 11.32	86.17 ± 4.22	85.54 ± 12.83	77.00 ± 4.49	68.65 ± 10.55	67.70 ± 10.15	89.92 ± 4.35	**93.38 ± 4.38**
Total Evidence									
ID Acc (%) ↑	98.96 ± 0.11	98.69 ± 0.16	**99.20 ± 0.07**	98.33 ± 0.38	99.18 ± 0.08	99.01 ± 0.12	98.69 ± 0.17	98.76 ± 0.16	98.69 ± 0.16
ID Cov (%) ↑	91.46 ± 1.57	84.41 ± 3.46	87.04 ± 3.94	86.05 ± 5.58	90.02 ± 2.31	**95.73 ± 0.31**	84.13 ± 5.26	86.37 ± 4.82	75.63 ± 12.73
OOD Cov (%) ↓	**6.45 ± 3.60**	17.91 ± 10.31	13.80 ± 11.71	28.22 ± 18.99	10.37 ± 7.26	16.83 ± 11.41	17.60 ± 10.01	12.02 ± 6.00	10.63 ± 6.49
Adv Cov (%) ↓	87.00 ± 10.32	72.79 ± 15.85	34.22 ± 15.05	22.35 ± 17.90	71.85 ± 15.70	77.64 ± 4.59	65.18 ± 15.98	25.83 ± 7.37	**8.64 ± 5.36**
AUROC (%) ↑	**97.19 ± 1.31**	90.86 ± 5.39	90.28 ± 8.88	80.85 ± 15.00	94.37 ± 4.20	82.96 ± 11.28	90.86 ± 5.37	93.76 ± 4.34	91.35 ± 3.82
AAUROC (%) ↑	38.15 ± 12.42	59.83 ± 11.71	85.72 ± 4.27	85.49 ± 12.84	75.11 ± 5.79	29.18 ± 5.93	65.74 ± 10.63	88.83 ± 4.90	**93.27 ± 4.46**

C.5 Adversarial Attack Analysis

Table 6 evaluates robustness across multiple adversarial attacks and perturbation strengths. CLEAR performs consistently strongly across the evaluated settings, achieving the lowest adversarial coverage in all but one case and the highest AAUROC in all but the weakest L_2-PGD setting, where it effectively matches PostNet (98.93% versus 98.98%). Its advantage becomes more pronounced as gradient-based attacks strengthen. Under L_2-PGD at strength 1.0, CLEAR retains adversarial coverage of only 7.21% and AAUROC of 97.66%, while several alternatives exhibit substantially greater degradation. A similar pattern is observed for FGSM, where CLEAR achieves near-zero adversarial coverage at strength 1.0 while maintaining an AAUROC of 97.88%.

CLEAR is also particularly strong under AutoAttack, achieving the highest AAUROC at all three perturbation strengths: 98.87%, 98.35%, and 96.14%. At the strongest setting, its adversarial coverage remains low at 23.79%, only slightly above H-EDL at 21.56%, while CLEAR maintains a substantially higher AAUROC of 96.1% compared with 88.45%. Thus, despite the small difference in coverage, CLEAR provides markedly stronger overall separation between clean and adversarial inputs. For Salt and Pepper corruption, CLEAR also achieves the lowest coverage and highest AAUROC across all three strengths, reaching 1.28% coverage and 98.63% AAUROC at strength 1.0. Overall, CLEAR maintains strong performance across qualitatively different attack mechanisms and degrades substantially less than most comparative approaches as adversarial strength increases.

C.6 Full Object Detection Results

To evaluate CLEAR beyond classification and regression, we additionally consider object detection on CelebA using YOLO11n (Khanam & Hussain, 2024). Object detection provides a useful test case because uncertainty should reflect not only whether a detection is correct, but also how accurately its bounding box is localised.

We use YOLO11n as a fixed base detector and train an Evidential Meta Model (EMM) (Shen et al., 2023) on its intermediate features to produce detection-level evidential uncertainty over bounding-box localisation. EMM is well suited to this setting because it adds an evidential uncertainty model on top of a pretrained detector without requiring the detector itself to be modified. This allows YOLO11n to remain unchanged while providing the evidential outputs required by CLEAR. Otherwise, YOLO11n either cannot provide preliminary uncertainty outputs for CLEAR to build upon or would require full retraining if the head is replaced with a Dirichlet head, which would be extremely computationally expensive.

Table 6: Mean adversarial coverage and AUROC performance with standard deviation of the comparative approaches across a variety of adversarial attack types and perturbation strengths for the MNIST → FashionMNIST setting. Highlighted cells denote the best performance for each attack and strength.

