FROM DISTRIBUTION TO SAMPLE: QUANTIFYING UNCERTAINTY FOR HETEROGENEOUS DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty quantification (UQ) becomes particularly challenging under heterogeneous distributions. In practice, it can be often observed that many UQ methods exhibit calibration degradation as the proportion of heterogeneous samples increases. We argue that this limitation arises mainly because covariate shift and out-of-distribution (OOD) are only defined with regard to distributions, failing to provide a direct criterion for assessing individual inputs. To address this, we propose a sample-based lightweight plug-in named Dual-signal Coverage-aware UQ (D-CUQ). D-CUQ leverages two key metrics: the Manifold-deviation Signal, capturing where an input lies relative to the ID manifold, and the Coverage-gap Signal, capturing whether the local tangent geometry around it is spanned by training data. We further establish a formal connection between the two levels, showing that distribution-level quantities arise as aggregates of their sample-level counterparts. Extensive experiments on classification and regression benchmarks demonstrate that D-CUQ significantly improves uncertainty estimation for heterogeneous test distributions.

1 INTRODUCTION

Much of the success of machine learning relies on the implicit assumption that training and test distributions are nearly identical (James et al., 2013). In practice, this assumption is routinely violated by complex and unpredictable heterogeneous distributions, which we define as the mixture of in-distribution (ID), covariate shift, out-of-distribution (OOD) samples (Liu et al., 2021a; Simchowitz et al., 2023). Such distributions frequently arise in real-world settings, such as noisy sensory recordings during inference, evolving user requirements, or entirely new objectives unforeseen at training time (Yang et al., 2023; Zhang et al., 2024; Gao et al., 2024).

Calibration aligns predicted confidence with true probabilities, serving as a key criterion for reliable UQ. UQ has matured into a set of standard estimators with established baselines and public implementations (Gawlikowski et al., 2023). Despite progress in methodologies (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), guaranteeing reliability across heterogeneous distributions remains a formidable challenge, given that prior work primarily targets simple distribution shifts (Hu et al., 2024; Slavutsky & Blei, 2026). Fig. 1 reports the averaged calibration results of 4 state-of-the-art UQ methods (MC Dropout (Gal & Ghahramani, 2016), Deep Ensemble (Lakshminarayanan et al., 2017), SWAG(Maddox et al., 2019), Temperature Scaling (Guo et al., 2017)) on a heterogeneous mixture of different datasets, with backbone details provided in the Appendix. Notably, we observed that Expected Calibration Error (ECE) (Guo et al., 2017) and Brier (Brier, 1950) exhibit degradation as ID samples are progressively replaced by covariate-shifted or semantic OOD samples.

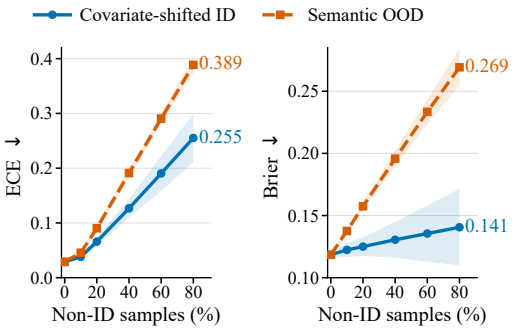

Figure 1: Calibration degradation under heterogeneous distributions. Averaged ECE and Brier score increase with the replacement of ID samples by covariate-shifted or semantic OOD samples.

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.

We argue this limitation arises mainly because non-ID concepts, such as covariate shift and semantic OOD, are currently defined at the distribution level (Shimodaira, 2000; Sugiyama et al., 2007; Zhang et al., 2024), rather than from a sample based perspective. To the best of our knowledge, these definitions concern a distribution that has no closed form, and thus it is difficult to determine how an isolated point relates to it. Revisiting the question at the sample level avoids this, but leaves us without a representation of where ID data lies and without a rule for separating covariate shift from semantic OOD. Specifically, uncertainty varies across forms of heterogeneity, making a unified distribution-level treatment of UQ inherently inaccurate.

To this end, we reformulate the heterogeneous problem from a distribution-based to a sample-based perspective, building on the manifold hypothesis. The representations of each ID class concentrate near a low-dimensional manifold which a finite training set can only partially cover (Narayanan & Mitter, 2010; Bengio et al., 2013). Consequently, each sample is characterized by two non-equivalent quantities: its distance to the manifold, and the local coverage provided by the training samples (Mukhoti et al., 2023; Bui & Liu, 2024).

Under this view, we propose **Dual-signal Coverage-aware UQ (D-CUQ)**, a lightweight plug-in provides each input with an individualized calibration to the predictive distribution of an existing UQ method. Then, we characterize these two signals in closed form and further theoretically establish a formal connection between them. Specifically, the **manifold-deviation signal**, defined via the class-conditional density to capture how far an input deviates from the ID manifold. The **coverage-gap signal**, formulated as a posterior variance that quantifies the sufficiency with which the local tangent geometry around the input is covered by the training data. The contributions of this paper are summarized as follows:

- We identify calibration degradation under heterogeneous distributions and propose D-CUQ, a plug-in module that combines manifold-deviation and coverage-gap signals to adjust the predictive distribution of existing methods without retraining the backbone.
- By reformulating the heterogeneous problem from a distribution-based to a sample-based perspective, we establish a formal connection between the two levels, revealing that distribution-level quantities are simply aggregates of their sample-level counterparts.
- Extensive experiments on classification, regression and further discussion benchmarks validate our theoretical derivation. The results show that D-CUQ lowers the average ECE from 0.37 to 0.26 in classification and raises the macro AUROC from 0.66 to 0.84 in regression.

2 Related Work

Predictive uncertainty. Predictive uncertainty methods are commonly grouped by the number of networks used and whether they are stochastic (Gawlikowski et al., 2023), which is largely a difference in inference cost. MC Dropout samples forward passes from one trained network (Gal & Ghahramani, 2016), SWAG fits a Gaussian to the SGD iterates and averages over its samples (Maddox et al., 2019), and Deep Ensembles trains several networks independently (Lakshminarayanan et al., 2017). Sampling is avoided by low-overhead variants of the Laplace approximation (Daxberger et al., 2021), by variational Bayesian last layers, which restrict inference to the output layer (Harrison et al., 2024), and by evidential methods, which place a Dirichlet distribution on the class probabilities (Sensoy et al., 2018).

Uncertainty calibration. When it comes to uncertainty calibration, temperature scaling rescales the logits of a trained classifier post hoc (Guo et al., 2017), and uncertainty-weighted gradients act during training instead (Lin et al., 2025). In regression, Kuleshov et al. (2018) recalibrate predicted quantiles so that a stated interval level matches its empirical coverage. Ovadia et al. (2019) evaluated such methods under dataset shift at scale and found that post-hoc calibration degrades as the shift grows. However, the problem becomes more pronounced under heterogeneous distributions, which existing methods have not considered.

UQ under distribution shift. Under distribution shift, recent research notices that standard architectures are sensitive to the circumstance, whether from common corruptions or from shifts in subpopulation (Hendrycks & Dietterich, 2019; Santurkar et al., 2021). The uncertainty they report

degrades along with their accuracy (Ovadia et al., 2019). One line of work responds using unlabeled data from the shifted distribution. Park et al. (2020) combine importance weighting with a learned feature map that equalizes the two distributions, Chan et al. (2020) regularize the posterior using unlabeled target data as pseudo-labels of model confidence, and PseudoCal, a post-hoc framework, turns the problem into a supervised one by synthesizing a labeled pseudo-target set through inference-stage mixup (Hu et al., 2024). A second line uses in-distribution data only, as in Density-Softmax, which couples the softmax layer with a density built on a Lipschitz-constrained feature extractor (Bui & Liu, 2024), and DAC, which derives a sample-wise temperature from k-NN feature densities (Tomani et al., 2023).

However, the first line requires unlabeled data from a single target domain and the second requires a training-time constraint or targets one shift type at a time, neither of which is available for a heterogeneous distribution handled by an already-trained predictor.

3 Proposed Method

In this section, we first reformulate the heterogeneous problem from a distribution-based to a sample-based perspective, then construct two signals that measure positional mismatch and directional undercoverage on frozen representations, and finally propose our method.

3.1 Problem Setting and Reformulation

Consider a predictor f_θ trained on $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^N$, drawn i.i.d. from a distribution P_{ID} over $\mathcal{X} \times \mathcal{Y}_{\text{ID}}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}_{\text{ID}} = \{1, \dots, K\}$ is the label set. Throughout, f_θ is frozen, and $z = h_\theta(x) \in \mathbb{R}^d$ and $\ell_\theta(x) \in \mathbb{R}^K$ denote the feature and the logits it produces at an input x. Both signals below are functions of z only, so their construction is independent of the architecture of f_θ. Applying any UQ method to f_θ yields a predictive distribution $p_0(\cdot \mid x)$ over $\mathcal{Y}_{\text{ID}}$, which D-CUQ takes as given and calibrate.

We say a distribution Q over $\mathcal{X} \times \mathcal{Y}$ is covariate shift with respect to P_{ID} if $Q_X \neq P_{\text{ID},X}$ while $Q(y \mid x) = P_{\text{ID}}(y \mid x)$ for all x, and semantic OOD if its label set $\mathcal{Y}_Q$ satisfies $\mathcal{Y}_Q \cap \mathcal{Y}_{\text{ID}} = \emptyset$, where $\mathcal{Y} \supseteq \mathcal{Y}_{\text{ID}}$ is the ambient label space. A heterogeneous distribution is a mixture of different distributions together with P_{ID},

$$Q_{\text{mix}} = \pi_0 P_{\text{ID}} + \sum_{r=1}^{R_{\text{cov}}} \pi_r^{\text{cov}} Q_r^{\text{cov}} + \sum_{s=1}^{R_{\text{OOD}}} \pi_s^{\text{OOD}} Q_s^{\text{OOD}}, \tag{1}$$

where Q_r^{cov} and Q_s^{OOD} denote covariate-shifted and semantic OOD distributions, respectively, as defined above, with counts R_{cov} and R_{OOD}. The mixing proportions satisfy $\pi_0, \pi_r^{\text{cov}}, \pi_s^{\text{OOD}} \geq 0$ and $\pi_0 + \sum_r \pi_r^{\text{cov}} + \sum_s \pi_s^{\text{OOD}} = 1$. Every test input is drawn from one of the distributions in Eq.1, but cannot be determined (Liu et al., 2021a).

Definition 1 (Sample-based Heterogeneity)**.** Given thresholds $\varepsilon_{\mathcal{M}} > 0$ and $\tau_{\mathcal{S}} \in (0, 1)$, the regime of an input x is

$$\mathcal{R}(x) = \begin{cases} \text{ID}, & \delta_{\mathcal{M}}(x) \leq \varepsilon_{\mathcal{M}},\ U_{\mathcal{S}}(x) \leq \tau_{\mathcal{S}}, \\ \text{covariate-shift}, & \delta_{\mathcal{M}}(x) \leq \varepsilon_{\mathcal{M}},\ U_{\mathcal{S}}(x) > \tau_{\mathcal{S}}, \\ \text{semantic OOD}, & \delta_{\mathcal{M}}(x) > \varepsilon_{\mathcal{M}}. \end{cases} \tag{2}$$

By the manifold hypothesis (Bengio et al., 2013), the features of inputs whose labels lie in $\mathcal{Y}_{\text{ID}}$ do not fill $\mathbb{R}^d$ but concentrate near a low-dimensional set $\mathcal{M}_{\text{ID}} \subset \mathbb{R}^d$, which is determined by $\mathcal{Y}_{\text{ID}}$ and not by the input marginal. The manifold deviation is the distance from z to that set,

$$\delta_{\mathcal{M}}(x) = \text{dist}\big(z, \mathcal{M}_{\text{ID}}\big) = \inf_{a \in \mathcal{M}_{\text{ID}}} \|z - a\|_2. \tag{3}$$

The coverage deficiency is measured against a reference set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{D}_{\text{tr}}$, the record of where in-distribution inputs have been observed. $U_{\mathcal{S}}(x)$ is large where $\mathcal{S}$ has few samples near z and decreases as $\mathcal{S}$ grows.

We discuss how these coordinates relate to the distribution-level definitions and where the two differ in appendix.

3.2 COVERAGE-GAP SIGNAL

A finite reference set leaves some directions of the feature space sparsely populated, and an input may lie close to an ID class while pointing along one of them. We measure this with a regularized leverage quantity. Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ contain the standardized reference representations. For $\lambda > 0$,

$$\mathbf{A}_\lambda = \mathbf{Z}^\top \mathbf{Z} + \lambda I_d, v_\lambda(x) \quad = \lambda\, z^\top \mathbf{A}_\lambda^{-1} z, G_{\mathcal{S}}(x) = \frac{v_\lambda(x)}{\|z\|_2^2 + \epsilon}, \tag{4}$$

where λ is a regularization parameter and $\epsilon > 0$ ensures numerical stability. The factor $z^\top \mathbf{A}_\lambda^{-1} z$ is the regularized leverage associated with kernel ridge regression (Alaoui & Mahoney, 2015), and $v_\lambda(x)$ has the standard posterior-variance form under a kernel model with regularization λ (Rasmussen & Williams, 2006). The kernel is the one induced by the frozen linear head, whose last-layer empirical kernel has scalar factor $\kappa(x, x') = h_\theta(x)^\top h_\theta(x')$ (Jacot et al., 2018). Theorem 2 shows that $G_{\mathcal{S}}(x)$ takes values in $[0, 1]$ and does not increase when a reference sample is added, so we use it as the computable form of $U_{\mathcal{S}}(x)$.

Theorem 2 (Properties of the coverage-gap signal)**.** *Consider $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top$, $\mathbf{k}_x = \mathbf{Z}z$ and $\kappa_{xx} = z^\top z$, and let $\mathbf{Z}^\top \mathbf{Z} = \mathbf{U}\,\mathrm{diag}(\xi_1, \ldots, \xi_d)\mathbf{U}^\top$ be an eigendecomposition with $\mathbf{U} = [u_1, \ldots, u_d]$ orthogonal. Then for every $\lambda > 0$,*

$$v_\lambda(x) = \kappa(x, x) - \mathbf{k}_x^\top (\mathbf{K} + \lambda I_n)^{-1} \mathbf{k}_x = \sum_{j=1}^{d} \frac{\lambda}{\xi_j + \lambda} \big(u_j^\top z\big)^2, \tag{5}$$

and $0 \le G_{\mathcal{S}}(x) \le 1$. Moreover, if a further input x_+ is added to $\mathcal{S}$ and λ is unchanged, then

$$G_{\mathcal{S} \cup \{x_+\}}(x) \le G_{\mathcal{S}}(x), \tag{6}$$

with strict inequality whenever $z^\top \mathbf{A}_\lambda^{-1} h_\theta(x_+) \neq 0$.

Remark.

- **Labels play no role.** Eq.5 identifies the signal with the posterior variance of the kernel model induced by the frozen head, a quantity built from the reference inputs and never from their labels.
- **Proximity does not imply coverage.** The spectral form weights direction u_j by $\lambda/(\xi_j+\lambda)$, with ξ_j the reference energy along u_j, so an input close to the data is still flagged when it points along a weakly explored direction.
- **New data helps only along the direction it populates.** A new reference sample lowers the signal exactly when its representation is not $\mathbf{A}_\lambda^{-1}$-orthogonal to z, that is, when it populates a direction in which x is currently unsupported.

3.3 MANIFOLD-DEVIATION SIGNAL

An input may lie far from every ID class region in feature space, which is a different failure from lying near one but along a direction the reference set does not cover. We measure it by distance to the class means. Let $\mathcal{I}_k = \{i : y_i = k\}$ be the indices of the class-k samples of $\mathcal{S}$, let $z_i = h_\theta(x_i)$, and let $\mu_k = |\mathcal{I}_k|^{-1} \sum_{i \in \mathcal{I}_k} z_i$ be the class-k mean feature. Let

$$\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i, \qquad \nu_j = \frac{1}{n} \sum_{i=1}^{n} \big(z_{ij} - \overline{z}_j\big)^2, \qquad \mathbf{D}_\nu = \mathrm{diag}(\nu_1, \ldots, \nu_d) + \epsilon_\nu I_d \tag{7}$$

be a diagonal covariance tied across classes and estimated from all of $\mathcal{S}$, where z_{ij} is the j-th coordinate of z_i and $\epsilon_\nu > 0$ ensures numerical stability. Modeling each class by a Gaussian with a tied covariance turns the Mahalanobis distance to the nearest class mean into a confidence score (Lee et al., 2018). Such a density can be fitted in the feature space of an already trained network (Mukhoti et al., 2023). We accordingly define

$$M_{\mathcal{S}}(x) = \frac{1}{d} \min_{1 \le k \le K} \big(z - \mu_k\big)^\top \mathbf{D}_\nu^{-1} \big(z - \mu_k\big), \tag{8}$$

normalized by d so that its scale does not grow with the feature dimension, and use it as the computable form of $\delta_{\mathcal{M}}(x)$.

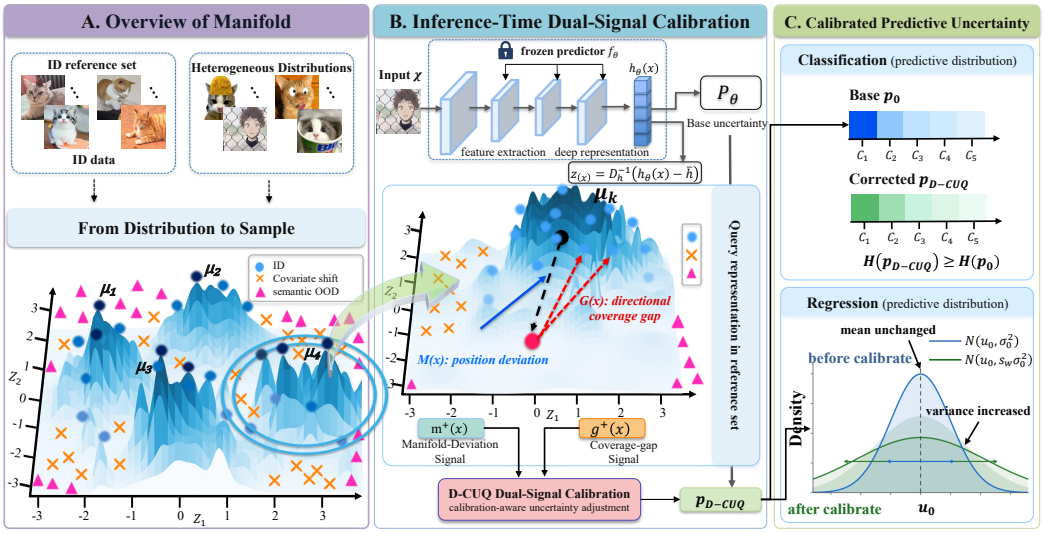

Figure 2: Overview of D-CUQ. (A) A sample-based view of heterogeneous distributions relative to ID manifolds. (B) Manifold-deviation and coverage-gap signals are computed from frozen representations against an ID reference set. (C) The signals guide conservative, sample-adaptive correction of predictive distributions for classification and regression while preserving point predictions.

Proposition 3 (Log-density form of the manifold-deviation signal)**.** *Let $\mu_1, \ldots, \mu_K$ and $\mathbf{D}_\nu$ be as in Eq.7, where $\mathbf{D}_\nu$ is positive definite by construction, and set $C_\nu = \frac{1}{2}\log\det(2\pi\mathbf{D}_\nu)$. For $u \in \mathbb{R}^d$ let $\widehat{p}_k(u) = \mathcal{N}(u; \mu_k, \mathbf{D}_\nu)$ denote the value at u of the Gaussian density with mean μ_k and covariance $\mathbf{D}_\nu$, so that the K densities differ only in their means. Then, for every input x with feature $z = h_\theta(x)$,*

$$
\begin{aligned}
-\log \max_{1\le k\le K} \widehat{p}_k(z) &= \frac{d}{2} M_{\mathcal{S}}(x) + C_\nu, \\
\max_{1\le k\le K} \widehat{p}_k(z) &= e^{-C_\nu} \exp\Big(-\tfrac{d}{2} M_{\mathcal{S}}(x)\Big),
\end{aligned}
\tag{9}
$$

with $M_{\mathcal{S}}$ given by Eq.8. Consequently, for any two inputs x and x' with features $z = h_\theta(x)$ and $z' = h_\theta(x')$,

$$
M_{\mathcal{S}}(x) > M_{\mathcal{S}}(x') \iff \max_{1\le k\le K} \widehat{p}_k(z) < \max_{1\le k\le K} \widehat{p}_k(z'). \tag{10}
$$

Remark.

- **Distance and improbability coincide.** Under a Gaussian assumption, lying far from every class mean and being improbable under every class are one statement, which is what $\delta_{\mathcal{M}}$ asks for.
- **The factor $1/d$ changes no ordering.** The function of $M_{\mathcal{S}}$ in Eq.9 is increasing, so by Eq.10 the factor $1/d$ leaves every ranking-based metric unchanged while making the signal comparable across feature dimensions.
- **The signal is on established ground.** Up to a monotone transform, $M_{\mathcal{S}}$ is the class-conditional Gaussian density score of (Lee et al., 2018), so this side of D-CUQ introduces no assumption beyond the one that line of work already makes.

3.4 Coverage-aware Predictive Calibration

Building on the two signals, we propose D-CUQ, summarised in Algorithm 1.