		PostNet	EDL	I-EDL	H-EDL	R-EDL	RED	S-EDL	C-EDL	CLEAR
		In-Training						Post-Hoc		
Adversarial Coverage (%) ↓										
L_2-PGD	0.1	0.80 ± 0.64	8.12 ± 7.34	8.58 ± 9.30	20.21 ± 16.57	3.38 ± 3.69	1.54 ± 0.81	8.59 ± 7.64	13.27 ± 9.31	$\mathbf{0.65 \pm 0.45}$
	0.5	53.91 ± 16.17	7.07 ± 6.38	9.60 ± 6.35	17.66 ± 15.24	28.46 ± 15.85	19.03 ± 4.72	5.77 ± 5.75	6.21 ± 5.89	$\mathbf{1.16 \pm 0.87}$
	1.0	86.16 ± 10.82	26.73 ± 13.70	31.36 ± 16.33	19.44 ± 14.92	64.58 ± 12.73	28.79 ± 9.32	23.05 ± 13.12	9.20 ± 7.30	$\mathbf{7.21 \pm 3.90}$
FGSM	0.1	9.59 ± 1.92	29.87 ± 5.88	18.31 ± 7.79	26.51 ± 16.57	24.28 ± 7.80	8.22 ± 4.64	29.88 ± 6.31	11.23 ± 3.10	$\mathbf{4.23 \pm 3.70}$
	0.5	16.97 ± 7.94	31.42 ± 13.40	38.04 ± 14.34	61.58 ± 25.86	52.62 ± 17.53	35.16 ± 18.41	31.78 ± 13.17	22.40 ± 12.71	$\mathbf{2.88 \pm 5.05}$
	1.0	15.44 ± 14.65	91.85 ± 9.20	28.13 ± 23.86	59.22 ± 31.95	49.57 ± 30.35	68.40 ± 24.68	92.35 ± 8.39	80.22 ± 17.23	$\mathbf{0.01 \pm 0.01}$
S&P	0.1	5.00 ± 2.60	8.22 ± 1.95	13.19 ± 10.47	21.79 ± 16.94	7.75 ± 6.02	3.51 ± 0.98	8.39 ± 2.11	7.89 ± 2.18	$\mathbf{1.61 \pm 0.40}$
	0.5	4.48 ± 2.46	8.04 ± 1.94	12.97 ± 10.58	21.49 ± 16.92	7.45 ± 6.19	3.11 ± 0.91	8.19 ± 2.03	7.35 ± 2.06	$\mathbf{1.47 \pm 0.43}$
	1.0	3.70 ± 2.27	7.84 ± 1.85	12.75 ± 10.73	21.06 ± 16.91	7.16 ± 6.41	2.51 ± 0.86	8.04 ± 1.94	6.57 ± 1.78	$\mathbf{1.28 \pm 0.47}$
AA (L_2)	0.1	4.40 ± 1.98	12.07 ± 7.66	10.86 ± 8.34	21.25 ± 12.44	6.23 ± 3.70	3.33 ± 0.72	8.91 ± 5.88	11.98 ± 6.23	$\mathbf{1.59 \pm 0.07}$
	0.5	80.64 ± 12.58	15.51 ± 5.96	16.14 ± 6.54	20.73 ± 12.62	47.19 ± 13.70	26.86 ± 9.45	10.09 ± 4.71	5.96 ± 4.18	$\mathbf{4.49 \pm 3.25}$
	1.0	99.76 ± 0.41	81.16 ± 6.09	81.19 ± 11.77	$\mathbf{21.56 \pm 13.71}$	85.53 ± 4.98	68.42 ± 8.36	77.43 ± 6.00	26.55 ± 5.86	23.79 ± 8.59
AAUROC (%) ↑										
L_2-PGD	0.1	98.98 ± 0.36	94.93 ± 4.14	94.24 ± 5.72	87.02 ± 12.54	97.86 ± 1.64	97.65 ± 0.85	94.70 ± 4.32	92.88 ± 5.38	$\mathbf{98.93 \pm 0.30}$
	0.5	78.24 ± 8.03	95.65 ± 1.87	95.17 ± 1.92	88.83 ± 11.23	93.53 ± 1.04	86.71 ± 5.43	96.07 ± 1.73	96.00 ± 2.51	$\mathbf{98.59 \pm 0.37}$
	1.0	40.19 ± 12.45	88.78 ± 5.29	88.75 ± 4.11	88.59 ± 10.79	79.67 ± 3.17	77.72 ± 8.60	90.23 ± 4.71	94.91 ± 2.13	$\mathbf{97.66 \pm 0.53}$
FGSM	0.1	96.63 ± 0.57	88.72 ± 3.31	92.05 ± 2.75	85.76 ± 11.74	91.14 ± 3.18	93.53 ± 5.44	88.81 ± 3.29	95.07 ± 1.79	$\mathbf{98.16 \pm 0.39}$
	0.5	94.74 ± 2.53	88.71 ± 3.54	81.26 ± 7.36	66.26 ± 22.34	78.11 ± 10.34	84.98 ± 8.05	88.66 ± 3.55	91.98 ± 2.70	$\mathbf{97.70 \pm 0.46}$
	1.0	94.92 ± 3.55	53.19 ± 14.44	88.36 ± 10.34	65.80 ± 32.64	83.19 ± 14.41	63.24 ± 14.52	53.02 ± 14.47	69.35 ± 12.22	$\mathbf{97.88 \pm 0.64}$
S&P	0.1	97.62 ± 1.10	95.92 ± 1.24	91.96 ± 7.23	86.20 ± 12.94	95.98 ± 3.23	97.38 ± 0.85	95.91 ± 1.25	96.12 ± 1.53	$\mathbf{98.47 \pm 0.46}$
	0.5	97.90 ± 1.05	95.99 ± 1.22	92.02 ± 7.28	86.30 ± 12.94	96.13 ± 3.27	97.60 ± 0.86	95.98 ± 1.22	96.32 ± 1.47	$\mathbf{98.54 \pm 0.46}$
	1.0	98.18 ± 0.97	96.06 ± 1.19	92.10 ± 7.35	86.46 ± 12.93	96.26 ± 3.32	97.83 ± 0.85	96.05 ± 1.20	96.56 ± 1.37	$\mathbf{98.63 \pm 0.47}$
AA (L_2)	0.1	$\mathbf{98.07 \pm 0.58}$	93.56 ± 4.62	93.71 ± 5.47	87.48 ± 5.51	97.06 ± 1.77	97.71 ± 0.35	95.48 ± 2.77	94.41 ± 3.42	98.87 ± 0.31
	0.5	58.96 ± 12.05	92.48 ± 3.66	92.48 ± 3.43	87.91 ± 5.68	89.39 ± 1.53	89.61 ± 4.40	95.00 ± 1.56	96.19 ± 2.01	$\mathbf{98.35 \pm 0.06}$
	1.0	1.82 ± 1.16	59.46 ± 1.88	57.38 ± 6.41	88.45 ± 5.80	40.13 ± 4.43	71.40 ± 21.68	65.47 ± 2.05	91.02 ± 1.72	$\mathbf{96.14 \pm 0.14}$

Table 7: Uncertainty quality on clean CelebA detections. Pearson's r is reported between predictive uncertainty and localisation error ($1-\mathrm{IoU}$), such that higher correlation indicates that poorly localised detections receive greater uncertainty.

Method	AUROC (%) ↑	AUPRC (%) ↑	FPR95 (%) ↓	ECE (%) ↓	Pearson r ↑
EMM	0.503 ± 0.030	0.125 ± 0.028	0.907 ± 0.030	0.844 ± 0.023	0.003 ± 0.084
CLEAR	$\mathbf{0.585 \pm 0.042}$	$\mathbf{0.211 \pm 0.087}$	$\mathbf{0.824 \pm 0.095}$	$\mathbf{0.750 \pm 0.160}$	$\mathbf{0.168 \pm 0.099}$

After EMM training, both YOLO11n and EMM are frozen. CLEAR is then calibrated using held-out CelebA detections and applied post-hoc to the EMM latent representation and evidential outputs. The comparison between EMM and EMM+CLEAR therefore isolates the effect of CLEAR: the underlying detector and bounding-box predictions remain unchanged, and only the evidential uncertainty is adjusted according to latent conflict.

We evaluate whether this adjusted uncertainty better reflects localisation quality using localisation error, defined as $1-\mathrm{IoU}$, together with AUROC, AUPRC, FPR95, calibration error, and the correlation between uncertainty and localisation error. We additionally evaluate the same relationship under brightness, blur, noise, compression, and resize/crop covariates.

Table 7 shows that CLEAR consistently improves uncertainty quality over EMM on clean CelebA detections. CLEAR increases AUROC from 0.503 to 0.585 and AUPRC from 0.125 to 0.211, while reducing FPR95 from 0.907 to 0.824 and ECE from 0.844 to 0.750. Importantly, the Pearson correlation between predictive uncertainty and localisation error increases from approximately zero for EMM ($r = 0.003$) to $r = 0.168$ with CLEAR. This indicates that CLEAR produces uncertainty that is more informative of bounding-box quality, with poorly localised detections more consistently receiving higher uncertainty.

Table 8: Linear association between predictive uncertainty and localisation error under covariate corruption on CelebA. Pearson's r is reported, where higher is better.

Corruption	EMM r ↑	CLEAR r ↑	Δr
Brightness	0.009 ± 0.017	$\mathbf{0.046 \pm 0.028}$	$+0.037$
Blur	0.008 ± 0.025	$\mathbf{0.052 \pm 0.035}$	$+0.045$
Noise	0.022 ± 0.027	$\mathbf{0.060 \pm 0.022}$	$+0.039$
Compression	0.002 ± 0.022	$\mathbf{0.048 \pm 0.024}$	$+0.046$
Resize/Crop	0.009 ± 0.013	$\mathbf{0.037 \pm 0.024}$	$+0.027$
Overall	0.010 ± 0.019	$\mathbf{0.049 \pm 0.023}$	$+0.039$

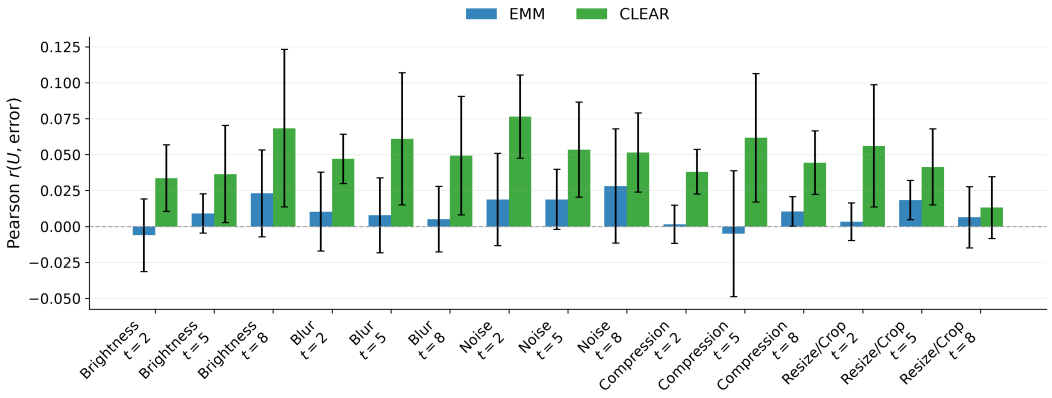


Figure 9: Pearson correlation between predictive uncertainty and localisation error under covariate corruption on CelebA. Higher correlation indicates that detections with larger localisation error receive greater uncertainty.

This behaviour remains consistent under covariate shift. As shown in Table 8, CLEAR improves the uncertainty-localisation correlation for every evaluated corruption, increasing the overall Pearson correlation from 0.010 to 0.049. The improvement is observed across brightness, blur, noise, compression, and resize/crop corruptions, with the largest gains occurring for compression and blur.

Figure 9 further shows that this stronger relationship is generally maintained across corruption severities and all types, whereas EMM frequently remains close to zero correlation. Although the absolute correlations remain modest, the consistent improvement across clean data, corruption types, and corruption strengths suggests that CLEAR provides a more reliable localisation-aware uncertainty signal than the underlying EMM alone.

Overall, these results show that CLEAR extends effectively to object detection without modifying the underlying detector or its bounding-box predictions. Across both clean and corrupted CelebA inputs, CLEAR consistently makes the evidential uncertainty more reflective of localisation quality than EMM alone. This supports the broader task-agnostic design of CLEAR, showing that its latent-conflict mechanism can transfer from standard classification and regression settings to structured prediction tasks where uncertainty must capture the reliability of individual detections.

C.7 Held-Out Analysis of the Latent Whitening Geometry

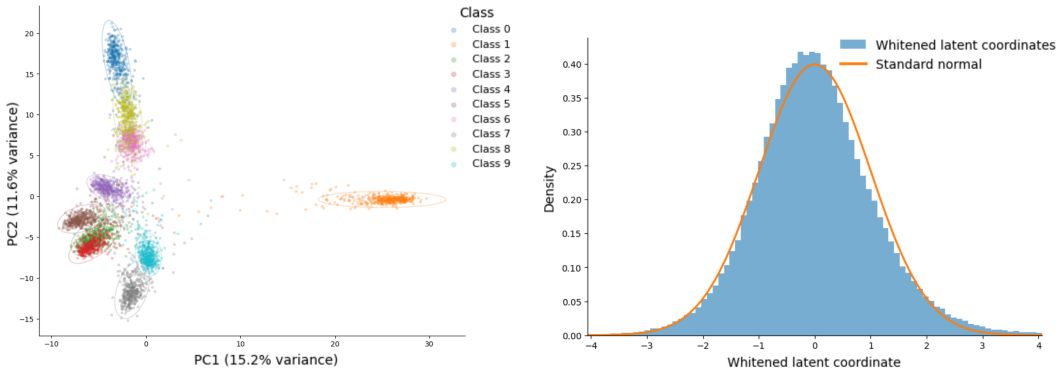
CLEAR uses class/binned-conditional latent statistics to define the local geometry in which latent-view conflict is measured. For each group g, the training representations are used to estimate a mean μ_g and covariance $\Sigma_g = L_g L_g^\top$, and latent representations are subsequently centred and whitened as:

$$\tilde{z} = L_g^{-1}(z - \mu_g) \tag{43}$$

We therefore evaluate whether this training-derived geometry remains meaningful on held-out ID data.

Figure 10 provides a qualitative view of the learned geometry on the MNIST dataset. In the PCA projection, the class- conditional covariance contours broadly align with the orientation and spread of the corresponding held-out latent clusters, indicating that the estimated second-order structure captures the main class-wise geometry. After whitening, the pooled latent coordinates are centred and approximately bell-shaped, but remain broader and heavier-tailed than a standard normal density. Whitening removes class-specific location, scale, and linear covariance structure, but does not impose Gaussianity on the underlying latent representations, which are learned for the prediction task rather than explicitly constrained to follow a Gaussian distribution. Accordingly, CLEAR uses the fitted statistics as a local second-order geometry rather than as an exact generative model of the latent distribution.

We next quantify the effectiveness of the whitening transformation on held-out data. Whitening centres each representation using its class-specific mean and rescales it according to the corresponding



(a) Held-out latent geometry with fitted covariance contours.

(b) Whitened latent coordinates compared with a standard normal density.

Figure 10: Held-out validation of the CLEAR latent geometry on MNIST. **Left:** PCA projection of held-out latent representations together with class-conditional covariance contours estimated from the training set. **Right:** Distribution of held-out latent coordinates after class-conditional whitening. The whitened coordinates are centred and approximately bell-shaped, but exhibit heavier tails than the standard normal.

Table 9: Latent-geometry diagnostics for CLEAR on MNIST. Whitening statistics are computed per class on held-out ID representations using geometry estimated from the training set.

Diagnostic	Ideal	Result
Whitened mean norm / $\sqrt{d}$	0	0.072
Mean absolute off-diagonal covariance	0	0.071
Mean diagonal covariance	1	1.286
Empirical coverage of 50% Gaussian region	0.50	0.529
Empirical coverage of 95% Gaussian region	0.95	0.634
Radial KS statistic	0	0.358
Uniform vs. Gaussian conflict Spearman ρ	1	0.99998
Rademacher vs. Gaussian conflict Spearman ρ	1	0.99998

Cholesky covariance factor; if the estimated geometry transfers to unseen samples, the transformed representations should have a mean close to zero and covariance close to the identity matrix. Across classes, the normalised whitened mean norm was 0.072, the mean absolute off-diagonal covariance was 0.071, and the mean diagonal covariance was 1.286. These values indicate that the whitening geometry transfers reasonably well to unseen ID samples: the transformed representations remain approximately centred and largely decorrelated, with only moderate deviation from unit variance.

We also test the stronger distributional condition that the whitened latent representations follow an exact multivariate standard normal distribution. This condition is only partially satisfied. The nominal 50% Gaussian region contains 52.9% of held-out samples, but the nominal 95% region contains only 63.4%, and the radial Kolmogorov-Smirnov statistic is 0.358. Thus, the fitted Gaussian geometry captures the central structure of the latent space more faithfully than its tails. The network objective does not explicitly constrain the class-conditional latent representations to be Gaussian, so these deviations primarily concern higher-order distributional structure rather than the mean and covariance geometry used by CLEAR.

Finally, we test whether CLEAR depends critically on the use of Gaussian perturbations for generating latent views. Replacing the Gaussian perturbations with variance-matched Uniform or Rademacher noise leaves the conflict scores essentially unchanged, with Spearman rank correlations of 0.99998 in both cases relative to the Gaussian baseline. This shows that CLEAR does not rely on exact Gaussianity of either the latent distribution or the perturbation law. Instead, its conflict mechanism primarily depends on the fitted local second-order geometry providing a stable coordinate system in which small perturbations expose latent inconsistency.