The raw scales of $M_{\mathcal{S}}$ and $G_{\mathcal{S}}$ depend on the feature space, so we standardize them on a calibration set $\mathcal{C}$ of ID samples disjoint from $\mathcal{S}$,

$$m_+(x) = \left[\frac{M_{\mathcal{S}}(x) - \bar{M}_{\mathcal{C}}}{s_{M,\mathcal{C}} + \epsilon_s}\right]_+, \qquad g_+(x) = \left[\frac{G_{\mathcal{S}}(x) - \bar{G}_{\mathcal{C}}}{s_{G,\mathcal{C}} + \epsilon_s}\right]_+, \tag{11}$$

where $\bar{M}_{\mathcal{C}}, s_{M,\mathcal{C}}$ and $\bar{G}_{\mathcal{C}}, s_{G,\mathcal{C}}$ are the means and standard deviations of the two signals on $\mathcal{C}$, $\epsilon_s > 0$ ensures numerical stability, and $[a]_+ = \max(a, 0)$, so only values above the ID level contribute. The rectified signals set an input-dependent scale

$$s_\omega(x) = 1 + \mathrm{softplus}\big(a_M m_+(x) + a_G g_+(x) + b\big), \qquad a_M = \mathrm{softplus}(\widetilde{a}_M),\ a_G = \mathrm{softplus}(\widetilde{a}_G), \tag{12}$$

where $\widetilde{a}_M, \widetilde{a}_G, b$ are free parameters. The outer softplus gives $s_\omega \geq 1$ and the inner ones give $a_M, a_G \geq 0$, so s_ω is non-decreasing in both signals and never sharpens the logits. For classification this defines an auxiliary distribution

$$q_\omega(y = k \mid x) = \frac{\exp\big(\ell_{\theta,k}(x)/s_\omega(x)\big)}{\sum_{j=1}^{K} \exp\big(\ell_{\theta,j}(x)/s_\omega(x)\big)}, \tag{13}$$

which has the form of temperature scaling (Guo et al., 2017) with an input-dependent temperature, but states the uncertainty indicated by coverage rather than replacing the base predictive distribution.

We fit $\omega = (\widetilde{a}_M, \widetilde{a}_G, b)$ on $\mathcal{C}$ together with a set $\mathcal{V}$ of virtual outliers sampled from the low-likelihood region of the class-conditional feature distribution (Du et al., 2022). The fit balances four terms, i.e.,

$$\begin{aligned} \widehat{\omega} \in \arg\min_{\omega}\ & \mathcal{L}_{\mathrm{ID}}(\omega) + \beta_{\mathrm{rank}}\big[\delta_H + \bar{H}_{\mathcal{C}}(\omega) - \bar{H}_{\mathcal{V}}(\omega)\big]_+ \\ & + \beta_{\mathrm{guard}}\big[\mathcal{L}_{\mathrm{ID}}(\omega) - \mathcal{L}_{\mathrm{base}} - \delta_{\mathrm{ID}}\big]_+ + \gamma\big(a_M^2 + a_G^2\big), \end{aligned} \tag{14}$$

where $\mathcal{L}_{\mathrm{ID}}(\omega)$ and $\mathcal{L}_{\mathrm{base}}$ are the negative log-likelihoods of q_ω and of the frozen head on $\mathcal{C}$, $\bar{H}_{\mathcal{C}}(\omega)$ and $\bar{H}_{\mathcal{V}}(\omega)$ are the average predictive entropies of q_ω on the two sets, $\delta_H > 0$ is the entropy margin the virtual outliers must exceed, $\delta_{\mathrm{ID}} > 0$ is the loss increase tolerated on $\mathcal{C}$, and $\beta_{\mathrm{rank}}, \beta_{\mathrm{guard}}, \gamma > 0$ are weights.

D-CUQ treats the uncertainty indicated by coverage as a floor rather than a replacement, so the uncertainty p_0 already carries is retained. Consider the power-scaling map $\mathcal{P}_\tau[p]_k = p_k^{1/\tau} / \sum_j p_j^{1/\tau}$ for $\tau \in [1, \infty]$, where $\mathcal{P}_1[p] = p$ and $\mathcal{P}_\infty[p] = \mathrm{Unif}(K)$. The calibration moves p_0 along this path to the larger of the two entropies, i.e.,

$$\begin{aligned} H_*(x) &= \max\big\{H(p_0(\cdot \mid x)),\ H(q_{\widehat{\omega}}(\cdot \mid x))\big\}, \\ \tau_*(x) &= \inf\big\{\tau \geq 1 : H(\mathcal{P}_\tau[p_0(\cdot \mid x)]) \geq H_*(x)\big\}, \\ p_{\text{D-CUQ}}(\cdot \mid x) &= \mathcal{P}_{\tau_*(x)}[p_0(\cdot \mid x)], \end{aligned} \tag{15}$$

where H is the Shannon entropy.

Algorithm 1: D-CUQ

Require: Frozen f_θ, baseline p_0, ID sets $\mathcal{S}, \mathcal{C}$
Ensure: Fitted module Ω, calibrated distribution

1: Build $\mathbf{Z}$ from $\mathcal{S}$; factorize $\mathbf{A}_\lambda$
2: Fit class means μ_k and shared covariance $\mathbf{D}_\nu$
3: Compute calibration statistics of M, G on $\mathcal{C}$
4: Generate $\widetilde{\sqsubseteq}$, fit $\widehat{\omega}$ by 14, store as Ω
Inference
5: Obtain $p_0(\cdot \mid x)$; compute $M(x), G(x)$
6: Rectify to $m_+(x), g_+(x)$; construct $q_{\widehat{\omega}}(\cdot \mid x)$
7: **if** $H(p_0) \geq H(q_{\widehat{\omega}})$ **return** $p_0(\cdot \mid x)$
8: **else return** $\mathcal{P}_{\tau_*(x)}[p_0(\cdot \mid x)]$, τ_* from 15

Assume $p_0(\cdot \mid x)$ has strictly positive coordinates. We show in Appendix that $\tau_*(x) \in [1, \infty]$ exists, that

$$H\big(p_{\text{D-CUQ}}(\cdot \mid x)\big) = H_*(x), \tag{16}$$

and that the finite solution is unique when $p_0(\cdot \mid x)$ is non-uniform with $H_*(x) < \log K$. Holding $p_0(\cdot \mid x)$ and $\ell_\theta(x)$ fixed, $H_*(x)$ is non-decreasing in both $m_+(x)$ and $g_+(x)$, and the calibration returns $p_0(\cdot \mid x)$ untouched whenever $H(p_0(\cdot \mid x)) \geq H(q_{\widehat{\omega}}(\cdot \mid x))$. The calibration therefore changes the uncertainty alone, leaving the point estimate and hence the accuracy unchanged, since every pairwise class ordering of $p_0(\cdot \mid x)$ is preserved.

Table 1: Classification results on the heterogeneous test sets. All metrics are averaged over 7 architectures and 5 random seeds, reported with and without (w/ vs. w/o) D-CUQ. Bold indicates superior performance.

Method	Corr. NLL ↓		Corr. Brier ↓		ECE ↓		AUROC ↑		AURC ↓	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Temp. Scaling	1.0784	**0.8404**	0.3283	**0.2497**	0.4531	**0.3568**	0.8764	**0.8931**	0.6792	**0.6706**
MC Dropout	2.2398	**0.9507**	0.4702	**0.2778**	0.5785	**0.3930**	0.8593	**0.8905**	0.6871	**0.6712**
Deep Ensemble	1.8839	**0.8750**	0.4137	**0.2538**	0.5314	**0.3658**	0.8704	**0.8910**	0.6754	**0.6660**
SWAG	2.3232	**0.9597**	0.4865	**0.2807**	0.5925	**0.3966**	0.8587	**0.8903**	0.6865	**0.6711**
VBLL	0.6588	**0.4982**	0.2075	**0.1561**	0.3098	**0.2289**	0.8759	**0.8966**	0.6829	**0.6724**
SNGP	0.2747	**0.2520**	0.0802	**0.0738**	0.0697	**0.0559**	0.8163	**0.8716**	0.7876	**0.7747**
EDL	0.3291	**0.2903**	0.1039	**0.0910**	0.0801	**0.0631**	0.8653	**0.8961**	0.6821	**0.6633**

4 EXPERIMENTS

Our evaluation spans both regression and classification benchmarks, with a primary focus on calibration performance under heterogeneous distributions. All experiments are conducted on 4 Hygon K100-AI 64GB DCUs. Additional details and results are provided in the Appendix.

Heterogeneous test sets. Each task is evaluated on Q_{mix} from Eq. 1 with equal mixing proportions and hidden component labels. (1) *Classification:* 2,500 samples from each of 7 datasets, CIFAR-100 (ID), CIFAR-100-C (covariate shift), CIFAR-10 and Tiny-ImageNet (near-OOD), MNIST, SVHN, Places365 (far-OOD), following the OpenOOD v1.5 (Zhang et al., 2024). (2) *Regression:* given the scarcity of shift benchmarks (Gardner et al., 2023), we construct the test set via biased selection (Huang et al., 2006), decomposing each input space into principal and residual subspaces, scoring samples with inverse-variance-weighted energy, and thresholding at the 60% to define 4 equally sampled shift regions (Jackson & Mudholkar, 1979). This is a standard decomposition in multivariate statistics.

4.1 CLASSIFICATION TASK

We use 4 CNNs (ResNet-18, ResNet-50, VGG-16, DenseNet-121) and 3 Transformers (ViT-Tiny, ViT-B/16, Swin-T), and for each backbone all methods start from the same frozen CIFAR-100 checkpoint. We report the mean and standard deviation over 5 random seeds.

We consider 7 baselines for classification: Temperature Scaling, MC Dropout, Deep Ensemble, SWAG, VBLL, SNGP (Liu et al., 2020), and EDL. We evaluate performance using the negative log-likelihood (NLL) (Gneiting & Raftery, 2007), the Brier score, ECE, the area under the receiver operating characteristic curve (AUROC) (Hanley & McNeil, 1982) and the area under the risk-coverage curve (AURC) (Geifman et al., 2019). On the heterogeneous test set, inputs whose labels fall outside $\mathcal{Y}_{\text{ID}}$ have no true class, so NLL and the Brier score are undefined for them. Following Hendrycks et al. (2019), we treat such inputs as incorrectly classified and evaluate the predicted confidence against the binary event that the prediction is correct (Guo et al., 2017; Le Coz et al., 2024), and report the result as Corr. NLL and Corr. Brier.

D-CUQ improves upon every baseline across all metrics on average (Tab. 1). The poorer a baseline's calibration, the greater improvement it receives. Across all the baselines, the ordering by correctness NLL exactly matches the ordering by its reduction. SWAG improves from 2.32 to 0.96 in NLL, while SNGP, already the best calibrated of the 7, moves only from 0.27 to 0.25. The practical effect is that the choice of baselines matters much less after calibration, since all 7 then lie between 0.25 and 0.96 rather than between 0.27 and 2.32. None of this costs accuracy, which the calibration leaves untouched by preserving class ordering.

4.2 REGRESSION TASK

We apply D-CUQ to 4 regression baselines: a deterministic Gaussian network that outputs a mean and a variance (Nix & Weigend, 1994), MC Dropout, Deep Ensembles and SWAG. For evaluation, we use RMSE to assess point prediction and NLL as the primary metric for probabilistic predic-

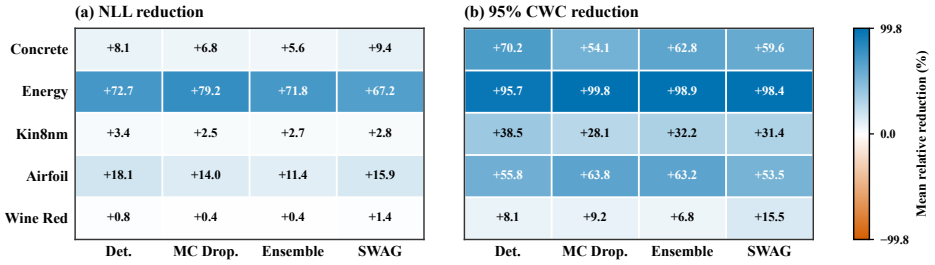

Figure 3: Mean relative reductions in predictive NLL and the 95% CWC after applying D-CUQ to 4 baselines across 5 regression datasets. Panels (a) and (b) report NLL and CWC reductions, respectively. Each cell shows the mean over 5 random seeds. Positive values indicate improvement.

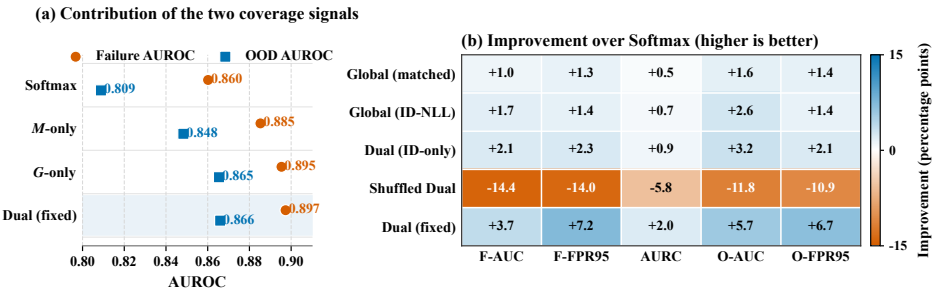

Figure 4: Result of ablation study. (a) Contributions of the individual and combined coverage signals. (b) Improvements over Softmax, where positive values indicate better performance.

tion (Gneiting & Raftery, 2007). We construct 95% prediction intervals and report CWC, together with the AUROC of predictive uncertainty as a detector of shifted inputs.

Notably, D-CUQ lowers both NLL and 95% CWC in Fig. 3. The magnitude of the gain is determined by the dataset rather than by the baseline. Averaged over methods, the NLL reduction runs from 0.8% on Wine Red to 72.7% on Energy, while averaged over datasets it stays between 18.4% and 20.6%. The CWC reduction exceeds the NLL reduction on 5 datasets, so what the calibration repairs first is interval coverage rather than the full density fit. The dual variant also gives the highest OOD AUROC for all baselines and beats the better of its two single-signal ablations on 19 of the 20 dataset and method combinations. None of this costs accuracy, since the calibration leaves the predictive mean and therefore the RMSE unchanged.

4.3 Ablation Study

On the classification heterogeneous test set, we isolate the two signals, averaging results over 6 backbones with 5 random seeds each. The single-signal variants use the same predictive mapping as the fixed dual variant, with scaling strengths matched on ID sample.

Each signal on its own improves over Softmax, but neither is a substitute for the other. G is the stronger single signal on 5 of the 6 backbones. M is stronger on DenseNet-121, where G is weakest. Using G therefore gives up 1.6 of failure AUROC on that backbone, and using M gives up 2.8 on ResNet-18. The four-regime probe measures the difference directly. Both signals together separate the regimes 10.6 of Macro-F1 above either signal (Fig. 4). The dual variant is best on all 5 macro metrics. It raises failure AUROC by 3.7 and OOD AUROC by 5.7 over Softmax, and lowers failure FPR95 by 7.2.

Four controls locate where this gain comes from. Replacing the input-dependent strength by a single global scale leaves OOD AUROC at 0.825 when the scale is matched on ID data and 0.835 when it is fitted to ID NLL, against 0.866 for the dual variant. The effect is therefore not generic logit smoothing. Fitting the dual mapping itself to ID NLL reaches 0.840, so ID calibration does not by

itself recover the coverage ordering. Shuffling the dual-signal scaling factors across inputs leaves their distribution unchanged but drops OOD AUROC to 0.691. That is 11.8 below the uncalibrated Softmax baseline. Misassigned uncertainty is worse than none. What the calibration contributes is the assignment of strength to individual inputs, not the amount of uncertainty it adds.

5 Discussion

The two signals were introduced for UQ under heterogeneous distributions, but we find that they are not confined to it. This section develops that point.

We freeze a ResNet-50 trained on CIFAR-100, and place each input at $[z_M(x), z_G(x)]$ in Figure 5. The 7 datasets occupy distinct regions ordered by distance from the ID cluster. That order does not follow the near and far grouping, but it tracks the ranking by OOD AUROC (Spearman $\rho = 0.89$, $n = 6$, one-sided exact $p = 0.017$). MNIST is the sharpest case, a far-OOD dataset that lies close to the ID cluster and is the least detectable of the six. Ridge probes over the 4 coverage regimes leave predictive entropy at chance, 20.1% Macro-F1 against a chance level of 25%, while z_M and z_G reach 26.3% and 33.3% alone and 43.9% together.

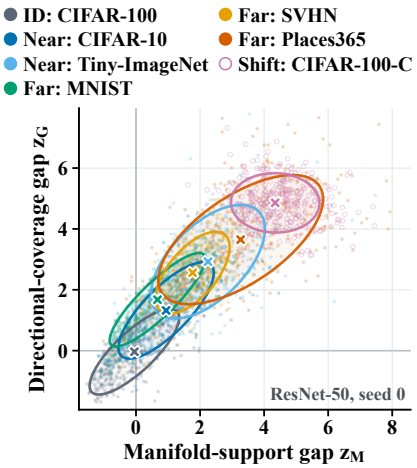

Figure 5: Datasets in the dual-signal plane.

That means, the gap between those probes is not just a gap between two uncertainty estimates. Neither signal is specific to UQ. Both are functions of the feature that every forward pass already produces, so they are present in any single inference, with or without a UQ method attached. The figure shows that inputs from 7 datasets fall into separate regions before any label is attached, and the dataset labels only name those regions rather than create them. A trained network already separates the kinds of data it is given, in its own representations, and nothing in the predictive distribution passes that on. In other words, the network already knows something about what it is being given, and has never told its user.

6 Conclusion

In this paper, we identify calibration degradation under heterogeneous distributions, and propose D-CUQ that combines manifold-deviation and coverage-gap signals to adjust the predictive distribution of existing methods without retraining the backbone. On the theoretical side, we establish a formal connection between the sample and distribution levels, revealing that distribution-level quantities are simply aggregates of their sample-level counterparts.

Extensive experiments confirm our theoretical derivations and demonstrate the superior calibration performance for different tasks. In classification , D-CUQ improves all five metrics for all seven base UQ methods, lowering the average ECE from 0.3736 to 0.2657 and by 33.1% on the most overconfident baseline. In regression, the dual signal raises the five-dataset macro OOD AUROC of all four base methods, from 0.661 to 0.838 on average.

Use of Large Language Models

Large language models were used in two ways in the preparation of this work. First, for polishing the language and the mathematical exposition of text written by the authors, including grammar, wording, and the presentation of definitions, statements, and proofs. Second, as coding assistants during the implementation of the experiments. The research idea, the method, the theoretical results, and the experimental design are the authors' own, and all text in this paper has been reviewed and verified by the authors. The authors take full responsibility for all content of this paper.

REFERENCES

Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28*, pp. 775–783, 2015.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Thomas F. Brooks, D. Stuart Pope, and Michael A. Marcolini. Airfoil self-noise and prediction. Technical Report NASA-RP-1218, NASA Langley Research Center, July 1989.

Ha Manh Bui and Anqi Liu. Density-softmax: Efficient test-time model for uncertainty estimation and robustness under distribution shifts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4822–4853. PMLR, 2024.

Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela van der Schaar. Unlabelled data improves Bayesian uncertainty calibration under covariate shift. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1392–1402. PMLR, 2020.

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.

Erik A. Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux - effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20089–20103, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *The Tenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016.

Zhitong Gao, Bingnan Li, Mathieu Salzmann, and Xuming He. Generalize or detect? towards robust semantic segmentation under multiple distribution shifts. In *Advances in Neural Information Processing Systems 37*, 2024.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023. doi: 10.1007/s10462-023-10562-9.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *International Conference on Learning Representations*. OpenReview.net, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Dapeng Hu, Jian Liang, Xinchao Wang, and Chuan-Sheng Foo. Pseudo-calibration: Improving predictive uncertainty estimation in unsupervised domain adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19304–19326. PMLR, 2024.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, 2017.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

J. E. Jackson and G. S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3088–3098, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 10–15 Jul 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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. doi: 10.1109/5.726791.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7167–7177, 2018.

Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Uncertainty weighted gradients for model calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15497–15507. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01444.

Jeremiah Z. 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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7498–7512, 2020.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6804–6814. PMLR, 18–24 Jul 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002, 2021b. doi: 10.1109/ICCV48922.2021.00986.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13132–13143, 2019.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023. doi: 10.1109/CVPR52729.2023.02336.

Hariharan Narayanan and Sanjoy K. Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23, pp. 1786–1794, 2010.

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

D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13969–13980. Curran Associates, Inc., 2019.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3219–3229. PMLR, 2020.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006. ISBN 026218253X.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: benchmarks for subpopulation shift. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3183–3193, 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. doi: 10.1016/S0378-3758(00)00115-4.

Max Simchowitz, Anurag Ajay, Pulkit Agrawal, and Akshay Krishnamurthy. Statistical learning under heterogeneous distribution shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Yuli Slavutsky and David Blei. Quantifying uncertainty in the presence of distribution shifts. *Advances in Neural Information Processing Systems*, 38:96792–96819, 2026.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007. doi: 10.5555/1314498.1390324.

Christian Tomani, Futa Waseda, Yuesong Shen, Marvin Eisenberger, Daniel Cremers, and Florian Buettner. Beyond in-domain scenarios: Robust density-aware calibration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012. doi: 10.1016/j.enbuild.2012.03.003.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008. doi: 10.1561/2200000001.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023. doi: 10.1007/s11263-023-01811-z.