Table 10: Sensitivity of CLEAR to the latent representation used for conflict estimation on MNIST→FashionMNIST over 10 runs. The same trained evidential model is shared across all representations within each run, with only the CLEAR latent geometry recalibrated. AAUROC denotes mean ID-Adv. AUROC across the evaluated L_2-PGD attack strengths.

	Layer 1	**Layer 2**	**Layer 3**	**Penultimate**
Latent Dim.	64	128	256	512
ID Acc (%) ↑	99.25 ± 0.06	99.25 ± 0.06	99.25 ± 0.06	99.25 ± 0.06
ID Cov (%) ↑	98.98 ± 0.22	98.82 ± 0.27	97.90 ± 0.55	93.81 ± 1.44
OOD Cov (%) ↓	0.41 ± 0.22	0.53 ± 0.16	0.85 ± 0.40	1.82 ± 0.81
AUROC (%) ↑	99.88 ± 0.04	99.91 ± 0.02	99.80 ± 0.05	98.42 ± 0.45
AAUROC (%) ↑	99.88 ± 0.02	99.93 ± 0.02	99.83 ± 0.04	98.29 ± 0.13

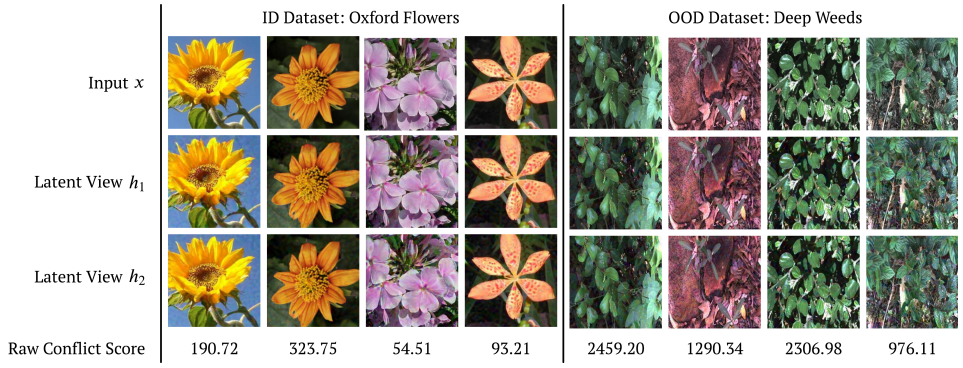


Figure 11: Qualitative visualisation of CLEAR's latent perturbation views for representative Oxford Flowers ID and DeepWeeds OOD inputs. For each input x, two covariance-shaped latent views h_1 and h_2 are visualised through feature inversion. The reconstructed views preserve the dominant semantic content of the input, while the corresponding raw conflict scores are substantially larger for the shown OOD examples than for the ID examples.

C.8 Sensitivity to Latent Representation Depth

CLEAR uses the penultimate representation by default, as this is the final task-relevant feature representation provided to the evidential head. To assess whether its performance depends on this particular choice, we additionally evaluate CLEAR using representations extracted at progressively deeper stages of the same ResNet-18 backbone. Specifically, we consider the outputs of residual layers 1-3 and the penultimate representation, corresponding to latent dimensionalities of 64, 128, 256, and 512, respectively. For each of 10 runs on MNIST→FashionMNIST, the underlying evidential model is trained only once and then held fixed across all four settings.

Table 10 shows that CLEAR remains highly effective across all evaluated representation depths. As expected, ID accuracy is identical across layers because the frozen predictive model is unchanged. The earlier representations perform particularly strongly on this benchmark: layer 1 achieves the highest ID coverage of 98.98% and lowest OOD coverage of 0.41%, while layer 2 achieves the highest ID-OOD AUROC of 99.91% and AAUROC of 99.93%. The penultimate representation nevertheless retains strong separation, with AUROC and AAUROC of 98.42% and 98.29%, respectively.

Importantly, these results show that the penultimate representation used throughout the main experiments was not selected through benchmark-specific layer optimisation; indeed, shallower representations perform better for this particular dataset pair. We retain the penultimate layer as the default because it provides the representation directly consumed by the evidential head and gives a consistent, architecture-independent definition of the latent space across the different models and tasks considered in this work. The results further indicate that alternative intermediate representations could provide additional gains when architecture-specific layer selection is appropriate.

C.9 Qualitative Analysis of Latent Perturbation Views

Figure 11 provides an input-space interpretation of the latent perturbations used by CLEAR. For each input, two latent views are sampled using the predicted group's calibrated covariance geometry, following Equation 11. To visualise these otherwise high-dimensional representations, we perform feature inversion by optimising a copy of the input image such that its representation under the frozen encoder approaches the corresponding target latent view, while applying pixel-distance and total-variation regularisation to retain an interpretable image. Across the Oxford Flowers examples, the resulting feature-inverted views remain visually close to the corresponding input, preserving the dominant flower structure, colour, and semantic content. The DeepWeeds examples similarly retain their overall appearance, showing that the perturbations remain local rather than moving to arbitrary regions of the representation space. Importantly, however, their latent behaviour differs substantially when evaluated relative to the calibrated Flowers geometry. For the examples shown, raw conflict ranges from 54.51 to 323.75 for ID inputs, compared with 976.11 to 2459.20 for OOD inputs. Thus, even when the perturbed views remain visually plausible, CLEAR can identify substantially greater incompatibility with the latent structure expected for the predicted group.

This distinction illustrates that CLEAR's conflict score does not simply measure the visible magnitude of a perturbation. Instead, it captures the cross-dimensional behaviour of the perturbed representations after conditioning on the calibrated group geometry. The qualitative examples therefore provide an intuitive view of how locally similar latent perturbations can remain compatible for familiar ID inputs but produce substantially greater conflict for unsupported OOD representations. Feature inversion is used here only as an interpretive visualisation and is not part of CLEAR inference. Moreover, the perturbation magnitude is deliberately enlarged for this analysis to make the corresponding image-space changes perceptible; the absolute conflict values shown should therefore be interpreted qualitatively rather than as reproductions of the operating scores used in the main experiments.

D Additional Discussion

Trade-offs between OOD and adversarial uncertainty. Some comparative approaches achieve stronger *individual* metrics than CLEAR, but Figure 2 shows that these gains are often accompanied by considerably larger losses on complementary uncertainty objectives. Importantly, the trade-off is frequently asymmetric: when CLEAR does not rank first, it generally remains close to the strongest result, whereas methods that outperform CLEAR on one metric can fall substantially behind on another. On MNIST→KMNIST, for example, CLEAR's AAUROC of 97.36% is only 0.68 percentage points below H-EDL and 0.15 below C-EDL, while still achieving the strongest ID coverage, OOD coverage, and ID-OOD AUROC. The contrast is more pronounced on CIFAR10→CIFAR100, where CLEAR is only 1.76 points below the best OOD AUROC and 4.41 points above the best adversarial coverage, yet achieves an AAUROC of 71.93%, compared with only 12.22% for C-EDL, the method attaining the highest OOD AUROC. Figure 2 therefore illustrates that CLEAR's occasional losses tend to be comparatively small, while the complementary performance sacrificed by the corresponding best-performing alternatives can be substantially larger. This motivates evaluating uncertainty methods jointly across ID retention, OOD detection, and adversarial detection rather than selecting them according to a single favourable metric.