I.-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998. doi: 10.1016/S0008-8846(98)00165-3.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2:(3):1–32, 2024.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

A Appendix

A.1 Distribution-Level Interpretation of the Sample-Level Regimes

This section relates Definition 1 to the distribution-level notions of covariate shift and semantic OOD, and states what each direction of that relation requires.

Let $Z = h_\theta(X)$ be the representation induced by the trained predictor. For measurable $A \subseteq \mathbb{R}^d$ and any components $Q_{\mathrm{cov}} = Q_r^{\mathrm{cov}}$ and $Q_{\mathrm{OOD}} = Q_s^{\mathrm{OOD}}$ of Eq. 1, define the pushforwards

$$P_{\mathrm{ID}}^Z(A) = P_{\mathrm{ID}}\big(h_\theta(X) \in A\big), \quad Q_{\mathrm{cov}}^Z(A) = Q_{\mathrm{cov}}\big(h_\theta(X) \in A\big), \quad Q_{\mathrm{OOD}}^Z(A) = Q_{\mathrm{OOD}}\big(h_\theta(X) \in A\big). \tag{17}$$

Under the manifold hypothesis, let $\mathcal{M}_k \subset \mathbb{R}^d$ be the representation manifold of the k-th ID class, let $\mathcal{M}_{\mathrm{ID}} = \bigcup_{k=1}^K \mathcal{M}_k$, and let

$$\mathcal{M}_{\mathrm{ID}}^{(\varepsilon_{\mathcal{M}})} = \big\{z : \mathrm{dist}(z, \mathcal{M}_{\mathrm{ID}}) \le \varepsilon_{\mathcal{M}}\big\} \tag{18}$$

be the neighborhood at the threshold of Definition 1, so that $\delta_{\mathcal{M}}(x) > \varepsilon_{\mathcal{M}}$ holds exactly when $z \notin \mathcal{M}_{\mathrm{ID}}^{(\varepsilon_{\mathcal{M}})}$.

The two coordinates are measured against different objects. $\delta_{\mathcal{M}}$ is measured against $\mathcal{M}_{\mathrm{ID}}$, the set where ID representations lie, and $U_{\mathcal{S}}$ against $\mathcal{S}$, a finite set that has actually been realized. We define the coverage deficiency by its behaviour. $U_{\mathcal{S}}(x)$ is high for an input that $\mathcal{S}$ has not observed near z, and does not increase when an input is added,

$$\mathcal{S} \subseteq \mathcal{S}' \implies U_{\mathcal{S}'}(x) \le U_{\mathcal{S}}(x). \tag{19}$$

Theorem 2 establishes this property for $G_{\mathcal{S}}$ at fixed ridge parameter and normalization, and we take $U_{\mathcal{S}} = G_{\mathcal{S}}$ throughout this section.

Assumption 4 (Semantic separation, idealized)**.** The ID representation law is supported on the neighborhood of the ID manifold, and semantic OOD places positive mass outside it,

$$P_{\mathrm{ID}}^Z\big(\mathcal{M}_{\mathrm{ID}}^{(\varepsilon_{\mathcal{M}})}\big) = 1, \qquad Q_{\mathrm{OOD}}^Z\big(\mathcal{M}_{\mathrm{ID}}^{(\varepsilon_{\mathcal{M}})}\big) < 1. \tag{20}$$

Order the eigenpairs of $\mathbf{Z}^\top\mathbf{Z}$ in Theorem 2 as $\xi_1 \ge \cdots \ge \xi_d$, and for a distribution Q over inputs let

$$w_j(Q) = \mathbb{E}_Q\Big[\frac{(u_j^\top z)^2}{\|z\|_2^2 + \epsilon}\Big] \tag{21}$$

be the normalized energy it places along u_j.

Assumption 5 (Directional shift)**.** Relative to P_{ID}, Q_{cov} places relatively more normalized energy along the directions the reference set covers least. For every $k < d$,

$$\sum_{j\le k} w_j(Q_{\mathrm{cov}}) \le \sum_{j\le k} w_j(P_{\mathrm{ID}}), \qquad \sum_{j\le d} w_j(Q_{\mathrm{cov}}) \ge \sum_{j\le d} w_j(P_{\mathrm{ID}}), \tag{22}$$

with at least one strict inequality at an index k where $\xi_{k+1} < \xi_k$ or at $k = d$.

Proposition 6 (One-way consistency)**.** *Under Assumption 4,*

$$Q_{\mathrm{OOD}}\big(\delta_{\mathcal{M}}(X) > \varepsilon_{\mathcal{M}}\big) > 0 = P_{\mathrm{ID}}\big(\delta_{\mathcal{M}}(X) > \varepsilon_{\mathcal{M}}\big). \tag{23}$$

Under Assumption 5,

$$\mathbb{E}_{Q_{\mathrm{cov}}}\big[U_{\mathcal{S}}(X)\big] > \mathbb{E}_{P_{\mathrm{ID}}}\big[U_{\mathcal{S}}(X)\big]. \tag{24}$$

Neither converse is claimed. A distribution that is not semantic OOD may still place mass at $\delta_{\mathcal{M}}(X) > \varepsilon_{\mathcal{M}}$, and one that is not covariate-shifted may still raise $U_{\mathcal{S}}$.

Proof. The first claim follows from Assumption 4, since $\delta_{\mathcal{M}}(x) > \varepsilon_{\mathcal{M}} \iff z \notin \mathcal{M}_{\mathrm{ID}}^{(\varepsilon_{\mathcal{M}})}$. For the second, the spectral form in Eq. 5 gives $\mathbb{E}_Q[G_{\mathcal{S}}(X)] = \sum_{j=1}^d c_j w_j(Q)$ with $c_j = \lambda/(\xi_j + \lambda)$, which is non-decreasing in j. Let $\Delta_j = w_j(Q_{\mathrm{cov}}) - w_j(P_{\mathrm{ID}})$ and $D_k = \sum_{j\le k} \Delta_j$. Abel summation gives

$$\sum_{j=1}^d c_j \Delta_j = c_d D_d - \sum_{k<d} (c_{k+1} - c_k) D_k, \tag{25}$$

where $c_d > 0$, $D_d \ge 0$, $c_{k+1} - c_k \ge 0$ and $D_k \le 0$, so every term is non-negative. The strict inequality in Assumption 5 makes the sum positive. □

Each statement in Proposition 6 is an expectation of a sample-level quantity under one component of Eq. 1. By linearity the same holds for the mixture, for instance

$$Q_{\text{mix}}\big(\delta_{\mathcal{M}}(X) > \varepsilon_{\mathcal{M}}\big) = \pi_0 P_{\text{ID}}(\,\cdot\,) + \sum_r \pi_r^{\text{cov}} Q_r^{\text{cov}}(\,\cdot\,) + \sum_s \pi_s^{\text{OOD}} Q_s^{\text{OOD}}(\,\cdot\,), \tag{26}$$

where $(\,\cdot\,)$ denotes the same event. Distribution-level quantities are thus aggregates of their sample-level counterparts, weighted by the mixing proportions.

Remark 7. The manifold statement in Section 3.1 is to be read for inputs drawn from P_{ID}, as formalized in Assumption 4. It need not hold for every input whose label lies in $\mathcal{Y}_{\text{ID}}$. Under severe corruption an input keeps its label in $\mathcal{Y}_{\text{ID}}$, so its generating component is covariate shift, while the evidence for any ID class can be destroyed and its representation can leave $\mathcal{M}_{\text{ID}}^{(\varepsilon_{\mathcal{M}})}$. Definition 1 then places it in the third case. The regime reflects the evidence available for the input rather than its generating component, and such inputs are the ones whose predictions should carry high uncertainty.

The cases in Definition 1 are mutually exclusive and exhaustive. If $\delta_{\mathcal{M}}(x) \le \varepsilon_{\mathcal{M}}$ the input lies in the neighborhood of the ID manifold and its case is decided by coverage, a small $U_{\mathcal{S}}(x)$ giving the first and a large one the second. If $\delta_{\mathcal{M}}(x) > \varepsilon_{\mathcal{M}}$ the input lies outside that neighborhood and falls in the third.

The regimes classify the evidence the reference set provides about an input, which is the quantity a calibration to the predictive distribution responds to. A covariate-shifted input that lands in a densely covered region is supported by the reference set as well as a clean ID input is, and calls for no calibration. A rare ID input in a weakly covered region is supported less well, and calls for one. Grouping inputs by the evidence available for them is therefore different from grouping them by their generating component, and it is the grouping the calibration acts on.

The manifold deviation is replaced by its computable proxy, $\delta_{\mathcal{M}}(x) \longmapsto M_{\mathcal{S}}(x)$, while $U_{\mathcal{S}} = G_{\mathcal{S}}$ is already computable, and D-CUQ keeps the pair $\mathbf{c}(x) = \big(M_{\mathcal{S}}(x), G_{\mathcal{S}}(x)\big)$ as a continuous coordinate. The thresholds make the partition explicit in Definition 1 and are not used in the computation; D-CUQ acts on the continuous coordinate and makes no decision at a regime boundary.

A.2 Proof and Further Analysis of the Coverage-Gap Signal

This section proves Theorem 2 and analyses the coverage-gap signal further. Let

$$z = z(x), \qquad \mathbf{S} = \mathbf{Z}^\top \mathbf{Z}, \qquad \mathbf{A}_\lambda = \mathbf{S} + \lambda I_d, \tag{27}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times d}$ contains the standardized ID reference representations and $\lambda > 0$, and let

$$\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top, \qquad \mathbf{k}_x = \mathbf{Z}z, \qquad \kappa_{xx} = z^\top z. \tag{28}$$

Both $\mathbf{A}_\lambda$ and $\mathbf{K} + \lambda I_n$ are positive definite and therefore invertible.

The ridge parameter is scaled with the average reference energy,

$$\lambda = \rho_\lambda \frac{\operatorname{tr}(\mathbf{Z}^\top \mathbf{Z})}{d}, \qquad \rho_\lambda > 0, \tag{29}$$

so that it tracks the magnitude of the reference scatter matrix. All results below condition on the fitted normalization and the resulting value of λ.

Let the linear prediction head be $\ell_\theta(x) = \mathbf{W}h_\theta(x) + b$. Since $h_\theta(x) = \mathbf{D}_h z(x) + \overline{h}$, the same head reads in standardized coordinates as

$$\ell_\theta(x) = \widetilde{\mathbf{W}} z(x) + \widetilde{b}, \qquad \widetilde{\mathbf{W}} = \mathbf{W}\mathbf{D}_h, \qquad \widetilde{b} = \mathbf{W}\overline{h} + b. \tag{30}$$

The Jacobian of the logits with respect to the standardized last-layer weight block is

$$\mathbf{J}_{\widetilde{\mathbf{W}}}(x) = I_K \otimes z(x)^\top, \tag{31}$$

so that

$$\mathbf{J}_{\widetilde{\mathbf{W}}}(x)\mathbf{J}_{\widetilde{\mathbf{W}}}(x')^\top = \big(z(x)^\top z(x')\big) I_K = \kappa(x, x') I_K, \tag{32}$$

where $\kappa(x, x') = z(x)^\top z(x')$ is the scalar kernel induced by the standardized last-layer weight block (Jacot et al., 2018). The bias block is excluded; including it adds a constant kernel term.

Proof. We establish the kernel identity, the spectral representation, boundedness and coverage monotonicity in turn.

Kernel residual identity. From

$$(\mathbf{K} + \lambda I_n)\mathbf{Z} = (\mathbf{Z}\mathbf{Z}^\top + \lambda I_n)\mathbf{Z} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z} + \lambda I_d) = \mathbf{Z}\mathbf{A}_\lambda, \tag{33}$$

multiplying by the corresponding inverses gives the push-through identity

$$(\mathbf{K} + \lambda I_n)^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{A}_\lambda^{-1}. \tag{34}$$

Hence

$$\begin{aligned}\mathbf{k}_x^\top(\mathbf{K} + \lambda I_n)^{-1}\mathbf{k}_x &= z^\top\mathbf{Z}^\top(\mathbf{K} + \lambda I_n)^{-1}\mathbf{Z}z \\ &= z^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{A}_\lambda^{-1}z = z^\top\mathbf{S}\mathbf{A}_\lambda^{-1}z.\end{aligned} \tag{35}$$

Because $\mathbf{S} = \mathbf{A}_\lambda - \lambda I_d$, we have $I_d - \mathbf{S}\mathbf{A}_\lambda^{-1} = \lambda\mathbf{A}_\lambda^{-1}$, and therefore

$$\kappa_{xx} - \mathbf{k}_x^\top(\mathbf{K} + \lambda I_n)^{-1}\mathbf{k}_x = z^\top\big(I_d - \mathbf{S}\mathbf{A}_\lambda^{-1}\big)z = \lambda z^\top\mathbf{A}_\lambda^{-1}z = v_\lambda(x), \tag{36}$$

which is the kernel-residual equality in Eq.5.

Spectral representation and boundedness. Let $\mathbf{S} = \mathbf{U}\,\mathrm{diag}(\xi_1, \ldots, \xi_d)\mathbf{U}^\top$ with $\mathbf{U}$ orthogonal and $\xi_j \geq 0$. Then

$$\mathbf{A}_\lambda^{-1} = \mathbf{U}\,\mathrm{diag}\Big(\tfrac{1}{\xi_1+\lambda}, \ldots, \tfrac{1}{\xi_d+\lambda}\Big)\mathbf{U}^\top, \tag{37}$$

and writing u_j for the j-th column of $\mathbf{U}$,

$$v_\lambda(x) = \lambda z^\top\mathbf{A}_\lambda^{-1}z = \sum_{j=1}^{d}\frac{\lambda}{\xi_j + \lambda}(u_j^\top z)^2, \tag{38}$$

which is the spectral equality in Eq.5. Since $0 < \lambda/(\xi_j + \lambda) \leq 1$,

$$0 \leq G_{\mathbf{Z}}(x) \leq \frac{\sum_{j=1}^{d}(u_j^\top z)^2}{\|z\|_2^2 + \epsilon} = \frac{\|z\|_2^2}{\|z\|_2^2 + \epsilon} \leq 1. \tag{39}$$

Coverage monotonicity. Let $z_+ \in \mathbb{R}^d$ be an additional standardized reference representation and let $\widetilde{\mathbf{Z}} = [\mathbf{Z}^\top,\, z_+]^\top$. Holding the standardization and λ fixed,

$$\widetilde{\mathbf{S}} = \mathbf{S} + z_+z_+^\top, \qquad \widetilde{\mathbf{A}}_\lambda = \mathbf{A}_\lambda + z_+z_+^\top, \tag{40}$$

and the Sherman–Morrison identity gives

$$\widetilde{\mathbf{A}}_\lambda^{-1} = \mathbf{A}_\lambda^{-1} - \frac{\mathbf{A}_\lambda^{-1}z_+z_+^\top\mathbf{A}_\lambda^{-1}}{1 + z_+^\top\mathbf{A}_\lambda^{-1}z_+}. \tag{41}$$

Hence

$$v_{\lambda,\mathbf{Z}}(x) - v_{\lambda,\widetilde{\mathbf{Z}}}(x) = \lambda z^\top\big(\mathbf{A}_\lambda^{-1} - \widetilde{\mathbf{A}}_\lambda^{-1}\big)z = \frac{\lambda(z^\top\mathbf{A}_\lambda^{-1}z_+)^2}{1 + z_+^\top\mathbf{A}_\lambda^{-1}z_+} \geq 0. \tag{42}$$

The normalization $\|z\|_2^2 + \epsilon$ is unchanged, so

$$G_{\mathbf{Z}}(x) - G_{\widetilde{\mathbf{Z}}}(x) = \frac{\lambda(z^\top\mathbf{A}_\lambda^{-1}z_+)^2}{(\|z\|_2^2 + \epsilon)(1 + z_+^\top\mathbf{A}_\lambda^{-1}z_+)} \geq 0, \tag{43}$$

with strict inequality exactly when $z^\top\mathbf{A}_\lambda^{-1}z_+ \neq 0$. This proves Eq.6. □

The spectral representation settles two limiting cases. If z lies in the null space of $\mathbf{Z}$, then $\mathbf{S}z = 0$ and $\mathbf{A}_\lambda^{-1}z = \lambda^{-1}z$, so

$$v_\lambda(x) = \|z\|_2^2, \qquad G_{\mathbf{Z}}(x) = \frac{\|z\|_2^2}{\|z\|_2^2 + \epsilon}, \tag{44}$$

and a nonzero direction absent from the ID reference representations receives a value close to one. If z is aligned with an eigenvector u_j carrying reference energy ξ_j, then

$$G_{\mathbf{Z}}(x) = \frac{\lambda}{\xi_j + \lambda} \frac{\|z\|_2^2}{\|z\|_2^2 + \epsilon}, \tag{45}$$

which decreases as ξ_j/λ grows. More generally, if an augmented reference scatter matrix satisfies $\widetilde{\mathbf{S}} \succeq \mathbf{S}$ at fixed λ, then $(\widetilde{\mathbf{S}} + \lambda I_d)^{-1} \preceq (\mathbf{S} + \lambda I_d)^{-1}$ and therefore $G_{\widetilde{\mathbf{Z}}}(x) \leq G_{\mathbf{Z}}(x)$, so the monotonicity extends from a single added sample to any positive-semidefinite increase in reference coverage.

Boundedness, spectral weighting and monotonicity together are the properties required of a coverage deficiency, and we therefore take

$$U_{\mathcal{S}}(x) := G_{\mathbf{Z}}(x) \tag{46}$$

as its computable form in Definition 1, at fixed standardization statistics and fixed ridge parameter.

We now relate the unnormalized residual $v_\lambda(x)$ to the predictive variance of a last-layer Laplace approximation. Consider one scalar output of the standardized linear head, $f_w(x) = w^\top z(x)$, with isotropic Gaussian prior $p(w) = \mathcal{N}(0, \alpha^{-1} I_d)$ and $\alpha > 0$, and let $\widehat{w}$ be the maximum-a-posteriori estimate. The Laplace approximation builds a Gaussian posterior around $\widehat{w}$ from the local curvature of the negative log posterior (Khan et al., 2019; Daxberger et al., 2021). Assume the negative log-likelihood has constant latent-output curvature over the reference samples,

$$\left.\frac{\partial^2 \mathcal{L}_i(f)}{\partial f^2}\right|_{f=\widehat{w}^\top z(x_i)} = \beta, \qquad \beta > 0, \tag{47}$$

under which the posterior precision is

$$\mathbf{\Lambda} = \alpha I_d + \beta \mathbf{Z}^\top \mathbf{Z}. \tag{48}$$

Corollary 8 (Last-layer Laplace variance)**.** *Under Eq.47, let $\lambda = \alpha/\beta$. The latent predictive variance satisfies*

$$s_f^2(x) := \mathrm{Var}_{w|\mathcal{S}}\big[f_w(x)\big] = \frac{1}{\alpha} v_\lambda(x) = \frac{1}{\alpha}\Big[\kappa_{xx} - \mathbf{k}_x^\top (\mathbf{K} + \lambda I_n)^{-1} \mathbf{k}_x\Big], \tag{49}$$

and consequently

$$G_{\mathbf{Z}}(x) = \frac{\alpha s_f^2(x)}{\|z(x)\|_2^2 + \epsilon}. \tag{50}$$

Proof. Since $\lambda = \alpha/\beta$,

$$\mathbf{\Lambda} = \alpha I_d + \beta \mathbf{S} = \beta(\mathbf{S} + \lambda I_d) = \beta \mathbf{A}_\lambda, \qquad \mathbf{\Lambda}^{-1} = \frac{1}{\beta} \mathbf{A}_\lambda^{-1}. \tag{51}$$

The variance of $f_w(x) = w^\top z$ under the Gaussian Laplace posterior is therefore

$$s_f^2(x) = z^\top \mathbf{\Lambda}^{-1} z = \frac{1}{\beta} z^\top \mathbf{A}_\lambda^{-1} z = \frac{1}{\alpha} \lambda z^\top \mathbf{A}_\lambda^{-1} z = \frac{v_\lambda(x)}{\alpha}. \tag{52}$$

Substituting Eq.36 proves Eq.49, and dividing $v_\lambda(x) = \alpha s_f^2(x)$ by $\|z(x)\|_2^2 + \epsilon$ proves Eq.50. □

For homoscedastic Gaussian regression the constant-curvature assumption is exact, with β the inverse observation-noise variance, and $s_f^2(x)$ is the epistemic variance of the latent regression function, to which the observation-noise variance is added to obtain the full predictive variance. For multiclass classification the Softmax likelihood curvature is class-coupled and input-dependent. Writing $p_i \in \mathbb{R}^K$ for the predictive probability vector of reference sample i, its class-space curvature is $\mathbf{B}_i = \mathrm{diag}(p_i) - p_i p_i^\top$, and the exact last-layer precision

$$\mathbf{\Lambda}_{\mathbf{W}} = \alpha I_{Kd} + \sum_{i=1}^{n} \mathbf{B}_i \otimes z(x_i) z(x_i)^\top \tag{53}$$

does not in general reduce to $I_K \otimes (\alpha I_d + \beta \mathbf{Z}^\top \mathbf{Z})$. Corollary 8 therefore holds under the isotropic-curvature approximation.

A.3 Proof and Further Analysis of the Manifold-Deviation Signal

This subsection proves Proposition 3 and gives further properties of the manifold-deviation signal. Recall that $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ is the finite ID reference set and $z(x_i) \in \mathbb{R}^d$ the corresponding standardized representation.

For each ID class k, define

$$\mathcal{I}_k = \{i : y_i = k\}, \qquad n_k = |\mathcal{I}_k|, \qquad \mu_k = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} z(x_i), \tag{54}$$

with $n_k > 0$ for every class. The shared diagonal covariance is estimated from all standardized ID reference representations,

$$\overline{z} = \frac{1}{n} \sum_{i=1}^n z(x_i), \qquad \nu_j = \frac{1}{n} \sum_{i=1}^n \big(z_j(x_i) - \overline{z}_j\big)^2, \qquad \mathbf{D}_\nu = \mathrm{diag}(\nu_1, \ldots, \nu_d) + \epsilon_\nu I_d, \tag{55}$$

with $\epsilon_\nu > 0$, so that $\mathbf{D}_\nu \succ 0$ and its inverse and determinant are well defined. For an arbitrary representation $z \in \mathbb{R}^d$, let

$$q_k(z) = (z - \mu_k)^\top \mathbf{D}_\nu^{-1} (z - \mu_k), \tag{56}$$

so that

$$M_{\mathcal{S}}(x) = \frac{1}{d} \min_{1 \le k \le K} q_k(z(x)). \tag{57}$$

Where no ambiguity arises we also write $M_{\mathcal{S}}(z)$ for the same function evaluated directly at a representation z.

Proof. For each class k, the shared-covariance Gaussian reference is

$$\widehat{p}_k(z) = (2\pi)^{-d/2} |\mathbf{D}_\nu|^{-1/2} \exp\Big[-\tfrac{1}{2} q_k(z)\Big], \tag{58}$$

whose negative logarithm is

$$-\log \widehat{p}_k(z) = \tfrac{1}{2} q_k(z) + C_\nu, \qquad C_\nu = \tfrac{d}{2} \log(2\pi) + \tfrac{1}{2} \log |\mathbf{D}_\nu|. \tag{59}$$

Because every class uses the same covariance, C_ν depends on neither z nor k. The negative logarithm is monotonically decreasing, so

$$-\log \max_{1 \le k \le K} \widehat{p}_k(z) = \min_{1 \le k \le K} \big[-\log \widehat{p}_k(z)\big] = \tfrac{1}{2} \min_{1 \le k \le K} q_k(z) + C_\nu. \tag{60}$$

Substituting $\min_k q_k(z(x)) = d M_{\mathcal{S}}(x)$ into Eq.60 gives

$$-\log \max_k \widehat{p}_k(z(x)) = \tfrac{d}{2} M_{\mathcal{S}}(x) + C_\nu, \tag{61}$$

which proves Eq.9. □

The characterization can be written as

$$\max_k \widehat{p}_k(z(x)) = \exp(-C_\nu) \exp\Big[-\tfrac{d}{2} M_{\mathcal{S}}(x)\Big], \tag{62}$$

so $M_{\mathcal{S}}$ induces exactly the reverse ordering of the maximum class-conditional reference density. For any two inputs,

$$M_{\mathcal{S}}(x) - M_{\mathcal{S}}(x') = \frac{2}{d} \log \frac{\max_k \widehat{p}_k(z(x'))}{\max_k \widehat{p}_k(z(x))}, \tag{63}$$

and an input has a larger manifold-deviation signal precisely when its maximum density under the fitted class references is smaller.

For $\tau \ge 0$, let

$$\widehat{\mathcal{E}}_k(\tau) = \{z \in \mathbb{R}^d : q_k(z) \le d\tau\}, \qquad \widehat{\mathcal{R}}_{\mathrm{ID}}(\tau) = \bigcup_{k=1}^K \widehat{\mathcal{E}}_k(\tau). \tag{64}$$

From Eq.57,

$$M_{\mathcal{S}}(x) \le \tau \iff \min_k q_k(z(x)) \le d\tau \iff z(x) \in \widehat{\mathcal{R}}_{\mathrm{ID}}(\tau), \tag{65}$$

so each sublevel set of $M_{\mathcal{S}}$ is exactly a union of class-conditional ellipsoidal neighborhoods. Combining Eq.62 with Eq.65,

$$M_{\mathcal{S}}(x) \le \tau \iff \max_k \widehat{p}_k(z(x)) \ge \exp\Big(-C_\nu - \tfrac{d}{2}\tau\Big), \tag{66}$$

so thresholding $M_{\mathcal{S}}$ is equivalent to thresholding the largest class-conditional reference density.

Consider a path leaving the centre of class k, $z_t = \mu_k + ta$ with $t \ge 0$ and $a \in \mathbb{R}^d$, and suppose class k remains the closest reference over an interval of t. There,

$$M_{\mathcal{S}}(z_t) = \frac{1}{d}(z_t - \mu_k)^\top \mathbf{D}_\nu^{-1}(z_t - \mu_k) = \frac{t^2}{d} a^\top \mathbf{D}_\nu^{-1} a, \tag{67}$$

and since $\mathbf{D}_\nu \succ 0$, for every nonzero direction a,

$$\frac{\mathrm{d}}{\mathrm{d}t} M_{\mathcal{S}}(z_t) = \frac{2t}{d} a^\top \mathbf{D}_\nu^{-1} a > 0, \qquad t > 0. \tag{68}$$

The signal therefore increases as the representation moves away from its closest class region along a fixed direction. Where the closest class changes, $M_{\mathcal{S}}$ stays continuous and is piecewise differentiable, being the pointwise minimum of finitely many continuous quadratic functions.

If a representation from class k follows the fitted reference model $Z \mid Y = k \sim \mathcal{N}(\mu_k, \mathbf{D}_\nu)$, then $q_k(Z) \sim \chi^2_d$ and

$$\mathbb{E}\Big[\frac{q_k(Z)}{d}\Big] = 1, \qquad \mathrm{Var}\Big[\frac{q_k(Z)}{d}\Big] = \frac{2}{d}. \tag{69}$$

Taking the minimum over classes gives $M_{\mathcal{S}}(Z) \le q_k(Z)/d$ and $\mathbb{E}[M_{\mathcal{S}}(Z)] \le 1$. The factor $1/d$ therefore keeps the nominal scale of the signal from growing linearly with representation dimensionality, and it leaves the ordering of inputs unchanged.

The signal compares $z(x)$ with every ID class reference and takes the minimum, so it uses no label from the base model. The distinction matters when the base prediction is wrong. An input far from the predicted class but close to another ID class still receives a small manifold-deviation value, and a large value requires the input to be far from every fitted ID class region.

Equation Eq.65 gives

$$M_{\mathcal{S}}(x) = \inf\big\{\tau \ge 0 : z(x) \in \widehat{\mathcal{R}}_{\mathrm{ID}}(\tau)\big\}, \tag{70}$$

so $M_{\mathcal{S}}(x)$ is the smallest normalized ellipsoidal radius at which the representation enters an empirical class-conditional ID region. It supplies the positional component of the sample-wise manifold view, and $G_{\mathbf{Z}}(x)$ the directional one. An input can have small $M_{\mathcal{S}}(x)$ and large $G_{\mathbf{Z}}(x)$, lying close to an ID class region along a direction the finite reference set represents weakly, while a large $M_{\mathcal{S}}(x)$ indicates positional deviation from every fitted ID class region whatever its directional coverage.

The implementation computes the class means from the standardized ID reference representations and estimates each ν_j by the population variance over the complete reference set, which is the factor $1/n$ in Eq.55. The experiments use $\epsilon_\nu = 0.1$ after feature standardization. All reference statistics are fitted offline and stay fixed at inference, and no test-set or test-batch statistics are used. The diagonal construction costs $\mathcal{O}(nd)$ operations for the covariance estimate and $\mathcal{O}(nd)$ for the class means, stores $\mathcal{O}(Kd + d)$ statistics, and evaluating $M_{\mathcal{S}}(x)$ for one input costs $\mathcal{O}(Kd)$ operations.

The equality in Proposition 3 is exact for the fitted class-conditional Gaussian reference, and the proof uses only that all class references share one positive-definite covariance matrix. A full shared covariance or a pooled within-class covariance yields the same algebraic characterization; D-CUQ uses the regularized global diagonal estimator in Eq.55. The Gaussian reference describes the geometry of the learned representations rather than the distribution of the inputs themselves, and $M_{\mathcal{S}}(x)$ is accordingly an empirical positional quantity defined by the fitted class references. Class priors are omitted from $\widehat{p}_k$, so the signal compares the geometric compatibility of an input with each ID class without favouring classes that occur more often in the reference set. The complete dual-signal coordinate is

$$\mathbf{c}(x) = \big(M_{\mathcal{S}}(x), G_{\mathbf{Z}}(x)\big), \tag{71}$$

with $M_{\mathcal{S}}(x)$ measuring positional mismatch and $G_{\mathbf{Z}}(x)$ directional undercoverage.

A.4 Regression Instantiation of D-CUQ

This section extends D-CUQ to regression tasks. The core principle remains invariant: dual calibration signals are extracted from a frozen representation to calibrate baseline predictive uncertainty. To accommodate the absence of discrete class labels in regression, we replace the class-conditional reference used in classification with a target-conditioned counterpart.

Let $\mathcal{S}_{\mathrm{R}} = \{(x_i, y_i)\}_{i=1}^{n}$ be the regression reference set and let $h_\theta(x) \in \mathbb{R}^d$ be the representation given by a frozen reference regressor. Representations are standardized as

$$z(x) = \mathbf{D}_h^{-1}\big(h_\theta(x) - \overline{h}\big), \tag{72}$$

with $\overline{h}$ and $\mathbf{D}_h$ estimated from $\mathcal{S}_{\mathrm{R}}$. We partition the reference targets into B empirical quantile intervals, let $\mathcal{I}_b$ contain the indices whose targets fall in interval b, and define

$$n_b = |\mathcal{I}_b|, \qquad \mu_b = \frac{1}{n_b}\sum_{i\in\mathcal{I}_b} z(x_i), \qquad b = 1, \dots, B. \tag{73}$$

The shared regularized diagonal covariance is

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z(x_i), \qquad \nu_j = \frac{1}{n}\sum_{i=1}^{n}\big(z_j(x_i) - \overline{z}_j\big)^2, \qquad \mathbf{D}_\nu = \mathrm{diag}(\nu_1, \dots, \nu_d) + \epsilon_\nu I_d, \tag{74}$$

and the regression manifold-deviation signal is

$$M_{\mathrm{R}}(x) = \frac{1}{d}\min_{1\le b\le B}\big(z(x) - \mu_b\big)^\top \mathbf{D}_\nu^{-1}\big(z(x) - \mu_b\big). \tag{75}$$

The positional signal thus compares an input with the representation regions associated with different ranges of the continuous target. Taking the minimum over intervals makes the signal available from the representation, without the predicted target.

The coverage-gap signal is unchanged,

$$G_{\mathrm{R}}(x) = \frac{\lambda z(x)^\top\big(\mathbf{Z}^\top\mathbf{Z} + \lambda I_d\big)^{-1} z(x)}{\|z(x)\|_2^2 + \epsilon}, \tag{76}$$

where $\mathbf{Z}$ stacks the standardized regression reference representations. Both signals are standardized on a disjoint ID calibration set and rectified as in Eq.11.

Each baseline is represented by a predictive Gaussian

$$p_0(y \mid x) = \mathcal{N}\big(\mu_0(x), \sigma_0^2(x)\big), \tag{77}$$

where for sampling-based methods $\mu_0(x)$ and $\sigma_0^2(x)$ are the predictive moments. With the same monotone dual-signal mapping

$$s_\omega(x) = 1 + \mathrm{softplus}\big(a_M m_+(x) + a_G g_+(x) + b\big), \qquad a_M, a_G \ge 0, \tag{78}$$

the calibrated distribution is

$$p_{\text{D-CUQ}}(y \mid x) = \mathcal{N}\big(\mu_0(x), s_\omega(x)\sigma_0^2(x)\big). \tag{79}$$

The predictive mean, and therefore every point-prediction metric computed from it, is unchanged.

Proposition 9 (Properties of the Gaussian calibration)**.** *Let $s = s_\omega(x) \ge 1$. The calibration in Eq.79 satisfies*

$$\mathbb{E}_{\text{D-CUQ}}[Y \mid x] = \mu_0(x), \qquad \mathrm{Var}_{\text{D-CUQ}}[Y \mid x] = s\sigma_0^2(x), \qquad H\big(p_{\text{D-CUQ}}\big) - H\big(p_0\big) = \tfrac{1}{2}\log s, \tag{80}$$

and the width of every symmetric Gaussian prediction interval is multiplied by $\sqrt{s}$. For an observed target y, let $r(x, y) = (y - \mu_0(x))^2/\sigma_0^2(x)$ be the squared standardized residual. The change in Gaussian NLL is

$$\mathcal{L}_{\text{D-CUQ}}(x, y) - \mathcal{L}_0(x, y) = \frac{1}{2}\Big[\log s + r(x, y)\Big(\frac{1}{s} - 1\Big)\Big], \tag{81}$$

and for $s > 1$ the calibration lowers the NLL exactly when

$$r(x, y) > \frac{s\log s}{s - 1}. \tag{82}$$

Proof. The first two identities are the mean and variance of the Gaussian in Eq.79. The differential entropy of $\mathcal{N}(\mu, \sigma^2)$ is $\frac{1}{2}\log(2\pi e\sigma^2)$, so replacing $\sigma_0^2(x)$ by $s\sigma_0^2(x)$ raises it by $\frac{1}{2}\log s$. A symmetric Gaussian prediction interval has width proportional to the predictive standard deviation, which changes from $\sigma_0(x)$ to $\sqrt{s}\sigma_0(x)$. Subtracting the base Gaussian NLL

$$\mathcal{L}_0(x, y) = \tfrac{1}{2}\big[\log(2\pi\sigma_0^2(x)) + r(x, y)\big] \tag{83}$$

from the calibrated one gives Eq.81, and rearranging $\mathcal{L}_{\text{D-CUQ}}(x, y) < \mathcal{L}_0(x, y)$ gives Eq.82. □

For the learned regression variant the mapping parameters are estimated on an ID calibration set $\mathcal{C}_{\text{R}}$. Writing $\mu_{0,i}$ and $\sigma_{0,i}^2$ for the base predictive moments of calibration sample (x_i, y_i), we minimize

$$\widehat{\omega} \in \arg\min_{\omega} \frac{1}{2|\mathcal{C}_{\text{R}}|} \sum_{(x_i,y_i)\in\mathcal{C}_{\text{R}}} \Big[\log\big(2\pi s_\omega(x_i)\sigma_{0,i}^2\big) + \frac{(y_i - \mu_{0,i})^2}{s_\omega(x_i)\sigma_{0,i}^2}\Big] + \gamma(a_M^2 + a_G^2). \tag{84}$$

In the paired UCI experiments the mapping is fitted once using the deterministic Gaussian reference predictor and then applied unchanged to every baseline.

The regression experiments also evaluate a fixed mapping, which separates the contribution of the dual signals from that of parameter learning. Define

$$d_{\text{dual}}(x) = \sqrt{m_+^2(x) + g_+^2(x)}, \tag{85}$$

and let $q_{0.95}$ and s_d be the 95th percentile and standard deviation of d_{dual} on the ID calibration set. The fixed scale is

$$s_{\text{fixed}}(x) = 1 + 1.5\,\text{softplus}\Big(\frac{d_{\text{dual}}(x) - q_{0.95}}{s_d + \epsilon}\Big). \tag{86}$$

The experiments report this fixed dual-signal variant and the ID-learned mapping in Eq.84, together with the corresponding single-signal ablations.

The UCI experiments use at most ten target-quantile intervals, and fewer when the reference set is small or quantile boundaries coincide. After feature standardization the diagonal covariance regularizer is $\epsilon_\nu = 0.1$ and the ridge parameter is $\lambda = 0.05\,\text{tr}(\mathbf{Z}^\top\mathbf{Z})/d$. The learned mapping is optimized for 500 steps on the ID calibration set.

For regression, the instantiation scales the baseline predictive variance via multiplicative adjustment rather than composing through an entropy floor, leaving the predictive mean untouched. Proposition 9 gives the condition under which the calibration lowers the Gaussian NLL, namely $r(x, y) > s\log s/(s-1)$, so the regime in which inflation is beneficial is characterized in closed form. The evaluation reports NLL, CRPS, interval coverage, interval width and calibration alongside error-based risk ranking.

A.5 Properties of the Predictive Calibration

Proposition 10 (Properties of the predictive calibration)**.** *Assume $p_0(\cdot \mid x)$ has strictly positive coordinates and $a_M, a_G \geq 0$. Then $\tau_*(x) \in [1, \infty]$ exists and*

$$p_{\text{D-CUQ}}(\cdot \mid x) = \mathcal{P}_{\tau_*(x)}\big[p_0(\cdot \mid x)\big], \qquad H\big(p_{\text{D-CUQ}}(\cdot \mid x)\big) = H_*(x). \tag{87}$$

The calibration preserves every pairwise class ordering of $p_0(\cdot \mid x)$. Holding $p_0(\cdot \mid x)$ and $\ell_\theta(x)$ fixed, $H_*(x)$ is non-decreasing in both $m_+(x)$ and $g_+(x)$. If $H(p_0(\cdot \mid x)) \geq H(q_{\widehat{\omega}}(\cdot \mid x))$ then $p_{\text{D-CUQ}}(\cdot \mid x) = p_0(\cdot \mid x)$, and if $p_0(\cdot \mid x)$ is non-uniform with $H_*(x) < \log K$ the finite solution is unique.

Proof. Fix an input x and suppress its dependence where no ambiguity arises. We first establish a general entropy property. For $\mathbf{u} = (u_1, \ldots, u_K)^\top \in \mathbb{R}^K$ and $t > 0$, define

$$\rho_{t,k} = \frac{\exp(u_k/t)}{\sum_{j=1}^K \exp(u_j/t)}. \tag{88}$$

With $\eta = 1/t$ and $Z(\eta) = \sum_j \exp(\eta u_j)$, the entropy of ρ_t is $H(\rho_t) = \log Z(\eta) - \eta \mathbb{E}_{\rho_t}[u]$. The log-partition identities for exponential families (Wainwright & Jordan, 2008) give

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \log Z(\eta) = \mathbb{E}_{\rho_t}[u], \qquad \frac{\mathrm{d}}{\mathrm{d}\eta} \mathbb{E}_{\rho_t}[u] = \mathrm{Var}_{\rho_t}(u), \tag{89}$$

so that

$$\frac{\mathrm{d}H(\rho_t)}{\mathrm{d}t} = \frac{\mathrm{Var}_{\rho_t}(u)}{t^3} \geq 0, \tag{90}$$

with strict inequality at finite t whenever the coordinates of $\mathbf{u}$ are not all equal.

Applying this to the coverage-induced distribution $q_{\widehat{\omega}}$ and holding the frozen logits fixed, its scale satisfies

$$\frac{\partial s_{\widehat{\omega}}}{\partial m_+} = a_M \,\mathrm{sigmoid}(\psi_{\widehat{\omega}}) \geq 0, \qquad \frac{\partial s_{\widehat{\omega}}}{\partial g_+} = a_G \,\mathrm{sigmoid}(\psi_{\widehat{\omega}}) \geq 0, \tag{91}$$

because $a_M, a_G \geq 0$. With Eq.90, $H(q_{\widehat{\omega}})$ is therefore coordinate-wise non-decreasing in both m_+ and g_+.

Consider next the power path of the base predictive distribution. Since $p_{0,k} > 0$ for every class, set $u_k = \log p_{0,k}$, so that $\mathcal{P}_\tau[p_0] = \mathrm{softmax}(\mathbf{u}/\tau)$. By Eq.90, $H(\mathcal{P}_\tau[p_0])$ is continuous and non-decreasing in τ, with

$$H(\mathcal{P}_1[p_0]) = H(p_0), \qquad \lim_{\tau \to \infty} H(\mathcal{P}_\tau[p_0]) = \log K. \tag{92}$$

Since $H_* = \max\{H(p_0), H(q_{\widehat{\omega}})\} \in [H(p_0), \log K]$, continuity yields an extended solution $\tau_* \in [1, \infty]$. It is finite when $H_* < \log K$; when $H_* = \log K$ and p_0 is non-uniform it is attained in the extended sense at $\tau_* = \infty$; and when p_0 is uniform, $\tau_* = 1$. In every case $H(\mathcal{P}_{\tau_*}[p_0]) = H_*$, and therefore

$$H\big(p_{\text{D-CUQ}}\big) = \max\{H(p_0), H(q_{\widehat{\omega}})\} \geq H(p_0), \tag{93}$$

so the calibration returns at least the entropy supplied by the base method. The right-hand side is non-decreasing in m_+ and g_+, and the calibrated entropy inherits that monotonicity.

If $H(p_0) \geq H(q_{\widehat{\omega}})$ then $H_* = H(p_0)$. Since $\tau = 1$ already attains this value and no $\tau < 1$ is admitted, the infimum in Eq.15 is $\tau_* = 1$ and $p_{\text{D-CUQ}} = \mathcal{P}_1[p_0] = p_0$.

For every finite $\tau > 0$ and every pair of classes i, j,

$$\mathcal{P}_\tau[p_0]_i \geq \mathcal{P}_\tau[p_0]_j \iff p_{0,i}^{1/\tau} \geq p_{0,j}^{1/\tau} \iff p_{0,i} \geq p_{0,j}, \tag{94}$$

so every finite calibration preserves the complete pairwise class ordering, ties included, and hence the predicted class.