Behaviour on the more challenging benchmarks. The Oxford Flowers and ImageNet experiments expose several operating regimes that are less apparent on the smaller benchmarks. On Oxford Flowers→DeepWeeds, R-EDL achieves the highest OOD AUROC, but its adversarial separation is considerably weaker than CLEAR. PostNet reports an adversarial coverage close to zero, but this accompanies an ID accuracy of only 5.92% and ID coverage of 14.30%, indicating degradation of the underlying predictor rather than useful selective uncertainty. The low-shot nature of Oxford Flowers likely makes class-conditioned density estimation particularly difficult, highlighting the importance of interpreting coverage together with predictive performance.

On ImageNet→CUB, the much larger 1000-class output space produces more pronounced extremes. I-EDL accepts almost all ID and OOD samples, whereas S-EDL and C-EDL reject almost all samples, including ID inputs. These behaviours likely reflect increased sensitivity of their uncertainty mechanisms at this scale: the substantially larger and more heterogeneous representation and output spaces can alter the magnitude and distribution of evidential, smoothing, and conflict-based uncertainty when method-specific settings are transferred unchanged from smaller benchmarks. These operating

points can therefore produce favourable values for individual coverage metrics while providing limited practical selectivity. We deliberately retain common method-specific settings across datasets rather than independently tuning each approach for every benchmark, so these results also reveal differences in how the methods transfer to larger-scale settings. CLEAR avoids either extreme, retaining substantial ID coverage while achieving the highest OOD and adversarial AUROC.

Applying CLEAR to conventional pretrained models. A practical limitation of CLEAR is that it requires an evidential model from which both a latent representation and evidential output can be obtained. Consequently, when only a conventional pretrained model is available, directly applying CLEAR may otherwise require retraining the predictor using an evidential objective and evidential head, which can be computationally prohibitive for large models or datasets. A lightweight alternative is to use an Evidential Meta Model (EMM) (Shen et al., 2023), which learns an evidential output from intermediate representations of a frozen pretrained model. This allows the original predictor to remain unchanged while providing the evidential representation required by CLEAR. Depending on the task, such a meta-model can parameterise an appropriate evidential distribution, such as a Dirichlet distribution for classification or a Normal-Inverse-Gamma distribution for regression. Our object-detection experiments, seen in Appendix C.6, demonstrate this strategy in practice: EMM is trained over frozen YOLO11n features to model bounding-box localisation uncertainty, after which CLEAR is applied post-hoc to the resulting evidential model. The resulting improvements over EMM alone show that CLEAR can therefore be used with existing pretrained predictors without requiring the underlying model itself to be retrained as an evidential network.

Regression coverage trade-offs. The lower ID coverage observed for CLEAR on some regression benchmarks arises because its uncertainty adjustment is driven by support in the calibrated latent geometry. Samples that are only weakly supported by this geometry receive increased uncertainty, which can include difficult or atypical ID examples as well as genuinely shifted inputs. In addition, the operating threshold is selected to maximise ID-OOD separation rather than to enforce a fixed ID retention rate, so improved rejection of OOD and adversarial inputs can be accompanied by reduced ID coverage. Importantly, CLEAR preserves the underlying regression prediction, meaning that this effect reflects a change in selective uncertainty rather than degraded predictive accuracy.

E Experimental Setup

This section describes our experimental setup in detail to support reproducibility and clarify the evaluation protocol. We first introduce the datasets used in our experiments, followed by the model architectures considered. We then describe the comparative methods and training procedure before presenting the task-specific configurations and implementation details used across each experimental setting.

E.1 Datasets

We evaluate CLEAR across a diverse set of established benchmarks spanning different domains and levels of task complexity. This enables a comprehensive assessment under varied experimental conditions. Dataset-specific details, including sample counts, input resolution, class composition, and data splits, are provided below, with representative examples shown in Figure 12.

- **MNIST (LeCun et al., 1998):** contains 28x28 greyscale images of handwritten digits from 0-9, spanning 10 classes. We use a split of 60,000 training samples, 8000 test samples, and 2000 validation samples. Pixel values are normalised to $[0, 1]$. In our experiments, MNIST serves as one of the ID datasets. It is paired with FashionMNIST and KMNIST as far-OOD datasets because they contain no overlapping classes while retaining similar visual characteristics, including greyscale appearance, low resolution, and a hand-drawn style. MNIST is also paired with EMNIST as a near-OOD dataset because the two datasets share similar visual characteristics and overlapping classes, with EMNIST containing handwritten digits in addition to handwritten letters.
- **FashionMNIST (Xiao et al., 2017):** contains 28x28 greyscale images of clothing items such as coats, bags, and t-shirts, spanning 10 classes. We use the widely adopted standard split of 60,000 training samples, 8000 test samples, and 2000 validation samples. Pixel



(a) MNIST



(b) FMNIST



(c) KMNIST



(d) EMNIST



(e) MNIST-C



(f) CIFAR10



(g) CIFAR100



(h) CIFAR-10-C (severity 1)



(i) CIFAR-10-C (severity 5)



(j) SVHN



(k) Oxford Flowers



(l) Deep Weeds



(m) Caltech-UCSD Birds-200-2011 (CUB)



(n) ImageNet

MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
1.6812	25	4.1922	1.0223	1392	3.8774	36.06	-119.01	0.47700
2.5313	30	5.0394	1.1935	1565	2.6798	35.14	-119.46	0.45800
3.4801	52	3.9772	1.1859	1310	1.3603	37.80	-122.44	5.00001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(o) California Housing

Fix. acid.	Vol. acid.	Citric	Sugar	Chlor.	F-SO_2	T-SO_2	Density	pH	Sulph.	Alcohol	Quality
7.0	0.17	0.74	12.8	0.045	24	126	0.99420	3.26	0.38	12.2	8
7.7	0.64	0.21	2.2	0.077	32	133	0.99560	3.27	0.45	9.9	5
6.8	0.39	0.34	7.4	0.020	38	133	0.99212	3.18	0.44	12.0	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(p) Wine Quality

Frequency	Attack angle	Chord length	Flow velocity	Displacement thickness	Sound pressure
400	0.0	0.3048	31.7	0.003313	125.045
1250	12.3	0.1016	31.7	0.041876	118.767
2500	4.0	0.3048	39.6	0.005796	120.233
⋮	⋮	⋮	⋮	⋮	⋮

(q) Airfoil Self-Noise



(r) UTKFace



(s) CelebA

Figure 12: Example images and tables from the datasets used in our experimental evaluation.

values are normalised to $[0, 1]$. In our experiments, FashionMNIST is used as a far-OOD pairing with the ID MNIST dataset.

- **KMNIST (Clanuwat et al., 2018):** contains 28x28 greyscale images of handwritten Japanese characters, specifically Kuzushiji characters from classical Japanese literature, spanning 10 classes. We use the widely adopted standard split of 60,000 training samples,

8000 test samples, and 2000 validation samples. Pixel values are normalised to $[0, 1]$. In our experiments, KMNIST is used as a far-OOD pairing with the ID MNIST dataset.

- **EMNIST (Cohen et al., 2017):** contains 28x28 greyscale images of handwritten Latin letters and digits from 0-9, spanning 47 classes. We use a split of 112,800 training samples, 15,040 test samples, and 3760 validation samples. Pixel values are normalised to $[0, 1]$. In our experiments, EMNIST is used as a near-OOD pairing with the ID MNIST dataset because it shares similar visual characteristics and overlapping handwritten digit classes.
- **MNIST-C (Mu & Gilmer, 2019):** contains the same data points as the original MNIST dataset, with images corrupted using 15 different distributional shifts such as shot noise, glass blur, and zigzag artefacts. In our experiments, MNIST-C is used as a near-OOD pairing with the ID MNIST dataset because it contains the same underlying data under distributional shift.
- **CIFAR10 (Krizhevsky et al., 2009):** contains 64x64 RGB images of real-world objects such as birds, trucks, airplanes, and frogs, spanning 10 classes. We use the widely adopted standard split of 50,000 training samples, 8000 test samples, and 2000 validation samples. Pixel values are normalised to $[0, 1]$ across all three RGB channels. In our experiments, CIFAR10 serves as one of the ID datasets. CIFAR10 is paired with FashionMNIST and SVHN as far-OOD datasets because they contain no overlapping classes while exhibiting similar visual characteristics, including coloured appearance and real-world imagery. CIFAR-10 is also paired with CIFAR-100 as a near-OOD dataset because the two datasets share similar visual characteristics and overlapping classes. Specifically, CIFAR-10 contains broad object categories such as cat, dog, truck, and ship, which correspond to finer-grained classes such as house cat, beagle, pickup truck, and cruise ship in CIFAR-100.
- **CIFAR100 (Krizhevsky et al., 2009):** contains 64x64 RGB images of real-world objects spanning 100 classes, including fine-grained counterparts of classes found in CIFAR10. For example, CIFAR10 contains the class *truck*, whereas CIFAR100 contains classes such as *pickup truck* and *train*. We use the widely adopted standard split of 50,000 training samples, 8000 test samples, and 2000 validation samples. Pixel values are normalised to $[0, 1]$ across all three RGB channels. In our experiments, CIFAR100 is used as a near-OOD pairing with the ID CIFAR10 dataset because the datasets share similar visual characteristics and overlapping classes.
- **CIFAR-10-C (Hendrycks & 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 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.
- **SVHN (Netzer et al., 2011):** contains 64x64 RGB images of house numbers collected from Google Street View, spanning 10 digit classes from 0-9. We use a split of 73,257 training samples, 10,832 test samples, and 5200 validation samples. Pixel values are normalised to $[0, 1]$ across all three RGB channels. In our experiments, SVHN is used as a far-OOD pairing with the ID CIFAR10 dataset due to the substantial domain shift between street-level digit photographs and object-centric natural images.
- **Oxford Flowers (Nilsback & Zisserman, 2008):** contains 224x224 RGB images of flowers commonly found in the United Kingdom, spanning 102 classes such as Water Lily and Wild Pansy. We use a split of 4913 training samples, 819 test samples, and 819 validation samples. Pixel values are normalised to $[0, 1]$ across all three RGB channels. In our experiments, Oxford Flowers is used as an ID dataset for classification, and as an OOD pairing for the UTKFace regression task. It is paired with Deep Weeds because the two datasets contain no overlapping classes while sharing similar visual characteristics, including natural features such as leaves and flowers.
- **Deep Weeds (Olsen et al., 2019):** contains 224x224 RGB images of weed species found in Australia, spanning 8 classes such as Snake weed and Rubber vine. We use the standard split of 10,505 training samples, 3502 test samples, and 3502 validation samples. Pixel values are normalised to $[0, 1]$ across all three RGB channels. In our experiments, 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.
- **California Housing (Pace & Barry, 1997):** contains 20,640 California district records with eight numerical input features and median house value as the regression target. We use it as

an ID dataset and evaluate a target-shift setting in which districts with extreme house prices form the OOD subset.

- **Wine Quality (Cortez et al., 2009):** contains physicochemical measurements of red and white wines, with eleven numerical input features and a sensory quality score as the regression target. We train on white-wine samples and use red-wine samples as OOD inputs, testing whether uncertainty reflects a change in the input distribution while the prediction target remains wine quality.
- **Airfoil Self-Noise (Brooks et al., 1989):** contains 1,503 observations with five numerical input features describing airfoil operating conditions and scaled sound pressure level as the regression target. We use it to assess covariate shift, defining OOD samples by extreme values of the input frequency rather than by the regression target.
- **UTKFace (Zhang et al., 2017):** contains over 20,000 face images with age annotations spanning 0-116 years. We use age prediction as an image regression task and evaluate two OOD settings: Oxford Flowers images, which introduce a substantial image-domain shift, and faces at age extremes, which remain within the face domain but differ in target distribution.
- **CelebA Liu et al. (2015):** consists of 178×218 aligned RGB images of celebrity faces, spanning a single object-detection class (face). We utilise the standard split of 162,770 training images, 19,867 calibration images, and 19,962 test images. Images were resized to 640×640 pixels, with pixel normalisation bounded $[0, 1]$ across all three RGB channels.
- **ImageNet (Deng et al., 2009):** consists of 224x224 RGB images across 1000 object classes, each corresponding to a subset of ImageNet. The dataset contains 1,281,167 training samples, 50,000 validation samples, and 100,000 test samples without publicly available labels. Pixel values were normalised to the range $[0, 1]$ across all RGB channels. In our experimental evaluation, ImageNet is used as an ID dataset. ImageNet was paired with Caltech-UCSD Birds-200-2011 because they share very few overlapping classes but exhibit similar visual characteristics.
- **Caltech-UCSD Birds-200-2011 (CUB) (Welinder et al., 2010):** consists of 224x224 RGB images of 200 bird species, collected from various natural environments. The dataset contains a total of 11,788 images, with a standard split of 5,994 training samples and 5,794 test samples. Pixel values were normalised to the range $[0, 1]$ across all RGB channels. In our experimental evaluation, CUB is used as an OOD dataset paired with ImageNet.

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.

E.2 MODELS

To assess the performance and generalisability of CLEAR, we assess its performance against multiple models and architectures across our experimental evaluation. Below, we provide a concise description of each model together with the implementation details and hyperparameter settings adopted in our experiments:

- **ResNet-18 (He et al., 2016):** is used for the MNIST classification and image-based regression experiments. We train it from scratch, setting the initial convolution to stride one and removing the initial max-pooling layer for small images. The task-specific head is trained for 10 epochs, followed by 40 epochs of backbone training. The resulting model is frozen before CLEAR is applied.
- **EfficientNet-B0 (Tan & Le, 2019):** is used for the CIFAR experiments. We initialise the TorchVision model from ImageNet-1K pretrained weights and set its first convolution to stride one for small images. The task-specific head is trained for 10 epochs, followed by 40 epochs of backbone fine-tuning. The resulting model is frozen before CLEAR is applied.
- **DenseNet-121 (Huang et al., 2017):** is used for the Oxford Flowers and DeepWeeds experiments. We initialise the TorchVision model from ImageNet-1K pretrained weights and train it on the in-distribution Oxford Flowers task, with 10 epochs of head training followed by 40 epochs of backbone fine-tuning. The resulting model is frozen before CLEAR is applied.

- **ViT-B/16 (Dosovitskiy et al., 2021; Vaswani et al., 2017):** is used for the ImageNet and CUB experiments. We initialise the TorchVision model from ImageNet-1K pretrained weights and train its task-specific head for 10 epochs on ImageNet, keeping the backbone frozen. The resulting model is frozen before CLEAR is applied.
- **Tabular MLP:** is used for the tabular regression experiments. It comprises two fully connected layers of width 64, each followed by a ReLU activation. Its learned 64-dimensional representation is used by the regression head and CLEAR.
- **YOLO11n (Khanam & Hussain, 2024):** is used as the base object detector for the object detection experiments. We initialise the lightweight nano variant from COCO-pretrained weights and finetuned for 20 epochs for CelebA. Following fine-tuning, all detector parameters are frozen during CLEAR.

E.3 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 CLEAR alongside the baseline methods.