Finally, if p_0 is non-uniform then $\mathrm{Var}_{\mathcal{P}_\tau[p_0]}(\log p_0) > 0$ for every finite τ, so by Eq.90 the entropy is strictly increasing along the finite power path. Whenever $H_* < \log K$, the required entropy increase therefore has a unique finite solution. □

A.6 Additional Classification OOD and Failure-Detection Experiments

We additionally evaluate D-CUQ in a classification setting that isolates OOD and failure-detection performance. For CIFAR-10 as ID data, the test pool contains 2,500 ID samples, 5,000 near-OOD samples from CIFAR-100 and Tiny ImageNet, and 7,500 far-OOD samples from MNIST, SVHN, and Places365. The evaluation covers DenseNet-121, ResNet-50, VGG-16, and ViT-Tiny. For CIFAR-100 as ID data, the test pool contains 2,500 ID samples and 2,500 samples from each of CIFAR-10, Tiny ImageNet, MNIST, SVHN, and Places365. This evaluation additionally includes ResNet-18. Every backbone is evaluated with five random seeds.

Pure OOD detection treats ID and semantic OOD samples as the negative and positive classes, respectively. Failure detection additionally treats incorrect ID predictions as failures. We report AUROC and FPR95 for both settings, together with AURC for selective prediction. The aggregate results in Figure 6 summarize all backbone–seed evaluations; error bars denote their standard deviations.

Figure 6 shows that D-CUQ provides the strongest aggregate separation on both ID benchmarks. With CIFAR-10 as ID, D-CUQ increases Failure AUROC from 0.9717 to 0.9764 and OOD AUROC

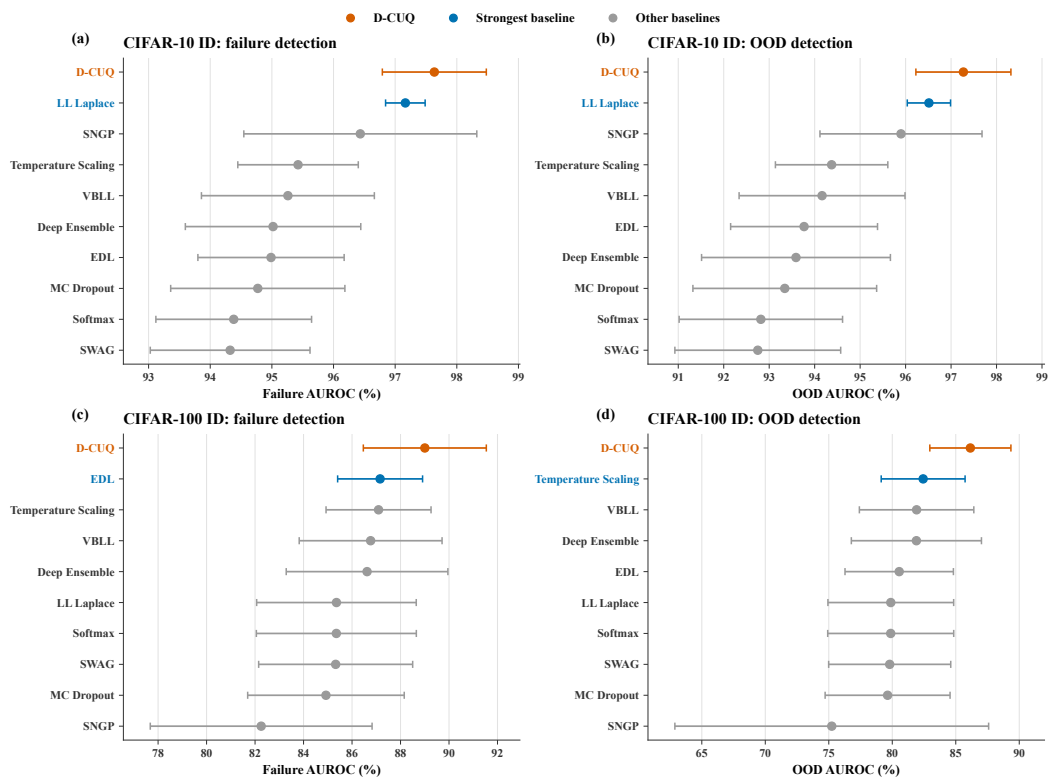

Figure 6: Classification OOD and failure-detection results. Markers and error bars denote means and standard deviations over backbone–seed evaluations. Orange denotes D-CUQ, blue the strongest baseline for each metric, and gray the remaining baselines.

from 0.9651 to 0.9727 relative to the strongest baseline for each metric. It also reduces Failure FPR95 from 0.1477 to 0.1246 and OOD FPR95 from 0.1753 to 0.1439. With CIFAR-100 as ID, D-CUQ obtains a Failure AUROC of 0.8900 and an OOD AUROC of 0.8615, compared with 0.8716 and 0.8243 for the corresponding strongest baselines. Its Failure FPR95 and OOD FPR95 are also reduced to 0.5059 and 0.5829, respectively. Across individual architectures, D-CUQ has a higher mean Failure AUROC than the strongest baseline on 7 of the 9 backbone comparisons.

A.7 Detail Information of Discussion

This section gives the full setup for the diagnostic reported in Section 5.

We adopt a frozen seed-0 ResNet-50 (He et al., 2016) checkpoint pretrained on CIFAR-100 (Krizhevsky, 2009). The evaluation suite comprises 17,500 test samples (2,500 per dataset across seven subsets): clean in-distribution (ID) data from CIFAR-100; near-OOD from CIFAR-10 and Tiny-ImageNet (derived from ImageNet (Deng et al., 2009)); far-OOD from MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), and Places365 (Zhou et al., 2018); and covariate-shift data from Gaussian-noise-corrupted CIFAR-100-C (following the standard corruption benchmark (Hendrycks & Dietterich, 2019)). Grouping into near-OOD and far-OOD strictly follows the generalized OpenOOD v1.5 protocol (Zhang et al., 2024).

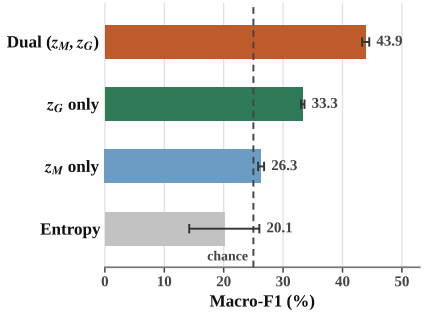

Figure 7: Separability results.

Each sample is represented directly as

$$\mathbf{s}(x) = \big[z_M(x),\, z_G(x)\big]^\top, \tag{95}$$

where the normalization statistics of both signals are estimated exclusively from the held-out ID calibration split. No t-SNE (van der Maaten & Hinton, 2008), UMAP (McInnes et al., 2018) or learned projection is applied, so both axes retain their original meanings. In Figure 5, 340 samples are drawn for readability, while the confidence ellipses are estimated from all 2500 samples of each kind. The axis limits are the 0.1th and 99.9th percentiles of the pooled coordinates and affect display only.

Table 2 reports, for each non-ID dataset, the Euclidean distance of its mean in the coverage plane from the ID mean, together with its OOD AUROC on the same checkpoint. Both coordinates are already standardized, so no further scaling enters the distance. The two orderings agree except for two adjacent swaps, CIFAR-10 with MNIST and SVHN with Tiny-ImageNet (Spearman $\rho = 0.89$, $n = 6$, one-sided exact $p = 0.017$).

Table 2: Distance of each dataset mean from the ID cluster in the coverage plane, against its OOD AUROC. Both are computed on the seed-0 ResNet-50 checkpoint shown in Figure 5. Datasets are listed by increasing distance.

Dataset	Group	Distance from ID	OOD AUROC↑
CIFAR-10	near	1.68	0.8274
MNIST	far	1.86	0.7754
SVHN	far	3.17	0.9513
Tiny-ImageNet	near	3.75	0.9391
Places365	far	4.96	0.9590
CIFAR-100-C	shift	6.58	0.9986

The four coverage regimes are ID, near-OOD, far-OOD and covariate shift. Ridge-regularized linear probes take predictive entropy, z_M, z_G, or the concatenation $[z_M, z_G]$ as input. Macro-F1 is averaged over 20 independently generated stratified $50/50$ train–test splits, and the error bars in Figure 7 are standard deviations over those splits. The four-class chance level is 25%. Predictive entropy reaches $20.1 \pm 5.9\%$, which the error bar places at chance. The two signals reach $26.3 \pm 0.5\%$ and $33.3 \pm 0.3\%$ on their own and $43.9 \pm 0.6\%$ together.

The probe measures coverage-regime structure. Dataset labels are used to build the heterogeneous test set and to score the probes. They are not required to compute either signal, and they are not available to D-CUQ at inference time.

A.8 Additional Details for the Regression Evaluation

This section gives the details of the regression experiment in Section 4.2. The experiment complements the main classification benchmark by testing whether the two signals remain useful when uncertainty is represented by a continuous predictive variance rather than class probabilities.

We use 5 regression benchmarks obtained through OpenML (Vanschoren et al., 2013): Concrete (Yeh, 1998), Energy (Tsanas & Xifara, 2012), Kin8nm, Airfoil (Brooks et al., 1989), and Wine Quality (Red) (Cortez et al., 2009). Features and targets used by the regressors are standardized using statistics estimated from the training split only. Table 3 reports the dataset dimensions and the resulting split sizes.

For benchmark construction, we decompose the standardized raw input matrix as $\widetilde{X} = U\Sigma V^\top$. Let $a_j(x)$ and ν_j denote the coordinate and empirical variance along the j-th right-singular direction. The major subspace is the smallest set of directions explaining 80% of the input variance. We define

$$q_M(x) = \frac{1}{d_M} \sum_{j \le d_M} \frac{a_j(x)^2}{\nu_j + \epsilon}, \qquad q_G(x) = \frac{1}{d - d_M} \sum_{j > d_M} \frac{a_j(x)^2}{\nu_j + \epsilon}. \tag{96}$$

Samples below the 60th percentile of both scores form the unshifted pool. Samples exceeding the threshold only in q_M, only in q_G, or in both scores form the major-subspace, minor-subspace, and

Table 3: Regression datasets and split sizes. The 4 test regimes are dataset-balanced, so the mixed-test size is 4 times the number of samples per regime. Split sizes are identical across the 5 seeds.

Dataset	OpenML ID	Samples	Features	Major dim.	Train	Calibration	Per regime	Mixed test
Concrete	44959	1030	8	5	257	64	82	328
Energy	44960	768	8	4	189	47	61	244
Kin8nm	189	8192	8	7	1709	427	512	2048
Airfoil	44957	1503	5	3	366	91	120	480
Wine Red	40691	1599	11	6	429	107	128	512

combined-shift pools, respectively. Equal numbers of samples from the four pools form the mixed test set. The remaining unshifted samples are divided into training (80%) and calibration (20%) subsets; unused samples are discarded. The partition scores and regime identities are used only to construct the benchmark and are never provided to the regressor or D-CUQ.

The shared regressor is a two-hidden-layer multilayer perceptron with 64units per layer and $\tanh$ activations. A linear head produces the predictive mean. Each network is trained for 300 full-batch epochs using AdamW with a learning rate of 3×10^{-3} and weight decay 10^{-4}. We repeat the complete split, training, calibration, and evaluation procedure with 5 random seeds.

We evaluate deterministic Gaussian regression, MC Dropout with dropout rate 0.1 and 30 stochastic predictions (Gal & Ghahramani, 2016), a five-member Deep Ensemble (Lakshminarayanan et al., 2017), and SWAG (Maddox et al., 2019). SWAG collects 20 snapshots over 120 additional SGD epochs using a learning rate of 0.01, momentum 0.9, and weight decay 10^{-4}; evaluation uses 30 posterior parameter samples. For each method, the ID calibration split estimates a shared aleatoric component and a global variance scale, so the comparison focuses on input-dependent uncertainty rather than arbitrary differences in global variance magnitude.

After training, the reference regressor is frozen, and its second hidden layer provides $d = 64$ dimensional representations. Standardizing the training representations produces the reference matrix Z. To form the regression analogue of the class-conditional reference, we divide the training targets into at most ten equal-frequency bins and let μ_b denote the feature center of bin b. The manifold-deviation signal is

$$M(x) = \min_b \frac{1}{d} \sum_{j=1}^{d} \frac{\big(z_j(x) - \mu_{b,j}\big)^2}{\widehat{v}_j + 0.1}, \tag{97}$$

where $\widehat{v}_j$ is the variance of the j-th training feature. The coverage-gap signal is

$$G(x) = \frac{\lambda z(x)^\top \big(Z^\top Z + \lambda I\big)^{-1} z(x)}{\|z(x)\|_2^2 + \epsilon}, \qquad \lambda = 0.05\, \frac{\mathrm{tr}(Z^\top Z)}{d}. \tag{98}$$

Both signals are standardized using their ID-calibration means and standard deviations, clipped to $[-20, 20]$, and rectified to obtain $m_+(x)$ and $g_+(x)$. For this controlled regression evaluation, the dual score and variance scale are

$$C(x) = \sqrt{m_+(x)^2 + g_+(x)^2}, \qquad s(x) = 1 + 1.5\ \mathrm{softplus}\left(\frac{C(x) - Q_{0.95}^{\mathrm{ID}}}{s_C + \epsilon}\right), \tag{99}$$

where $Q_{0.95}^{\mathrm{ID}}$ and s_C are the 95th percentile and standard deviation of the calibration dual scores. The M-only and G-only variants apply the same mapping to the corresponding single signal. Given a base prediction $\mathcal{N}(\mu_0(x), \sigma_0^2(x))$, D-CUQ returns

$$p_{\text{D-CUQ}}(y \mid x) = \mathcal{N}\big(\mu_0(x), s(x)\sigma_0^2(x)\big)\,. \tag{100}$$

The predictive mean is therefore identical across Base, M-only, G-only, and Dual; only the predictive variance changes.

The unshifted region is treated as ID and the three shifted regions as OOD. Predictive standard deviation is used as the common uncertainty score. OOD AUROC measures separation between unshifted and shifted inputs. Failure AUROC defines a prediction failure when its absolute error exceeds the 95th percentile of errors in the unshifted test region. We additionally evaluate Gaussian negative log-likelihood, continuous ranked probability score, RMSE, MAE, 90% and 95% interval coverage,

95% interval width, regression calibration error, Spearman correlation between uncertainty and absolute error, and AURC. Metrics are computed on the training-standardized target scale. For each seed, results are first macro-averaged over the 5 datasets; the reported mean and standard deviation are then computed over 5 seeds.

Figure 3 shows that dual-signal attains the highest five-dataset average OOD AUROC for all four baselines. It also exceeds the stronger single-signal variant in 19 of the 20 dataset–method combinations. The only exception is Deep Ensemble on Concrete, where Dual is lower than the stronger single signal by 0.004 OOD AUROC. Because the predictive mean is preserved, RMSE and MAE are unchanged across the coverage variants by construction. The experiment is a controlled cross-task test of the dual-signal mechanism.

A.9 Complete Experimental Results

The classification experiments use four CNNs, ResNet-18 and ResNet-50 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2015) and DenseNet-121 (Huang et al., 2017), together with three Transformers, ViT-Tiny and ViT-B/16 (Dosovitskiy et al., 2021) and Swin-T (Liu et al., 2021b). CIFAR-100 is the ID dataset and CIFAR-10 one of the semantic OOD (Krizhevsky, 2009), CIFAR-100-C supplies the covariate-shifted inputs (Hendrycks & Dietterich, 2019), and Tiny-ImageNet (Deng et al., 2009), MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011) and Places365 (Zhou et al., 2018) supply the remaining semantic OOD inputs. The regression experiments use 5 benchmarks obtained through OpenML (Vanschoren et al., 2013), namely Concrete (Yeh, 1998), Energy (Tsanas & Xifara, 2012), Kin8nm, Airfoil (Brooks et al., 1989) and Wine Quality (Red) (Cortez et al., 2009).

Unless stated otherwise, all entries report mean $\pm$ sample standard deviation over 5 random seeds. D-CUQ method labels and the best value within each protocol-comparable block are shown in bold. We do not compare boldface across blocks that use different datasets, backbones, or training protocols.

Table 4 gives the numerical values underlying the introductory ECE-degradation figure. This diagnostic uses ResNet-18 and 5 seeds.

Table 4: ECE (%) as the proportion of shifted samples increases. Bold values are lowest at each mixture ratio.

Shift type	UQ method	0%	10%	20%	40%	60%	80%
Covariate shift	Temperature Scaling	2.1 ± 0.1	3.9 ± 0.6	7.1 ± 1.1	13.7 ± 2.5	20.3 ± 4.5	26.9 ± 6.2
	MC Dropout	1.9 ± 0.6	2.7 ± 0.2	5.8 ± 0.8	12.3 ± 2.1	18.8 ± 3.9	25.4 ± 5.5
	Deep Ensemble	4.9 ± 0.7	3.9 ± 0.5	5.4 ± 0.3	9.7 ± 0.4	15.1 ± 0.5	21.0 ± 0.6
	SWAG	2.7 ± 0.5	4.8 ± 0.6	8.0 ± 1.1	15.0 ± 2.7	21.9 ± 4.8	28.8 ± 6.6
Semantic OOD	Temperature Scaling	2.1 ± 0.1	5.5 ± 0.7	10.5 ± 0.9	20.7 ± 1.1	30.7 ± 1.2	40.5 ± 1.9
	MC Dropout	1.9 ± 0.6	4.0 ± 0.5	9.0 ± 0.7	19.1 ± 0.9	29.0 ± 1.0	38.8 ± 1.7
	Deep Ensemble	4.9 ± 0.7	2.5 ± 0.8	5.3 ± 0.6	15.0 ± 0.5	25.0 ± 0.4	34.7 ± 0.7
	SWAG	2.7 ± 0.5	6.3 ± 0.5	11.3 ± 0.6	21.6 ± 1.0	31.6 ± 1.2	41.5 ± 1.9

Table 5 reports the architecture-macro results for the heterogeneous test sets. Correctness NLL, Brier, and ECE evaluate the predicted probability that an input is a valid calibrated prediction; failure metrics use predictive uncertainty to rank invalid or incorrect predictions. Each macro-level result first averages the 7 architectures within a seed and then reports the mean and standard deviation over 5 seeds.

Table 5: Main classification UQ results. Bold values are best within each baseline pair.

Baseline	Variant	Corr. NLL↓	Brier↓	ECE↓	AUROC↑	FPR95↓	AURC↓
Temp. Scale	Base	1.0784 ± 0.0228	0.3283 ± 0.0056	0.4531 ± 0.0054	0.8764 ± 0.0034	0.5630 ± 0.0210	0.6792 ± 0.0042
	+ D-CUQ	0.8404 ± 0.0229	0.2497 ± 0.0061	0.3568 ± 0.0069	0.8931 ± 0.0034	0.5329 ± 0.0234	0.6706 ± 0.0042
MC Drop.	Base	2.2398 ± 0.0321	0.4702 ± 0.0037	0.5785 ± 0.0034	0.8593 ± 0.0029	0.5819 ± 0.0155	0.6871 ± 0.0033
	+ D-CUQ	0.9507 ± 0.0285	0.2778 ± 0.0063	0.3930 ± 0.0064	0.8905 ± 0.0036	0.5294 ± 0.0214	0.6712 ± 0.0040
Deep Ens.	Base	1.8839 ± 0.0064	0.4137 ± 0.0017	0.5314 ± 0.0013	0.8704 ± 0.0031	0.5541 ± 0.0106	0.6754 ± 0.0033
	+ D-CUQ	0.8750 ± 0.0263	0.2538 ± 0.0052	0.3658 ± 0.0052	0.8910 ± 0.0035	0.5284 ± 0.0205	0.6660 ± 0.0039
SWAG	Base	2.3232 ± 0.0311	0.4865 ± 0.0049	0.5925 ± 0.0042	0.8587 ± 0.0031	0.5781 ± 0.0159	0.6865 ± 0.0030
	+ D-CUQ	0.9597 ± 0.0291	0.2807 ± 0.0066	0.3966 ± 0.0066	0.8903 ± 0.0039	0.5296 ± 0.0217	0.6711 ± 0.0041
VBLL	Base	0.6588 ± 0.2062	0.2075 ± 0.0551	0.3098 ± 0.0621	0.8759 ± 0.0102	0.5432 ± 0.0400	0.6829 ± 0.0067
	+ D-CUQ	0.4982 ± 0.0868	0.1561 ± 0.0254	0.2289 ± 0.0343	0.8966 ± 0.0051	0.5030 ± 0.0263	0.6724 ± 0.0041
SNGP	Base	0.2747 ± 0.0038	0.0802 ± 0.0014	0.0697 ± 0.0012	0.8163 ± 0.0041	0.6591 ± 0.0151	0.7876 ± 0.0017
	+ D-CUQ	0.2520 ± 0.0036	0.0738 ± 0.0013	0.0559 ± 0.0014	0.8716 ± 0.0045	0.5952 ± 0.0235	0.7747 ± 0.0021
EDL	Base	0.3291 ± 0.0061	0.1039 ± 0.0023	0.0801 ± 0.0041	0.8653 ± 0.0051	0.5410 ± 0.0142	0.6821 ± 0.0033
	+ D-CUQ	0.2903 ± 0.0047	0.0910 ± 0.0016	0.0631 ± 0.0020	0.8961 ± 0.0053	0.4966 ± 0.0131	0.6633 ± 0.0035

Table 6: Per-backbone main classification UQ results for CNN and ViT-Tiny backbones. In each metric cell, the upper value is Base and the lower value is **+D-CUQ**. Calibration and failure-detection metrics are grouped separately. Values are mean ± standard deviation over 5 seeds. Bold values are best within each pair.