- **Differential Entropy:** quantifies the dispersion, or uncertainty, represented by the Dirichlet distribution. It is defined as:
$$\sum_{k=1}^{K} \ln \Gamma(\alpha_k) - \ln \Gamma(S) - \sum_{k=1}^{K} (\alpha_k - 1)(\Psi(\alpha_k) - \Psi(S))$$
where $B(\boldsymbol{\alpha})$ denotes the multivariate Beta function, S is the total concentration parameter, and $\Psi(\cdot)$ denotes the digamma function. Higher entropy values correspond to greater uncertainty and are therefore typically associated with OOD inputs. This score is applicable only to Dirichlet-based uncertainty models, such as posterior networks and evidential networks.
- **Predictive Entropy:** quantifies uncertainty in the model's predictive class distribution. Higher values indicate greater uncertainty, with the maximum attained when all classes are equally probable. This score is applicable to any model that provides predictive class probabilities.
- **Mutual Information:** measures epistemic uncertainty by quantifying how much information the model parameters contribute to the predictive distribution. It is defined as the difference between the predictive entropy and the expected conditional entropy under the posterior over model 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 $q(\omega)$ denotes the explicit or implicit posterior over model parameters.
- **Maximum Probability:** provides a simple confidence score based on the model's most probable prediction. It is defined as the maximum softmax probability across all classes:
$$\max_{k \in \{1,\ldots,K\}} \; p(y = k \mid x)$$
Higher values indicate greater confidence in the predicted class, whereas lower values indicate greater uncertainty.

Following prior work (Shen et al., 2023; Barker et al., 2026), the optimal ID-OOD threshold is determined using the validation datasets. For a selected scoring metric from those described above, we first compute scores for both ID and OOD validation samples. These scores are then used to construct a receiver operating characteristic (ROC) curve, treating ID samples as positive and OOD samples as negative. The ROC curve characterises how effectively the selected metric separates ID and OOD data. The decision threshold is chosen to maximise TPR $-$ FPR, corresponding to the point of optimal separation. This criterion provides a principled balance between retaining ID inputs and rejecting OOD inputs, avoids ad hoc threshold tuning, and enables clear reporting of ID and OOD coverage under deployment-like conditions.

Following calibration, the selected threshold is fixed and applied to the corresponding 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.

E.4 Comparative Approaches

To evaluate the effectiveness of CLEAR, we compare it against a range of recent EDL and uncertainty quantification approaches that are representative of the current state-of-the-art. Each approach is briefly described below, alongside any implementation-specific details or hyperparameter choices used in our experiments:

- **Posterior Networks (Charpentier et al., 2020):** provide a foundational approach to uncertainty estimation in classification that predates and informs later evidential deep learning methods. They avoid requiring OOD data during training by directly modelling a closed-form posterior over categorical distributions. Specifically, the model predicts Dirichlet concentration parameters $\alpha = \beta_{\text{prior}} + \beta$ for each input x, where β denotes pseudo-counts obtained from class-conditional density estimates in a learned latent space and β_{prior} is a fixed symmetric prior. The resulting Dirichlet distribution represents both aleatoric and epistemic uncertainty in closed form.
- **Evidential Deep Learning (EDL) (Sensoy et al., 2018):** provides a deterministic alternative to Bayesian neural networks for uncertainty estimation by modelling class probabilities through a Dirichlet distribution whose parameters are derived from the model's non-negative outputs. The model is trained using an objective that combines squared prediction error, predictive variance, and a KL divergence to a uniform prior. This evidential formulation enables both aleatoric and epistemic uncertainty to be quantified in closed form without requiring sampling or OOD examples during training. EDL serves as the base approach for the subsequent comparative methods. As the base method, its hyperparameters are shared across the remaining comparative approaches to ensure a fair comparison and are described in Appendix E.6.
- **Fisher Information-Based Evidential Deep Learning ($\mathcal{I}$-EDL) (Deng et al., 2023):** extends EDL by incorporating the Fisher Information Matrix (FIM) to adaptively weight the loss according to the informativeness of the predicted evidence. The central idea is that classes associated with greater evidence contain less Fisher information and should therefore be regularised less strongly. The method introduces a regularisation term based on the log-determinant of the FIM to discourage overconfident predictions:

  $$\mathcal{L}_i^{|\mathcal{I}|} = \sum_{j=1}^{K} \log \psi^{(1)}(\alpha_{ij}) + \log\left(1 - \frac{\psi^{(1)}(\alpha_{i0})}{\sum_{j=1}^{K} \psi^{(1)}(\alpha_{ij})}\right)$$

  In our experiments, we set $\lambda_{|\mathcal{I}|} = 0.001$ and anneal it linearly over the first 10 epochs, following the guidance in Appendix C.2 of the original paper together with our manual tuning.
- **Hyper-Opinion Evidential Deep Learning (H-EDL) (Qu et al., 2024):** extends classical EDL through hyper-opinions, which generalise multinomial opinions from Subjective Logic by representing uncertainty over both singleton classes and composite subsets. This enables H-EDL to capture both sharp evidence, which supports a single class, and vague evidence, which supports multiple plausible classes, thereby improving robustness to ambiguous inputs. The method introduces an opinion projection mechanism that maps hyper-opinions to standard Dirichlet-based predictions, allowing training within the conventional EDL framework. This projection also addresses the vanishing gradient problem that can limit traditional EDL on fine-grained tasks. In our experiments, we adopt the two-stage training procedure described in the original paper without introducing additional hyperparameters.
- **Relaxed Evidential Deep Learning (R-EDL) (Chen et al., 2024):** addresses overconfidence in standard EDL by relaxing two nonessential assumptions. First, it replaces the fixed prior weight, typically equal to the number of classes, with a tunable scalar hyperparameter λ that controls the contribution of the base rate α in the Dirichlet construction:

  $$\alpha(x) = e(x) + \lambda$$

  Second, R-EDL removes the variance-penalising regularisation term from the EDL objective and instead directly optimises the projected class probabilities $P(x) = \frac{\alpha(x)}{S}$ against the one-hot labels. This simplification provides a better balance between evidence magnitude and evidence proportion, particularly in OOD settings. In our experiments, we follow the guidance from Appendix C.2 and Figure 1(b) of the original paper and use $\lambda = 0.1$ as a fixed prior weight throughout training.

- **Regularized Evidential Model (RED) (Pandey & Yu, 2023):** addresses the inability of conventional evidential models to learn effectively from samples mapped to zero-evidence regions. RED introduces a vacuity-guided correct-evidence regularisation term that encourages greater evidence for the ground-truth class, with its influence strongest when evidence is low. This enables the model to learn from samples that would otherwise produce negligible gradient updates while retaining the standard evidential learning objective.
- **Smoothed Evidential Deep Learning (S-EDL) (Kopetzki et al., 2021):** improves the robustness of standard evidential models by applying median smoothing to their uncertainty estimates. Given an input x, a collection of noisy samples $\{x_s\} \sim \mathcal{N}(x, \sigma)$ is generated. The final uncertainty estimate is obtained by taking the median across these samples, improving robustness to adversarial perturbations by reducing the influence of outliers. In our experiments, we follow the original setup and use a Gaussian noise scale of $\sigma = 0.01$ with 50 samples per input, based on guidance from the original paper and our manual tuning. S-EDL is a purely post-hoc method.
- **Conflict-aware Evidential Deep Learning (C-EDL) (Barker et al., 2026):** is a lightweight post-hoc extension of EDL that improves OOD and adversarial robustness without retraining the base model. C-EDL generates multiple label-preserving transformed views of an input and measures disagreement between their evidential outputs using intra-class variability and inter-class conflict. The resulting conflict score is used to reduce evidential strength when disagreement is high, increasing uncertainty while largely preserving predictions for consistent ID inputs.
- **Deep Evidential Regression (DER) (Amini et al., 2020):** extends evidential learning to continuous prediction by placing a Normal-Inverse-Gamma (NIG) distribution over the mean and variance of a Gaussian likelihood. A single network predicts the NIG parameters $(\gamma, \nu, \alpha, \beta)$, where γ represents the predictive mean and the remaining parameters determine the evidential strength and associated aleatoric and epistemic uncertainty. Training combines the negative log-likelihood of the resulting Student-t predictive distribution with an error-dependent evidence regulariser, which penalises high evidence when the prediction is inaccurate. This enables both forms of uncertainty to be obtained analytically from a single forward pass without requiring sampling or OOD examples during training. The setup is described in Section 3. In our experiments, we use an evidential regularisation weight of $\lambda_{\text{reg}} = 10^{-3}$.
- **Uncertainty Regularised Evidential Regression Network (UR-ERN) (Ye et al., 2024):** extends DER by addressing the high uncertainty area (HUA), in which the constrained activations used to ensure valid evidential parameters can produce vanishing gradients and prevent highly uncertain samples from contributing effectively to learning. UR-ERN introduces an additional uncertainty regularisation term that maintains a learning signal within this region, with its influence increasing with the discrepancy between the predicted value and ground truth. This allows the model to move samples away from poorly learned high-uncertainty regions while retaining the original evidential regression objective and NIG uncertainty formulation. In our experiments, we use an uncertainty regularisation weight of 10^{-3}.
- **Evidential Meta Model (EMM) (Shen et al., 2023):** is a post-hoc uncertainty approach that combines representations from multiple frozen intermediate layers using a lightweight evidential meta-model. We adapt EMM to object detection by applying ROIAlign to predefined detector features for each predicted bounding box and parameterising a Normal-Inverse-Gamma distribution over its four localisation coordinates. The meta-model is trained for 20 epochs using Adam with an initial learning rate of 10^{-2}, cosine annealing to 10^{-6}, an IoU matching threshold of 0.5, evidential regularisation weight $\lambda_{\text{reg}} = 0.1$, and gradient clipping at 1.0.

E.5 Adversarial Attacks

This section presents the adversarial attack strategies used to assess the robustness of uncertainty-aware models. Adversarial attacks introduce carefully constructed, often imperceptible perturbations that can cause a model to misclassify while remaining highly confident. These attacks are particularly challenging for uncertainty-aware methods because they can disrupt both aleatoric and epistemic uncertainty signals, causing adversarial inputs to appear in-distribution or evade uncertainty-based detection.

To provide a comprehensive robustness evaluation, we consider both gradient-based and gradient-free attacks. Gradient-based attacks represent white-box settings, where access to model gradients is used to construct adversarial perturbations. Conversely, gradient-free attacks apply random or structured perturbations without relying on gradient information and more closely reflect black-box settings. Where supported, attacks are implemented using Foolbox (Rauber et al., 2017) under the following configurations:

- L_2 **Projected Gradient Descent (L2PGD) (Madry et al., 2018):** 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) (Goodfellow et al., 2015):** 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 (Croce & Hein, 2020):** is an ensemble of complementary adversarial attacks comprising Auto-PGD with cross-entropy and difference-of-logits-ratio losses, the Fast Adaptive Boundary attack, and the black-box Square Attack. Adversarial perturbations are constrained by $\|x' - x\|_2 \leq \epsilon$. The attacks are applied sequentially to inputs that remain correctly classified, combining gradient-based and query-based strategies to evaluate robustness without model-specific hyperparameter tuning.

E.6 Training Details

To ensure that differences between uncertainty methods are not driven by method-specific optimisation, we use a common training configuration wherever possible. All models are trained with a batch size of 64 using the Adam optimiser with weight decay 10^{-4}. For MNIST, CIFAR-10, and Oxford Flowers, training follows the two-stage procedure described above: the task-specific head is first trained with the feature extractor frozen, followed by joint fine-tuning of the backbone and head. The head is initially trained with a learning rate of 10^{-3}; during fine-tuning, the backbone and head use learning rates of 10^{-5} and 10^{-4}, respectively. For ImageNet, the ImageNet-1K pretrained ViT-B/16 backbone remains frozen and only the task-specific head is trained, using a learning rate of 10^{-4}. The tabular MLP uses a learning rate of 10^{-3} throughout. Learning rates are reduced by a factor of 0.5 when validation loss does not improve for five epochs, with a minimum learning rate of 10^{-8}. A dropout rate of 0.25 is used in the shared predictive architecture. These optimisation settings are kept fixed across comparative approaches for a given dataset, with method-specific parameters varying only where required by the corresponding uncertainty formulation.

E.7 Runtime and Complexity

All experiments were implemented in PyTorch 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 or used pre-trained weights. All runs used random seeds.

CLEAR introduces a one-off post-hoc calibration stage followed by a lightweight adjustment at inference. During calibration, estimating the group-conditioned covariance structure over the d-dimensional latent representations requires $\mathcal{O}(|\mathcal{D}_{\text{cal}}|d^2)$ operations. Computing the Cholesky factor and latent scale for each group contributes $\mathcal{O}(Kd^3)$ for classification, or $\mathcal{O}(Bd^3)$ for regression. Estimating the reference conflict additionally evaluates T perturbation draws for each calibration

example, giving $\mathcal{O}(|\mathcal{D}_{\text{cal}}|T(Vd^2 + V^2d))$. The resulting calibration complexity is therefore:

$$\mathcal{O}\big(|\mathcal{D}_{\text{cal}}|d^2 + Kd^3 + |\mathcal{D}_{\text{cal}}|T(Vd^2 + V^2d)\big) \tag{44}$$

for classification, with K replaced by B for regression. This cost is incurred only once after training.

At inference, CLEAR requires only the original forward pass through the frozen evidential model, since the V perturbation views are generated directly in latent space. Whitening these views using the calibrated Cholesky factor requires $\mathcal{O}(Vd^2)$ operations. The $V(V-1)/2$ unordered view pairs are then evaluated in $\mathcal{O}(V^2d)$ time, since the cross-dimensional conflict in Equation 14 can be computed using sums over the d latent dimensions rather than explicitly evaluating all $d(d-1)$ interactions. The additional CLEAR inference complexity is therefore:

$$\mathcal{O}\big(Vd^2 + V^2d\big) \tag{45}$$

With the default $V = 3$, this reduces to $\mathcal{O}(d^2)$ additional computation per prediction.

The empirical inference times in Table 4 are consistent with this analysis. Across the evaluated benchmarks, CLEAR remains close to the corresponding single-pass EDL inference time. On MNIST→FashionMNIST, for example, inference increases from 1.96 s for EDL to 2.36 s for CLEAR, while C-EDL and S-EDL require 13.15 s and 57.56 s, respectively. The difference becomes particularly pronounced for larger inputs: on Oxford Flowers, CLEAR requires 18.69 s compared with 18.48 s for EDL, 74.46 s for C-EDL, and 660.19 s for S-EDL. Similarly, on ImageNet, CLEAR requires 423.66 s compared with 397.02 s for EDL and 7356.02 s for S-EDL. These results show that operating directly on the latent representation introduces a relatively small overhead compared with the frozen base model while avoiding the repeated full-network evaluations required by more computationally expensive post-hoc approaches.