		Calibration			Failure detection		
Backbone	Baseline	NLL↓	Brier↓	ECE↓	F-AUROC↑	F-FPR95↓	AURC↓
ResNet-18		0.9356 ± 0.0535	0.2913 ± 0.0148	0.4217 ± 0.0148	0.8700 ± 0.0087	0.5830 ± 0.0334	0.6972 ± 0.0055
	Temp. Scale	0.8361 ± 0.0480	0.2628 ± 0.0139	0.3894 ± 0.0143	0.8753 ± 0.0082	0.5697 ± 0.0323	0.6942 ± 0.0052
		0.8697 ± 0.0446	0.2750 ± 0.0128	0.4049 ± 0.0124	0.8722 ± 0.0096	0.5757 ± 0.0328	0.6957 ± 0.0056
	MC Drop.	0.8208 ± 0.0451	0.2586 ± 0.0131	0.3847 ± 0.0134	0.8755 ± 0.0092	0.5713 ± 0.0312	0.6941 ± 0.0054
		0.7684 ± 0.0061	0.2426 ± 0.0016	0.3675 ± 0.0010	0.8800 ± 0.0042	0.5637 ± 0.0107	0.6831 ± 0.0027
	Deep Ens.	0.7028 ± 0.0159	0.2209 ± 0.0046	0.3399 ± 0.0054	0.8806 ± 0.0032	0.5627 ± 0.0120	0.6827 ± 0.0023
		0.9772 ± 0.0591	0.3007 ± 0.0153	0.4312 ± 0.0144	0.8689 ± 0.0100	0.5845 ± 0.0360	0.6978 ± 0.0069
	SWAG	0.8418 ± 0.0488	0.2646 ± 0.0141	0.3913 ± 0.0145	0.8750 ± 0.0080	0.5714 ± 0.0312	0.6939 ± 0.0054
		0.7290 ± 0.4067	0.2232 ± 0.1088	0.3322 ± 0.1237	0.8721 ± 0.0163	0.5672 ± 0.0574	0.7009 ± 0.0124
	VBLL	0.5826 ± 0.1167	0.1848 ± 0.0371	0.2910 ± 0.0535	0.8822 ± 0.0092	0.5426 ± 0.0461	0.6951 ± 0.0071
		0.3481 ± 0.0033	0.1001 ± 0.0011	0.0741 ± 0.0009	0.8144 ± 0.0101	0.7116 ± 0.0156	0.7614 ± 0.0036
	SNGP	0.3480 ± 0.0032	0.1001 ± 0.0011	0.0745 ± 0.0009	0.8186 ± 0.0099	0.7097 ± 0.0156	0.7601 ± 0.0036
		0.3140 ± 0.0015	0.0985 ± 0.0005	0.0690 ± 0.0029	0.8791 ± 0.0062	0.5031 ± 0.0272	0.6790 ± 0.0047
	EDL	0.3133 ± 0.0015	0.0985 ± 0.0005	0.0700 ± 0.0030	0.8823 ± 0.0062	0.4961 ± 0.0262	0.6773 ± 0.0047
ResNet-50		0.9423 ± 0.0568	0.2895 ± 0.0136	0.4160 ± 0.0130	0.8947 ± 0.0090	0.4862 ± 0.0443	0.6679 ± 0.0097
	Temp. Scale	0.8451 ± 0.0563	0.2548 ± 0.0139	0.3722 ± 0.0143	0.9034 ± 0.0087	0.4751 ± 0.0431	0.6638 ± 0.0096
		1.5000 ± 0.0883	0.4036 ± 0.0133	0.5285 ± 0.0111	0.8806 ± 0.0116	0.5305 ± 0.0539	0.6779 ± 0.0117
	MC Drop.	0.9067 ± 0.0618	0.2707 ± 0.0147	0.3917 ± 0.0148	0.9023 ± 0.0087	0.4737 ± 0.0408	0.6644 ± 0.0094
		1.2227 ± 0.0116	0.3458 ± 0.0028	0.4785 ± 0.0025	0.8948 ± 0.0030	0.4945 ± 0.0173	0.6632 ± 0.0045
	Deep Ens.	0.8303 ± 0.0469	0.2467 ± 0.0108	0.3659 ± 0.0110	0.9025 ± 0.0076	0.4735 ± 0.0339	0.6599 ± 0.0085
		1.5595 ± 0.0748	0.4140 ± 0.0109	0.5375 ± 0.0085	0.8812 ± 0.0101	0.5216 ± 0.0465	0.6758 ± 0.0091
	SWAG	0.9122 ± 0.0607	0.2726 ± 0.0144	0.3939 ± 0.0141	0.9019 ± 0.0087	0.4736 ± 0.0417	0.6644 ± 0.0091
		0.6677 ± 0.2302	0.2089 ± 0.0642	0.3169 ± 0.0754	0.8880 ± 0.0135	0.5047 ± 0.0594	0.6760 ± 0.0118
	VBLL	0.5730 ± 0.1528	0.1780 ± 0.0442	0.2728 ± 0.0578	0.9017 ± 0.0105	0.4773 ± 0.0504	0.6689 ± 0.0100
		0.3130 ± 0.0129	0.1009 ± 0.0052	0.0682 ± 0.0044	0.8615 ± 0.0090	0.6023 ± 0.0346	0.7022 ± 0.0069
	SNGP	0.2992 ± 0.0130	0.0956 ± 0.0053	0.0589 ± 0.0050	0.8716 ± 0.0090	0.5824 ± 0.0381	0.6972 ± 0.0073
		0.3007 ± 0.0037	0.0904 ± 0.0013	0.0763 ± 0.0042	0.8901 ± 0.0055	0.4533 ± 0.0289	0.6549 ± 0.0039
	EDL	0.2853 ± 0.0037	0.0869 ± 0.0013	0.0720 ± 0.0039	0.9088 ± 0.0050	0.4195 ± 0.0250	0.6462 ± 0.0034
VGG-16		1.1841 ± 0.0373	0.3487 ± 0.0085	0.4740 ± 0.0085	0.8352 ± 0.0037	0.6097 ± 0.0115	0.7095 ± 0.0047
	Temp. Scale	0.7680 ± 0.0366	0.2237 ± 0.0086	0.3248 ± 0.0111	0.8832 ± 0.0050	0.5241 ± 0.0171	0.6849 ± 0.0069
		4.8199 ± 0.0893	0.6609 ± 0.0025	0.7337 ± 0.0021	0.7971 ± 0.0045	0.6300 ± 0.0057	0.7223 ± 0.0046
	MC Drop.	0.9909 ± 0.0392	0.2879 ± 0.0089	0.4074 ± 0.0101	0.8668 ± 0.0046	0.5520 ± 0.0126	0.6922 ± 0.0061
		4.0105 ± 0.0239	0.6041 ± 0.0023	0.6931 ± 0.0016	0.8063 ± 0.0047	0.6143 ± 0.0083	0.7160 ± 0.0052
	Deep Ens.	0.9622 ± 0.0363	0.2775 ± 0.0080	0.3943 ± 0.0093	0.8652 ± 0.0038	0.5552 ± 0.0115	0.6913 ± 0.0063
		4.6719 ± 0.0377	0.6679 ± 0.0029	0.7384 ± 0.0021	0.7956 ± 0.0047	0.6349 ± 0.0054	0.7235 ± 0.0047
	SWAG	1.0043 ± 0.0389	0.2929 ± 0.0091	0.4144 ± 0.0096	0.8667 ± 0.0055	0.5522 ± 0.0119	0.6921 ± 0.0063
		0.6521 ± 0.0848	0.2039 ± 0.0293	0.3035 ± 0.0397	0.8194 ± 0.0089	0.6358 ± 0.0218	0.7341 ± 0.0067
	VBLL	0.4108 ± 0.0474	0.1228 ± 0.0163	0.1767 ± 0.0289	0.8858 ± 0.0081	0.5248 ± 0.0272	0.7013 ± 0.0096
		0.0648 ± 0.0080	0.0121 ± 0.0019	0.0016 ± 0.0008	0.7438 ± 0.0216	0.6759 ± 0.1261	0.9684 ± 0.0072
	SNGP	0.0646 ± 0.0080	0.0121 ± 0.0019	0.0018 ± 0.0008	0.8711 ± 0.0202	0.5933 ± 0.1606	0.9624 ± 0.0080
		0.3726 ± 0.0352	0.1279 ± 0.0140	0.1391 ± 0.0204	0.8600 ± 0.0183	0.6033 ± 0.0230	0.7180 ± 0.0105
	EDL	0.2634 ± 0.0230	0.0824 ± 0.0088	0.0625 ± 0.0125	0.8963 ± 0.0152	0.4739 ± 0.0515	0.6930 ± 0.0113
DenseNet		0.8910 ± 0.0440	0.2843 ± 0.0127	0.4103 ± 0.0135	0.8835 ± 0.0078	0.5361 ± 0.0449	0.6862 ± 0.0081
	Temp. Scale	0.7728 ± 0.0504	0.2427 ± 0.0158	0.3557 ± 0.0194	0.8971 ± 0.0089	0.5097 ± 0.0523	0.6788 ± 0.0089
		1.1651 ± 0.0363	0.3543 ± 0.0119	0.4835 ± 0.0114	0.8742 ± 0.0077	0.5650 ± 0.0323	0.6921 ± 0.0053
	MC Drop.	0.8068 ± 0.0489	0.2517 ± 0.0149	0.3679 ± 0.0176	0.8972 ± 0.0099	0.5047 ± 0.0531	0.6781 ± 0.0086
		0.9195 ± 0.0041	0.2889 ± 0.0018	0.4219 ± 0.0018	0.8935 ± 0.0025	0.5030 ± 0.0131	0.6664 ± 0.0022
	Deep Ens.	0.6730 ± 0.0236	0.2115 ± 0.0080	0.3224 ± 0.0109	0.9024 ± 0.0054	0.4905 ± 0.0285	0.6623 ± 0.0051
		1.3170 ± 0.0504	0.3811 ± 0.0140	0.5070 ± 0.0126	0.8730 ± 0.0081	0.5555 ± 0.0352	0.6919 ± 0.0051
	SWAG	0.8203 ± 0.0504	0.2561 ± 0.0154	0.3729 ± 0.0177	0.8972 ± 0.0100	0.5043 ± 0.0522	0.6781 ± 0.0083
		0.7821 ± 0.4641	0.2380 ± 0.1166	0.3486 ± 0.1260	0.8781 ± 0.0244	0.5279 ± 0.0663	0.6895 ± 0.0133
	VBLL	0.5399 ± 0.1145	0.1725 ± 0.0386	0.2658 ± 0.0513	0.9026 ± 0.0149	0.4745 ± 0.0478	0.6768 ± 0.0080
		0.3428 ± 0.0042	0.1030 ± 0.0010	0.0758 ± 0.0030	0.8556 ± 0.0192	0.6240 ± 0.0377	0.7228 ± 0.0092
	SNGP	0.3412 ± 0.0039	0.1029 ± 0.0010	0.0788 ± 0.0030	0.8717 ± 0.0175	0.5873 ± 0.0466	0.7150 ± 0.0090
		0.3160 ± 0.0040	0.0993 ± 0.0009	0.0634 ± 0.0039	0.8701 ± 0.0067	0.5683 ± 0.0409	0.6897 ± 0.0074
	EDL	0.3068 ± 0.0045	0.0977 ± 0.0009	0.0698 ± 0.0047	0.8915 ± 0.0080	0.5216 ± 0.0339	0.6765 ± 0.0076
ViT-Tiny		1.2350 ± 0.0562	0.3679 ± 0.0118	0.4895 ± 0.0107	0.8703 ± 0.0070	0.5984 ± 0.0355	0.6776 ± 0.0049
	Temp. Scale	0.9470 ± 0.0839	0.2677 ± 0.0209	0.3631 ± 0.0218	0.8829 ± 0.0105	0.5813 ± 0.0539	0.6718 ± 0.0081
		2.2569 ± 0.0800	0.5137 ± 0.0111	0.6194 ± 0.0093	0.8496 ± 0.0073	0.6374 ± 0.0347	0.6903 ± 0.0059
	MC Drop.	1.1120 ± 0.1066	0.3013 ± 0.0216	0.4067 ± 0.0211	0.8830 ± 0.0109	0.5676 ± 0.0632	0.6708 ± 0.0097
		1.9411 ± 0.0095	0.4484 ± 0.0030	0.5665 ± 0.0024	0.8573 ± 0.0042	0.6194 ± 0.0102	0.6808 ± 0.0044
	Deep Ens.	1.0363 ± 0.0987	0.2780 ± 0.0187	0.3808 ± 0.0188	0.8796 ± 0.0111	0.5757 ± 0.0615	0.6683 ± 0.0103
		2.5346 ± 0.0863	0.5453 ± 0.0109	0.6437 ± 0.0092	0.8475 ± 0.0080	0.6296 ± 0.0324	0.6892 ± 0.0055
	SWAG	1.1216 ± 0.1076	0.3046 ± 0.0219	0.4108 ± 0.0214	0.8829 ± 0.0110	0.5680 ± 0.0630	0.6707 ± 0.0097
		0.7363 ± 0.2544	0.2359 ± 0.0647	0.3415 ± 0.0696	0.8818 ± 0.0172	0.5247 ± 0.0747	0.6640 ± 0.0127
	VBLL	0.5816 ± 0.1904	0.1806 ± 0.0461	0.2456 ± 0.0567	0.8885 ± 0.0157	0.5220 ± 0.0771	0.6617 ± 0.0127
		0.3177 ± 0.0054	0.0953 ± 0.0017	0.0479 ± 0.0054	0.8239 ± 0.0053	0.6623 ± 0.0417	0.7288 ± 0.0076
	SNGP	0.2952 ± 0.0053	0.0918 ± 0.0014	0.0548 ± 0.0074	0.8847 ± 0.0077	0.5730 ± 0.0199	0.6981 ± 0.0063
		0.3431 ± 0.0053	0.1057 ± 0.0015	0.0801 ± 0.0117	0.8577 ± 0.0120	0.5325 ± 0.0587	0.6694 ± 0.0081
	EDL	0.3076 ± 0.0077	0.0979 ± 0.0020	0.0639 ± 0.0096	0.8893 ± 0.0141	0.5354 ± 0.0642	0.6550 ± 0.0081

Table 7: Per-backbone main classification UQ results for ViT-B/16 and Swin-T. In each metric cell, the upper value is Base and the lower value is **+D-CUQ**. Calibration and failure-detection metrics are grouped separately. Values are mean ± standard deviation over 5 seeds. Bold values are best within each pair.

		Calibration			Failure detection		
Backbone	Baseline	NLL↓	Brier↓	ECE↓	F-AUROC↑	F-FPR95↓	AURC↓
ViT-B/16		1.1720 ± 0.0382	0.3631 ± 0.0099	0.4800 ± 0.0101	0.8903 ± 0.0105	0.5875 ± 0.0741	0.6595 ± 0.0117
	Temp. Scale	0.8876 ± 0.0366	0.2639 ± 0.0091	0.3575 ± 0.0103	0.9005 ± 0.0100	0.5754 ± 0.0775	0.6547 ± 0.0115
		2.3593 ± 0.0635	0.5293 ± 0.0106	0.6267 ± 0.0088	0.8805 ± 0.0083	0.5536 ± 0.0537	0.6614 ± 0.0088
	MC Drop.	1.0060 ± 0.0447	0.2925 ± 0.0106	0.3940 ± 0.0122	0.9015 ± 0.0097	0.5544 ± 0.0814	0.6530 ± 0.0110
		1.9750 ± 0.0158	0.4620 ± 0.0018	0.5759 ± 0.0014	0.8863 ± 0.0031	0.5233 ± 0.0167	0.6559 ± 0.0032
	Deep Ens.	0.9463 ± 0.0393	0.2730 ± 0.0088	0.3727 ± 0.0111	0.8998 ± 0.0090	0.5587 ± 0.0799	0.6522 ± 0.0111
		2.4028 ± 0.0713	0.5327 ± 0.0112	0.6292 ± 0.0092	0.8807 ± 0.0095	0.5505 ± 0.0635	0.6606 ± 0.0094
	SWAG	1.0086 ± 0.0458	0.2931 ± 0.0108	0.3947 ± 0.0122	0.9012 ± 0.0097	0.5546 ± 0.0815	0.6530 ± 0.0110
		0.5569 ± 0.0478	0.1864 ± 0.0179	0.2821 ± 0.0189	0.8866 ± 0.0191	0.5628 ± 0.0890	0.6639 ± 0.0138
	VBLL	0.4397 ± 0.0428	0.1435 ± 0.0159	0.2003 ± 0.0159	0.8991 ± 0.0175	0.5445 ± 0.1077	0.6576 ± 0.0141
		0.1486 ± 0.0028	0.0316 ± 0.0006	0.0510 ± 0.0023	0.7467 ± 0.0134	0.7118 ± 0.0297	0.9318 ± 0.0028
	SNGP	0.1255 ± 0.0021	0.0280 ± 0.0005	0.0340 ± 0.0015	0.8713 ± 0.0053	0.5881 ± 0.0174	0.9132 ± 0.0022
		0.3236 ± 0.0192	0.1022 ± 0.0059	0.0566 ± 0.0175	0.8416 ± 0.0258	0.5581 ± 0.0208	0.6926 ± 0.0133
	EDL	0.2815 ± 0.0115	0.0886 ± 0.0031	0.0483 ± 0.0058	0.8946 ± 0.0133	0.5645 ± 0.0305	0.6557 ± 0.0081
Swin-T		1.1885 ± 0.0304	0.3533 ± 0.0062	0.4805 ± 0.0056	0.8907 ± 0.0058	0.5401 ± 0.0339	0.6564 ± 0.0063
	Temp. Scale	0.8264 ± 0.0285	0.2322 ± 0.0060	0.3347 ± 0.0061	0.9094 ± 0.0049	0.4951 ± 0.0415	0.6462 ± 0.0058
		2.7077 ± 0.0653	0.5544 ± 0.0055	0.6530 ± 0.0039	0.8610 ± 0.0071	0.5808 ± 0.0367	0.6701 ± 0.0073
	MC Drop.	1.0115 ± 0.0413	0.2819 ± 0.0078	0.3985 ± 0.0077	0.9072 ± 0.0048	0.4818 ± 0.0362	0.6459 ± 0.0056
		2.3501 ± 0.0187	0.5042 ± 0.0034	0.6164 ± 0.0023	0.8747 ± 0.0031	0.5608 ± 0.0154	0.6627 ± 0.0025
	Deep Ens.	0.9744 ± 0.0402	0.2692 ± 0.0074	0.3847 ± 0.0073	0.9073 ± 0.0053	0.4822 ± 0.0365	0.6454 ± 0.0056
		2.7993 ± 0.0783	0.5638 ± 0.0053	0.6605 ± 0.0036	0.8640 ± 0.0063	0.5701 ± 0.0313	0.6670 ± 0.0070
	SWAG	1.0088 ± 0.0430	0.2810 ± 0.0085	0.3982 ± 0.0083	0.9069 ± 0.0044	0.4831 ± 0.0352	0.6455 ± 0.0057
		0.4877 ± 0.0136	0.1559 ± 0.0048	0.2437 ± 0.0079	0.9052 ± 0.0051	0.4792 ± 0.0330	0.6521 ± 0.0073
	VBLL	0.3596 ± 0.0111	0.1104 ± 0.0039	0.1497 ± 0.0081	0.9165 ± 0.0029	0.4354 ± 0.0184	0.6453 ± 0.0061
		0.3881 ± 0.0061	0.1181 ± 0.0026	0.1695 ± 0.0034	0.8680 ± 0.0065	0.6258 ± 0.0182	0.6978 ± 0.0051
	SNGP	0.2902 ± 0.0075	0.0860 ± 0.0031	0.0887 ± 0.0047	0.9122 ± 0.0063	0.5330 ± 0.0404	0.6766 ± 0.0059
		0.3335 ± 0.0086	0.1033 ± 0.0026	0.0762 ± 0.0039	0.8585 ± 0.0086	0.5683 ± 0.0226	0.6711 ± 0.0041
	EDL	0.2739 ± 0.0043	0.0850 ± 0.0011	0.0554 ± 0.0023	0.9096 ± 0.0038	0.4653 ± 0.0354	0.6395 ± 0.0021

Table 8: Architecture-macro predictive UQ on ID and covariate-shifted ID. Each value is mean ± standard deviation over 5 seed-wise architecture macros.

		Clean ID			Shift		
Baseline	Variant	NLL↓	Brier↓	ECE↓	NLL↓	Brier↓	ECE↓
Temp. Scale	Base	0.5932 ± 0.0135	0.2443 ± 0.0069	0.0241 ± 0.0019	6.0694 ± 0.1429	1.1705 ± 0.0233	0.3393 ± 0.0283
	+ D-CUQ	0.6097 ± 0.0150	0.2480 ± 0.0071	0.0291 ± 0.0005	5.1107 ± 0.0717	1.0480 ± 0.0109	0.1660 ± 0.0208
MC Drop.	Base	0.7300 ± 0.0202	0.2584 ± 0.0087	0.0702 ± 0.0040	9.2481 ± 0.2423	1.3397 ± 0.0241	0.4983 ± 0.0256
	+ D-CUQ	0.6015 ± 0.0152	0.2459 ± 0.0074	0.0296 ± 0.0024	5.3949 ± 0.1063	1.0846 ± 0.0148	0.2196 ± 0.0225
Deep Ens.	Base	0.6195 ± 0.0138	0.2315 ± 0.0073	0.0540 ± 0.0026	8.5878 ± 0.0284	1.2456 ± 0.0014	0.4095 ± 0.0011
	+ D-CUQ	0.5516 ± 0.0128	0.2295 ± 0.0066	0.0365 ± 0.0016	5.3633 ± 0.0877	1.0626 ± 0.0068	0.1830 ± 0.0096
SWAG	Base	0.7375 ± 0.0192	0.2603 ± 0.0084	0.0771 ± 0.0032	9.1764 ± 0.2381	1.3589 ± 0.0287	0.5172 ± 0.0306
	+ D-CUQ	0.6027 ± 0.0152	0.2458 ± 0.0072	0.0284 ± 0.0023	5.3838 ± 0.0954	1.0862 ± 0.0152	0.2233 ± 0.0240
VBLL	Base	0.7742 ± 0.0186	0.2932 ± 0.0073	0.1032 ± 0.0147	5.6190 ± 0.6427	1.0882 ± 0.0415	0.2311 ± 0.0465
	+ D-CUQ	0.8080 ± 0.0375	0.3029 ± 0.0110	0.1177 ± 0.0233	4.8876 ± 0.1136	1.0135 ± 0.0096	0.0957 ± 0.0168
SNGP	Base	2.6204 ± 0.0174	0.7506 ± 0.0041	0.3775 ± 0.0041	4.6424 ± 0.0126	0.9927 ± 0.0012	0.0333 ± 0.0029
	+ D-CUQ	2.6433 ± 0.0189	0.7546 ± 0.0042	0.3828 ± 0.0039	4.5721 ± 0.0071	0.9886 ± 0.0003	0.0131 ± 0.0013
EDL	Base	1.8447 ± 0.0216	0.5882 ± 0.0041	0.5446 ± 0.0031	4.6410 ± 0.0202	0.9999 ± 0.0067	0.0535 ± 0.0170
	+ D-CUQ	1.8946 ± 0.0237	0.6040 ± 0.0046	0.5593 ± 0.0025	4.6067 ± 0.0041	0.9906 ± 0.0004	0.0200 ± 0.0034

Table 9: Per-backbone clean-ID and covariate-shift UQ for CNN and ViT-Tiny backbones. In each metric cell, the upper value is Base and the lower value is **+D-CUQ**. ID and shift metrics are grouped separately. Values are mean ± standard deviation over 5 seeds. Bold values are best within each pair.

		Clean ID			Shift		
Backbone	Baseline	NLL↓	Brier↓	ECE↓	NLL↓	Brier↓	ECE↓
ResNet-18		0.7240 ± 0.0073	0.3048 ± 0.0059	0.0206 ± 0.0013	6.1306 ± 0.2563	1.1576 ± 0.0546	0.3340 ± 0.0792
	Temp. Scale	0.7273 ± 0.0082	0.3054 ± 0.0060	0.0216 ± 0.0046	5.7557 ± 0.1867	1.1102 ± 0.0400	0.2761 ± 0.0690
		0.7253 ± 0.0078	0.3045 ± 0.0056	0.0187 ± 0.0065	6.0581 ± 0.2628	1.1444 ± 0.0479	0.3182 ± 0.0694
	MC Drop.	0.7281 ± 0.0085	0.3051 ± 0.0059	0.0256 ± 0.0050	5.7559 ± 0.1920	1.1080 ± 0.0381	0.2720 ± 0.0643
		0.6544 ± 0.0089	0.2780 ± 0.0044	0.0494 ± 0.0072	5.6830 ± 0.0208	1.0979 ± 0.0023	0.2734 ± 0.0034
	Deep Ens.	0.6693 ± 0.0096	0.2827 ± 0.0047	0.0659 ± 0.0059	5.4665 ± 0.0762	1.0692 ± 0.0120	0.2305 ± 0.0211
		0.7325 ± 0.0080	0.3059 ± 0.0055	0.0270 ± 0.0046	6.2498 ± 0.2924	1.1799 ± 0.0628	0.3562 ± 0.0831
	SWAG	0.7309 ± 0.0090	0.3052 ± 0.0059	0.0245 ± 0.0073	5.7314 ± 0.1821	1.1116 ± 0.0404	0.2766 ± 0.0684
		0.8791 ± 0.0349	0.3496 ± 0.0051	0.0980 ± 0.0275	5.9850 ± 1.2279	1.1037 ± 0.0976	0.2560 ± 0.1199
	VBLL	0.8581 ± 0.0422	0.3458 ± 0.0101	0.0852 ± 0.0464	5.3660 ± 0.2590	1.0482 ± 0.0248	0.1827 ± 0.0493
		3.0079 ± 0.0183	0.8963 ± 0.0021	0.6212 ± 0.0071	4.5965 ± 0.0196	0.9897 ± 0.0006	0.0140 ± 0.0092
	SNGP	3.0102 ± 0.0186	0.8965 ± 0.0021	0.6213 ± 0.0071	4.5938 ± 0.0178	0.9897 ± 0.0005	0.0128 ± 0.0082
		2.3675 ± 0.0214	0.7639 ± 0.0025	0.6517 ± 0.0039	4.6117 ± 0.0186	0.9911 ± 0.0015	0.0340 ± 0.0115
	EDL	2.3713 ± 0.0220	0.7646 ± 0.0026	0.6522 ± 0.0038	4.6073 ± 0.0166	0.9906 ± 0.0012	0.0300 ± 0.0094
ResNet-50		0.6420 ± 0.0198	0.2617 ± 0.0088	0.0344 ± 0.0079	6.3882 ± 0.3861	1.1963 ± 0.0827	0.3677 ± 0.0966
	Temp. Scale	0.6471 ± 0.0201	0.2626 ± 0.0088	0.0356 ± 0.0088	5.5093 ± 0.2464	1.0903 ± 0.0460	0.2401 ± 0.0716
		0.6855 ± 0.0254	0.2695 ± 0.0110	0.0725 ± 0.0067	7.7070 ± 0.6144	1.3111 ± 0.0967	0.4809 ± 0.0940
	MC Drop.	0.6475 ± 0.0222	0.2627 ± 0.0097	0.0378 ± 0.0096	5.7258 ± 0.2863	1.1146 ± 0.0539	0.2707 ± 0.0756
		0.5956 ± 0.0226	0.2456 ± 0.0095	0.0435 ± 0.0040	7.1435 ± 0.0465	1.1711 ± 0.0040	0.3311 ± 0.0037
	Deep Ens.	0.5970 ± 0.0202	0.2473 ± 0.0088	0.0323 ± 0.0035	5.7115 ± 0.2619	1.0658 ± 0.0211	0.1908 ± 0.0341
		0.6996 ± 0.0244	0.2718 ± 0.0104	0.0787 ± 0.0047	7.7881 ± 0.5656	1.3248 ± 0.0927	0.4927 ± 0.0894
	SWAG	0.6494 ± 0.0216	0.2633 ± 0.0093	0.0389 ± 0.0059	5.7098 ± 0.2773	1.1162 ± 0.0537	0.2730 ± 0.0758
		0.7689 ± 0.0495	0.2960 ± 0.0145	0.0708 ± 0.0219	6.0489 ± 0.9982	1.1386 ± 0.0777	0.3021 ± 0.0816
	VBLL	0.7752 ± 0.0603	0.2973 ± 0.0172	0.0742 ± 0.0292	5.2890 ± 0.5318	1.0544 ± 0.0517	0.1815 ± 0.0672
		1.6437 ± 0.0342	0.4774 ± 0.0098	0.3356 ± 0.0069	4.5888 ± 0.0119	0.9895 ± 0.0012	0.0110 ± 0.0100
	SNGP	1.6530 ± 0.0351	0.4792 ± 0.0100	0.3378 ± 0.0069	4.5834 ± 0.0075	0.9889 ± 0.0005	0.0094 ± 0.0030
		1.7320 ± 0.0198	0.5825 ± 0.0032	0.5695 ± 0.0048	4.6755 ± 0.0399	1.0019 ± 0.0087	0.0952 ± 0.0411
	EDL	1.7495 ± 0.0210	0.5871 ± 0.0036	0.5732 ± 0.0046	4.6252 ± 0.0162	0.9926 ± 0.0019	0.0464 ± 0.0193
VGG-16		0.7377 ± 0.0291	0.2960 ± 0.0116	0.0243 ± 0.0031	7.4599 ± 0.4230	1.3484 ± 0.0447	0.5235 ± 0.0529
	Temp. Scale	0.7689 ± 0.0294	0.3047 ± 0.0116	0.0400 ± 0.0063	5.3134 ± 0.1809	1.0622 ± 0.0228	0.2049 ± 0.0417
		1.3782 ± 0.0663	0.3457 ± 0.0170	0.1424 ± 0.0081	21.1236 ± 0.7543	1.7368 ± 0.0767	0.8188 ± 0.0577
	MC Drop.	0.7462 ± 0.0298	0.2984 ± 0.0117	0.0309 ± 0.0028	6.3552 ± 0.3072	1.2043 ± 0.0400	0.3791 ± 0.0530
		1.1893 ± 0.0471	0.3208 ± 0.0140	0.1197 ± 0.0064	19.8218 ± 0.0923	1.6404 ± 0.0048	0.7463 ± 0.0038
	Deep Ens.	0.7325 ± 0.0293	0.2928 ± 0.0105	0.0345 ± 0.0074	6.4752 ± 0.3194	1.1858 ± 0.0327	0.3434 ± 0.0346
		1.3300 ± 0.0524	0.3454 ± 0.0156	0.1437 ± 0.0071	19.9983 ± 0.8413	1.7669 ± 0.0713	0.8425 ± 0.0518
	SWAG	0.7442 ± 0.0280	0.2972 ± 0.0111	0.0279 ± 0.0022	6.3191 ± 0.2947	1.2111 ± 0.0410	0.3963 ± 0.0509
		1.0575 ± 0.0680	0.3933 ± 0.0182	0.1044 ± 0.0288	6.2906 ± 0.3726	1.1941 ± 0.0804	0.3779 ± 0.0870
	VBLL	1.1305 ± 0.0765	0.4120 ± 0.0203	0.1412 ± 0.0297	4.9951 ± 0.0738	1.0156 ± 0.0086	0.1151 ± 0.0247
		4.5767 ± 0.0141	0.9890 ± 0.0010	0.0626 ± 0.0112	4.6062 ± 0.0006	0.9900 ± 0.0000	0.0015 ± 0.0014
	SNGP	4.5769 ± 0.0141	0.9890 ± 0.0010	0.0627 ± 0.0112	4.6056 ± 0.0003	0.9900 ± 0.0000	0.0014 ± 0.0016
		1.7621 ± 0.0398	0.4291 ± 0.0096	0.2477 ± 0.0031	4.7288 ± 0.1335	1.0330 ± 0.0483	0.1062 ± 0.1042
	EDL	1.8561 ± 0.0427	0.4658 ± 0.0107	0.2903 ± 0.0026	4.6170 ± 0.0140	0.9915 ± 0.0016	0.0180 ± 0.0166
DenseNet		0.7076 ± 0.0181	0.2898 ± 0.0079	0.0259 ± 0.0045	5.3405 ± 0.2159	1.0743 ± 0.0424	0.2226 ± 0.0658
	Temp. Scale	0.7114 ± 0.0189	0.2905 ± 0.0079	0.0261 ± 0.0027	4.7087 ± 0.0867	0.9973 ± 0.0078	0.0579 ± 0.0287
		0.7181 ± 0.0217	0.2918 ± 0.0086	0.0482 ± 0.0063	5.7958 ± 0.3092	1.1371 ± 0.0555	0.3080 ± 0.0679
	MC Drop.	0.7105 ± 0.0201	0.2898 ± 0.0082	0.0293 ± 0.0089	4.7551 ± 0.1071	1.0010 ± 0.0104	0.0712 ± 0.0308
		0.5804 ± 0.0172	0.2481 ± 0.0084	0.0216 ± 0.0020	5.4015 ± 0.0241	1.0764 ± 0.0026	0.2416 ± 0.0037
	Deep Ens.	0.6077 ± 0.0171	0.2560 ± 0.0079	0.0491 ± 0.0024	4.6796 ± 0.0499	0.9948 ± 0.0031	0.0533 ± 0.0134
		0.7566 ± 0.0235	0.2997 ± 0.0085	0.0698 ± 0.0065	5.8856 ± 0.3469	1.1611 ± 0.0594	0.3365 ± 0.0683
	SWAG	0.7206 ± 0.0199	0.2923 ± 0.0079	0.0299 ± 0.0081	4.7538 ± 0.1094	1.0013 ± 0.0105	0.0732 ± 0.0302
		0.8466 ± 0.0273	0.3298 ± 0.0051	0.0808 ± 0.0264	6.1827 ± 2.0250	1.0911 ± 0.1199	0.2148 ± 0.1560
	VBLL	0.8321 ± 0.0412	0.3285 ± 0.0128	0.0750 ± 0.0400	4.7053 ± 0.0763	0.9950 ± 0.0046	0.0416 ± 0.0257
		2.7767 ± 0.0256	0.8522 ± 0.0042	0.6548 ± 0.0063	4.5887 ± 0.0030	0.9896 ± 0.0001	0.0065 ± 0.0013
	SNGP	2.7886 ± 0.0274	0.8533 ± 0.0044	0.6555 ± 0.0062	4.5923 ± 0.0017	0.9897 ± 0.0000	0.0097 ± 0.0008
		2.1885 ± 0.0163	0.7224 ± 0.0026	0.6343 ± 0.0053	4.6121 ± 0.0318	0.9917 ± 0.0035	0.0292 ± 0.0339
	EDL	2.2071 ± 0.0188	0.7266 ± 0.0031	0.6375 ± 0.0051	4.5992 ± 0.0069	0.9899 ± 0.0003	0.0106 ± 0.0066
ViT-Tiny		0.4953 ± 0.0192	0.2037 ± 0.0106	0.0205 ± 0.0062	5.6875 ± 0.2045	1.0911 ± 0.0406	0.2438 ± 0.0648
	Temp. Scale	0.5210 ± 0.0201	0.2089 ± 0.0103	0.0294 ± 0.0073	4.8053 ± 0.0806	1.0028 ± 0.0104	0.0681 ± 0.0349
		0.5671 ± 0.0218	0.2147 ± 0.0118	0.0679 ± 0.0057	7.2774 ± 0.4043	1.2639 ± 0.0813	0.4377 ± 0.0838
	MC Drop.	0.5108 ± 0.0200	0.2067 ± 0.0106	0.0298 ± 0.0058	4.9702 ± 0.1205	1.0156 ± 0.0179	0.1025 ± 0.0454
		0.4507 ± 0.0158	0.1830 ± 0.0105	0.0364 ± 0.0039	6.5809 ± 0.0225	1.1714 ± 0.0028	0.3572 ± 0.0020
	Deep Ens.	0.4514 ± 0.0132	0.1883 ± 0.0090	0.0334 ± 0.0066	4.9266 ± 0.1285	1.0073 ± 0.0101	0.0867 ± 0.0284
		0.6028 ± 0.0251	0.2196 ± 0.0129	0.0775 ± 0.0062	7.4891 ± 0.4644	1.3003 ± 0.0941	0.4715 ± 0.0924
	SWAG	0.5112 ± 0.0207	0.2068 ± 0.0110	0.0288 ± 0.0064	4.9654 ± 0.1185	1.0164 ± 0.0183	0.1053 ± 0.0471
		0.6109 ± 0.0470	0.2296 ± 0.0143	0.0867 ± 0.0367	5.0993 ± 0.0949	1.0341 ± 0.0224	0.1546 ± 0.0546
	VBLL	0.6743 ± 0.0625	0.2465 ± 0.0181	0.1142 ± 0.0422	4.6746 ± 0.0325	0.9943 ± 0.0044	0.0392 ± 0.0275
		1.8602 ± 0.0278	0.6954 ± 0.0066	0.5755 ± 0.0100	4.6519 ± 0.0540	0.9909 ± 0.0022	0.0258 ± 0.0150
	SNGP	1.9294 ± 0.0323	0.7054 ± 0.0069	0.5848 ± 0.0099	4.5769 ± 0.0197	0.9890 ± 0.0006	0.0065 ± 0.0049
		1.7354 ± 0.0185	0.6052 ± 0.0041	0.6199 ± 0.0026	4.6363 ± 0.0431	0.9953 ± 0.0073	0.0528 ± 0.0491
	EDL	1.8169 ± 0.0229	0.6267 ± 0.0050	0.6374 ± 0.0025	4.6033 ± 0.0070	0.9900 ± 0.0005	0.0119 ± 0.0121

Table 10: Per-backbone clean-ID and covariate-shift UQ for ViT-B/16 and Swin-T. In each metric cell, the upper value is Base and the lower value is **+D-CUQ**. Clean-ID and shift metrics are grouped separately. Values are mean ± standard deviation over 5 seeds. Bold values are best within each pair.

		Clean ID			Shift		
Backbone	Baseline	NLL↓	Brier↓	ECE↓	NLL↓	Brier↓	ECE↓
ViT-B/16	Temp. Scale	0.4518 ± 0.0125	0.1795 ± 0.0059	0.0262 ± 0.0055	5.3594 ± 0.2632	1.1246 ± 0.0394	0.3002 ± 0.0465
		0.4709 ± 0.0132	0.1828 ± 0.0063	0.0263 ± 0.0053	4.8401 ± 0.1889	1.0385 ± 0.0282	0.1620 ± 0.0532
	MC Drop.	0.5342 ± 0.0219	0.1936 ± 0.0064	0.0740 ± 0.0048	6.8769 ± 0.5227	1.3156 ± 0.0459	0.4985 ± 0.0392
		0.4621 ± 0.0137	0.1816 ± 0.0063	0.0300 ± 0.0067	4.9788 ± 0.2440	1.0603 ± 0.0371	0.2007 ± 0.0596
	Deep Ens.	0.4298 ± 0.0100	0.1704 ± 0.0059	0.0509 ± 0.0034	6.4434 ± 0.0187	1.2291 ± 0.0029	0.4113 ± 0.0043
		0.4156 ± 0.0092	0.1692 ± 0.0055	0.0223 ± 0.0047	4.9670 ± 0.2350	1.0488 ± 0.0328	0.1746 ± 0.0575
	SWAG	0.5402 ± 0.0217	0.1942 ± 0.0062	0.0752 ± 0.0052	6.8833 ± 0.5226	1.3165 ± 0.0439	0.4990 ± 0.0381
		0.4638 ± 0.0136	0.1818 ± 0.0062	0.0292 ± 0.0041	4.9947 ± 0.2469	1.0613 ± 0.0370	0.2026 ± 0.0595
	VBLL	0.6372 ± 0.0116	0.2182 ± 0.0046	0.1363 ± 0.0081	4.8913 ± 0.1036	1.0261 ± 0.0129	0.1545 ± 0.0259
		0.6831 ± 0.0147	0.2312 ± 0.0053	0.1579 ± 0.0086	4.6577 ± 0.0689	0.9975 ± 0.0063	0.0709 ± 0.0221
	SNGP	3.4319 ± 0.0226	0.9265 ± 0.0013	0.0892 ± 0.0026	4.7124 ± 0.0696	0.9936 ± 0.0035	0.0701 ± 0.0067
		3.4367 ± 0.0227	0.9267 ± 0.0012	0.0929 ± 0.0025	4.5247 ± 0.0432	0.9861 ± 0.0017	0.0292 ± 0.0032
	EDL	1.6646 ± 0.0322	0.5327 ± 0.0059	0.5530 ± 0.0098	4.6140 ± 0.0375	0.9947 ± 0.0103	0.0281 ± 0.0386
		1.7098 ± 0.0311	0.5473 ± 0.0056	0.5656 ± 0.0099	4.6018 ± 0.0103	0.9905 ± 0.0017	0.0150 ± 0.0095
Swin-T	Temp. Scale	0.3939 ± 0.0111	0.1745 ± 0.0083	0.0164 ± 0.0049	6.1193 ± 0.1856	1.2014 ± 0.0608	0.3836 ± 0.0665
		0.4215 ± 0.0123	0.1811 ± 0.0080	0.0244 ± 0.0031	4.8420 ± 0.1012	1.0346 ± 0.0251	0.1529 ± 0.0539
	MC Drop.	0.5014 ± 0.0196	0.1893 ± 0.0104	0.0675 ± 0.0045	9.8982 ± 0.5199	1.4687 ± 0.0761	0.6264 ± 0.0613
		0.4050 ± 0.0130	0.1769 ± 0.0085	0.0235 ± 0.0015	5.2235 ± 0.1739	1.0882 ± 0.0489	0.2411 ± 0.0725
	Deep Ens.	0.4366 ± 0.0142	0.1750 ± 0.0091	0.0564 ± 0.0043	9.0404 ± 0.0622	1.3327 ± 0.0041	0.5054 ± 0.0032
		0.3879 ± 0.0121	0.1705 ± 0.0079	0.0180 ± 0.0034	5.3169 ± 0.3198	1.0666 ± 0.0294	0.2018 ± 0.0436
	SWAG	0.5007 ± 0.0213	0.1852 ± 0.0103	0.0678 ± 0.0045	9.9407 ± 0.5317	1.4628 ± 0.0753	0.6217 ± 0.0619
		0.3991 ± 0.0139	0.1736 ± 0.0085	0.0194 ± 0.0011	5.2122 ± 0.1690	1.0853 ± 0.0492	0.2363 ± 0.0739
	VBLL	0.6193 ± 0.0216	0.2356 ± 0.0072	0.1456 ± 0.0185	4.8354 ± 0.0680	1.0295 ± 0.0125	0.1578 ± 0.0330
		0.7025 ± 0.0261	0.2593 ± 0.0079	0.1766 ± 0.0186	4.5256 ± 0.0349	0.9899 ± 0.0026	0.0393 ± 0.0149
	SNGP	1.0455 ± 0.0218	0.4177 ± 0.0086	0.3035 ± 0.0089	4.7523 ± 0.0318	1.0057 ± 0.0059	0.1042 ± 0.0243
		1.1084 ± 0.0268	0.4321 ± 0.0092	0.3243 ± 0.0085	4.5283 ± 0.0150	0.9871 ± 0.0008	0.0230 ± 0.0085
	EDL	1.4626 ± 0.0249	0.4816 ± 0.0050	0.5358 ± 0.0030	4.6083 ± 0.0253	0.9920 ± 0.0032	0.0292 ± 0.0261
		1.5515 ± 0.0285	0.5101 ± 0.0060	0.5591 ± 0.0030	4.5929 ± 0.0103	0.9893 ± 0.0006	0.0078 ± 0.0043

Table 11: Architecture-macro ablation study. Values are mean ± standard deviation over 5 seed-wise architecture macros.

Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
Softmax	0.8602 ± 0.0033	0.5714 ± 0.0209	0.6864 ± 0.0047	0.8088 ± 0.0041	0.6443 ± 0.0185
Global (matched)	0.8702 ± 0.0036	0.5579 ± 0.0235	0.6818 ± 0.0050	0.8249 ± 0.0045	0.6299 ± 0.0215
Global (ID-NLL)	0.8774 ± 0.0036	0.5570 ± 0.0262	0.6795 ± 0.0052	0.8349 ± 0.0043	0.6308 ± 0.0237
M-only	0.8853 ± 0.0018	0.5238 ± 0.0140	0.6722 ± 0.0031	0.8483 ± 0.0028	0.6003 ± 0.0123
G-only	0.8954 ± 0.0047	0.5016 ± 0.0298	0.6682 ± 0.0057	0.8654 ± 0.0059	0.5792 ± 0.0285
Dual (ID-only)	0.8814 ± 0.0037	0.5483 ± 0.0273	0.6771 ± 0.0053	0.8404 ± 0.0039	0.6231 ± 0.0246
Shuffled Dual	0.7165 ± 0.0099	0.7113 ± 0.0113	0.7448 ± 0.0049	0.6905 ± 0.0104	0.7535 ± 0.0115
D-CUQ (Dual)	0.8973 ± 0.0029	0.4996 ± 0.0211	0.6664 ± 0.0042	0.8660 ± 0.0040	0.5770 ± 0.0192

Table 12: Per-backbone controlled-ablation results (ResNet-18, ResNet-50, VGG-16). Values are mean ± standard deviation over 5 seeds.

Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
ResNet-18					
Softmax	0.8711 ± 0.0089	0.5819 ± 0.0322	0.6967 ± 0.0056	0.8232 ± 0.0117	0.6721 ± 0.0274
Global (matched)	0.8800 ± 0.0077	0.5518 ± 0.0333	0.6910 ± 0.0050	0.8405 ± 0.0107	0.6418 ± 0.0323
Global (ID-NLL)	0.8700 ± 0.0087	0.5829 ± 0.0333	0.6972 ± 0.0055	0.8213 ± 0.0114	0.6738 ± 0.0277
M-only	0.8749 ± 0.0077	0.5433 ± 0.0294	0.6893 ± 0.0061	0.8323 ± 0.0109	0.6351 ± 0.0284
G-only	0.9031 ± 0.0036	0.4792 ± 0.0263	0.6754 ± 0.0026	0.8787 ± 0.0061	0.5712 ± 0.0235
Dual (ID-only)	0.8713 ± 0.0089	0.5815 ± 0.0321	0.6966 ± 0.0056	0.8235 ± 0.0117	0.6719 ± 0.0272
Shuffled Dual	0.7374 ± 0.0088	0.6863 ± 0.0179	0.7464 ± 0.0041	0.7131 ± 0.0075	0.7358 ± 0.0151
D-CUQ (Dual)	0.8999 ± 0.0054	0.4972 ± 0.0234	0.6774 ± 0.0042	0.8701 ± 0.0081	0.5902 ± 0.0229
ResNet-50					
Softmax	0.8823 ± 0.0089	0.5180 ± 0.0456	0.6753 ± 0.0095	0.8371 ± 0.0105	0.5981 ± 0.0409
Global (matched)	0.8930 ± 0.0086	0.4925 ± 0.0446	0.6690 ± 0.0095	0.8538 ± 0.0100	0.5732 ± 0.0408
Global (ID-NLL)	0.8947 ± 0.0090	0.4862 ± 0.0443	0.6679 ± 0.0097	0.8566 ± 0.0107	0.5666 ± 0.0416
M-only	0.8994 ± 0.0076	0.4962 ± 0.0364	0.6660 ± 0.0080	0.8654 ± 0.0096	0.5782 ± 0.0325
G-only	0.9223 ± 0.0070	0.4328 ± 0.0421	0.6540 ± 0.0089	0.9002 ± 0.0089	0.5167 ± 0.0432
Dual (ID-only)	0.8948 ± 0.0090	0.4860 ± 0.0446	0.6678 ± 0.0097	0.8567 ± 0.0107	0.5666 ± 0.0415
Shuffled Dual	0.7078 ± 0.0133	0.7008 ± 0.0257	0.7453 ± 0.0083	0.6888 ± 0.0132	0.7429 ± 0.0264
D-CUQ (Dual)	0.9163 ± 0.0071	0.4500 ± 0.0410	0.6571 ± 0.0085	0.8902 ± 0.0091	0.5350 ± 0.0395
VGG-16					
Softmax	0.7939 ± 0.0041	0.6347 ± 0.0081	0.7239 ± 0.0043	0.7060 ± 0.0072	0.7147 ± 0.0079
Global (matched)	0.8009 ± 0.0038	0.6325 ± 0.0081	0.7219 ± 0.0043	0.7149 ± 0.0067	0.7130 ± 0.0086
Global (ID-NLL)	0.8352 ± 0.0037	0.6096 ± 0.0115	0.7095 ± 0.0047	0.7668 ± 0.0071	0.6950 ± 0.0115
M-only	0.8370 ± 0.0035	0.5893 ± 0.0099	0.7060 ± 0.0058	0.7730 ± 0.0050	0.6767 ± 0.0110
G-only	0.8417 ± 0.0052	0.5786 ± 0.0132	0.7049 ± 0.0054	0.7779 ± 0.0077	0.6703 ± 0.0136
Dual (ID-only)	0.8505 ± 0.0053	0.5812 ± 0.0174	0.7008 ± 0.0052	0.7878 ± 0.0059	0.6716 ± 0.0156
Shuffled Dual	0.7453 ± 0.0079	0.6978 ± 0.0043	0.7453 ± 0.0034	0.6793 ± 0.0093	0.7626 ± 0.0027
D-CUQ (Dual)	0.8476 ± 0.0040	0.5766 ± 0.0112	0.7018 ± 0.0061	0.7862 ± 0.0060	0.6674 ± 0.0125

Table 13: Per-backbone controlled-ablation results (DenseNet-121, Swin-T, ViT-B/16). Values are mean ± standard deviation over 5 seeds.

Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
DenseNet-121					
Softmax	0.8729 ± 0.0068	0.5579 ± 0.0394	0.6917 ± 0.0074	0.8280 ± 0.0098	0.6389 ± 0.0336
Global (matched)	0.8887 ± 0.0081	0.5233 ± 0.0428	0.6832 ± 0.0079	0.8551 ± 0.0122	0.6005 ± 0.0406
Global (ID-NLL)	0.8835 ± 0.0078	0.5359 ± 0.0449	0.6862 ± 0.0081	0.8453 ± 0.0106	0.6178 ± 0.0383
M-only	0.9146 ± 0.0134	0.4208 ± 0.0440	0.6593 ± 0.0102	0.8924 ± 0.0158	0.5046 ± 0.0440
G-only	0.8987 ± 0.0160	0.4998 ± 0.0745	0.6763 ± 0.0139	0.8767 ± 0.0196	0.5757 ± 0.0763
Dual (ID-only)	0.8869 ± 0.0057	0.5263 ± 0.0407	0.6838 ± 0.0077	0.8498 ± 0.0074	0.6081 ± 0.0339
Shuffled Dual	0.6906 ± 0.0336	0.7056 ± 0.0305	0.7563 ± 0.0139	0.6749 ± 0.0326	0.7433 ± 0.0313
D-CUQ (Dual)	0.9165 ± 0.0113	0.4331 ± 0.0514	0.6602 ± 0.0102	0.8967 ± 0.0129	0.5121 ± 0.0505
Swin-T					
Softmax	0.8602 ± 0.0053	0.5838 ± 0.0304	0.6700 ± 0.0050	0.8164 ± 0.0058	0.6362 ± 0.0259
Global (matched)	0.8715 ± 0.0053	0.5709 ± 0.0314	0.6656 ± 0.0054	0.8321 ± 0.0061	0.6260 ± 0.0274
Global (ID-NLL)	0.8907 ± 0.0058	0.5399 ± 0.0340	0.6564 ± 0.0063	0.8596 ± 0.0067	0.5986 ± 0.0304
M-only	0.8940 ± 0.0034	0.5352 ± 0.0258	0.6551 ± 0.0035	0.8652 ± 0.0035	0.5960 ± 0.0225
G-only	0.9035 ± 0.0059	0.4885 ± 0.0373	0.6472 ± 0.0061	0.8792 ± 0.0068	0.5528 ± 0.0341
Dual (ID-only)	0.8932 ± 0.0054	0.5308 ± 0.0329	0.6548 ± 0.0061	0.8629 ± 0.0062	0.5901 ± 0.0310
Shuffled Dual	0.7490 ± 0.0082	0.7286 ± 0.0168	0.7245 ± 0.0053	0.7297 ± 0.0096	0.7610 ± 0.0168
D-CUQ (Dual)	0.9033 ± 0.0044	0.5042 ± 0.0329	0.6487 ± 0.0050	0.8783 ± 0.0049	0.5658 ± 0.0295
ViT-B/16					
Softmax	0.8806 ± 0.0080	0.5519 ± 0.0560	0.6610 ± 0.0087	0.8423 ± 0.0084	0.6059 ± 0.0482
Global (matched)	0.8869 ± 0.0094	0.5764 ± 0.0689	0.6600 ± 0.0106	0.8532 ± 0.0100	0.6250 ± 0.0624
Global (ID-NLL)	0.8903 ± 0.0105	0.5874 ± 0.0742	0.6595 ± 0.0117	0.8599 ± 0.0112	0.6330 ± 0.0671
M-only	0.8919 ± 0.0081	0.5582 ± 0.0611	0.6577 ± 0.0093	0.8617 ± 0.0090	0.6114 ± 0.0545
G-only	0.9030 ± 0.0083	0.5308 ± 0.0709	0.6513 ± 0.0094	0.8797 ± 0.0089	0.5886 ± 0.0635
Dual (ID-only)	0.8917 ± 0.0101	0.5837 ± 0.0732	0.6587 ± 0.0115	0.8618 ± 0.0107	0.6301 ± 0.0666
Shuffled Dual	0.6690 ± 0.0049	0.7486 ± 0.0062	0.7512 ± 0.0043	0.6572 ± 0.0070	0.7754 ± 0.0090
D-CUQ (Dual)	0.9004 ± 0.0082	0.5363 ± 0.0696	0.6528 ± 0.0093	0.8743 ± 0.0089	0.5914 ± 0.0607

Table 14: Dataset-wise architecture-macro OOD AUROC. Only aggregate means were retained for this breakdown, so no uncertainty intervals are reported. Gaussian corruption is evaluated separately as a covariate-shift diagnostic.

Dataset	Softmax	Global-M	Global-ID	M-only	G-only	Dual-ID	Shuffled	**D-CUQ**
CIFAR-10 (near-OOD)	0.8046	0.8144	0.8218	0.8121	0.8249	0.8231	0.6772	0.8223
Tiny-ImageNet (near-OOD)	0.8406	0.8541	0.8632	0.8937	0.9212	0.8723	0.7057	0.9192
MNIST (far-OOD)	0.6793	0.7053	0.7094	0.7493	0.7534	0.7161	0.6214	0.7623
SVHN (far-OOD)	0.8698	0.8862	0.9093	0.8837	0.8976	0.9101	0.7282	0.8956
Places365 (far-OOD)	0.8499	0.8646	0.8709	0.9029	0.9300	0.8805	0.7201	0.9306
Gaussian shift	0.9224	0.9286	0.9351	0.9691	0.9676	0.9434	0.9174	0.9742

Table 15: Classification OOD and failure-detection results with CIFAR-10 ID. Values are mean ± standard deviation over all backbone–seed evaluations.

Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
Softmax	0.9438 ± 0.0126	0.2424 ± 0.0497	0.5928 ± 0.0127	0.9282 ± 0.0180	0.2737 ± 0.0555
Temperature Scaling	0.9542 ± 0.0098	0.2312 ± 0.0518	0.5892 ± 0.0123	0.9437 ± 0.0124	0.2624 ± 0.0571
MC Dropout	0.9477 ± 0.0141	0.2312 ± 0.0463	0.5891 ± 0.0133	0.9334 ± 0.0202	0.2628 ± 0.0525
Deep Ensemble	0.9502 ± 0.0142	0.2067 ± 0.0450	0.5847 ± 0.0127	0.9359 ± 0.0208	0.2380 ± 0.0524
SWAG	0.9432 ± 0.0130	0.2434 ± 0.0501	0.5932 ± 0.0128	0.9275 ± 0.0182	0.2749 ± 0.0558
VBLL	0.9526 ± 0.0140	0.2362 ± 0.0647	0.5899 ± 0.0148	0.9416 ± 0.0182	0.2691 ± 0.0719
SNGP	0.9643 ± 0.0189	0.1941 ± 0.1054	0.5766 ± 0.0184	0.9590 ± 0.0178	0.2141 ± 0.0954
EDL	0.9499 ± 0.0119	0.2435 ± 0.0598	0.5896 ± 0.0109	0.9377 ± 0.0162	0.2694 ± 0.0574
D-CUQ (Drop)	0.9524 ± 0.0145	0.2077 ± 0.0464	0.5861 ± 0.0120	0.9388 ± 0.0207	0.2396 ± 0.0541
D-CUQ (learned)	0.9764 ± 0.0084	0.1246 ± 0.0422	0.5692 ± 0.0085	0.9727 ± 0.0105	0.1439 ± 0.0502

Table 16: Per-backbone Failure AUROC with CIFAR-10 ID. Backbone entries are mean ± standard deviation over 5 seeds.

Method	DenseNet-121	ResNet-50	VGG-16	ViT-Tiny	Macro
Softmax	0.9403 ± 0.0026	0.9431 ± 0.0010	0.9291 ± 0.0029	0.9628 ± 0.0019	0.9438 ± 0.0126
Temperature Scaling	0.9471 ± 0.0028	0.9496 ± 0.0007	0.9499 ± 0.0017	0.9704 ± 0.0016	0.9542 ± 0.0098
MC Dropout	0.9418 ± 0.0028	0.9577 ± 0.0032	0.9285 ± 0.0028	0.9629 ± 0.0019	0.9477 ± 0.0141
Deep Ensemble	0.9549 ± 0.0015	0.9514 ± 0.0018	0.9284 ± 0.0042	0.9661 ± 0.0014	0.9502 ± 0.0142
SWAG	0.9398 ± 0.0024	0.9418 ± 0.0020	0.9284 ± 0.0030	0.9629 ± 0.0018	0.9432 ± 0.0130
VBLL	0.9539 ± 0.0050	0.9446 ± 0.0019	0.9383 ± 0.0028	0.9736 ± 0.0032	0.9526 ± 0.0140
SNGP	0.9789 ± 0.0021	0.9670 ± 0.0016	0.9337 ± 0.0029	0.9778 ± 0.0014	0.9643 ± 0.0189
EDL	0.9525 ± 0.0059	0.9647 ± 0.0022	0.9338 ± 0.0027	0.9484 ± 0.0033	0.9499 ± 0.0119
D-CUQ (Drop)	0.9574 ± 0.0023	0.9482 ± 0.0009	0.9328 ± 0.0026	0.9713 ± 0.0015	0.9524 ± 0.0145
D-CUQ (learned)	0.9823 ± 0.0006	0.9811 ± 0.0010	0.9623 ± 0.0017	0.9797 ± 0.0005	0.9764 ± 0.0084

Table 17: Classification OOD and failure-detection results with CIFAR-100 ID. Values are mean ± standard deviation over all backbone–seed evaluations.

Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
Softmax	0.8535 ± 0.0330	0.5871 ± 0.0511	0.6955 ± 0.0169	0.7988 ± 0.0496	0.6633 ± 0.0447
Temperature Scaling	0.8709 ± 0.0217	0.5641 ± 0.0522	0.6875 ± 0.0156	0.8243 ± 0.0330	0.6421 ± 0.0488
MC Dropout	0.8492 ± 0.0323	0.6038 ± 0.0549	0.6979 ± 0.0202	0.7963 ± 0.0492	0.6770 ± 0.0480
Deep Ensemble	0.8662 ± 0.0333	0.5625 ± 0.0531	0.6823 ± 0.0191	0.8190 ± 0.0512	0.6363 ± 0.0478
SWAG	0.8532 ± 0.0318	0.5853 ± 0.0516	0.6958 ± 0.0164	0.7980 ± 0.0481	0.6623 ± 0.0448
VBLL	0.8677 ± 0.0295	0.5515 ± 0.0720	0.6931 ± 0.0269	0.8192 ± 0.0451	0.6353 ± 0.0740
SNGP	0.8225 ± 0.0457	0.6427 ± 0.0552	0.7733 ± 0.1007	0.7523 ± 0.1235	0.7285 ± 0.1166
EDL	0.8716 ± 0.0175	0.5335 ± 0.0722	0.6820 ± 0.0241	0.8055 ± 0.0427	0.6177 ± 0.0747
D-CUQ (Drop)	0.8659 ± 0.0281	0.5688 ± 0.0556	0.6899 ± 0.0153	0.8151 ± 0.0434	0.6485 ± 0.0473
D-CUQ (fixed)	0.8905 ± 0.0282	0.5144 ± 0.0723	0.6754 ± 0.0171	0.8556 ± 0.0433	0.5956 ± 0.0683
D-CUQ (NLL-rank)	0.8839 ± 0.0163	0.5379 ± 0.0523	0.6802 ± 0.0137	0.8429 ± 0.0248	0.6195 ± 0.0499
D-CUQ (learned)	0.8900 ± 0.0254	0.5059 ± 0.0862	0.6737 ± 0.0189	0.8615 ± 0.0319	0.5829 ± 0.0866

Table 18: Per-backbone Failure AUROC with CIFAR-100 ID. Backbone entries are mean ± standard deviation over 5 seeds.

Method	DenseNet-121	ResNet-18	ResNet-50	VGG-16	ViT-Tiny	Macro
Softmax	0.8769 ± 0.0119	0.8704 ± 0.0094	0.8784 ± 0.0076	0.7949 ± 0.0057	0.8470 ± 0.0063	0.8535 ± 0.0330
Temperature Scaling	0.8882 ± 0.0118	0.8694 ± 0.0093	0.8912 ± 0.0074	0.8354 ± 0.0054	0.8705 ± 0.0059	0.8709 ± 0.0217
MC Dropout	0.8775 ± 0.0102	0.8368 ± 0.0102	0.8831 ± 0.0119	0.7986 ± 0.0057	0.8503 ± 0.0068	0.8492 ± 0.0323
Deep Ensemble	0.8966 ± 0.0029	0.8808 ± 0.0035	0.8889 ± 0.0039	0.8066 ± 0.0050	0.8581 ± 0.0046	0.8662 ± 0.0333
SWAG	0.8763 ± 0.0101	0.8697 ± 0.0103	0.8761 ± 0.0097	0.7969 ± 0.0053	0.8473 ± 0.0071	0.8532 ± 0.0318
VBLL	0.8808 ± 0.0193	0.8709 ± 0.0211	0.8852 ± 0.0131	0.8188 ± 0.0076	0.8827 ± 0.0166	0.8677 ± 0.0295
SNGP	0.8670 ± 0.0103	0.8166 ± 0.0122	0.8599 ± 0.0075	0.7444 ± 0.0181	0.8247 ± 0.0058	0.8225 ± 0.0457
EDL	0.8808 ± 0.0043	0.8781 ± 0.0050	0.8920 ± 0.0095	0.8532 ± 0.0110	0.8537 ± 0.0089	0.8716 ± 0.0175
D-CUQ (Drop)	0.8905 ± 0.0119	0.8793 ± 0.0087	0.8820 ± 0.0077	0.8166 ± 0.0054	0.8609 ± 0.0074	0.8659 ± 0.0281
D-CUQ (fixed)	0.9182 ± 0.0088	0.8982 ± 0.0074	0.9125 ± 0.0070	0.8451 ± 0.0082	0.8784 ± 0.0112	0.8905 ± 0.0282
D-CUQ (NLL-rank)	0.8997 ± 0.0106	0.8746 ± 0.0087	0.8983 ± 0.0073	0.8637 ± 0.0085	0.8832 ± 0.0100	0.8839 ± 0.0163
D-CUQ (learned)	0.9242 ± 0.0116	0.8725 ± 0.0098	0.9066 ± 0.0056	0.8592 ± 0.0058	0.8878 ± 0.0136	0.8900 ± 0.0254

Table 19: Matched comparisons with recent OOD baselines on CIFAR-100. Bold values are best only within each protocol-matched block.

Scope	Method	F-AUROC↑	F-FPR95↓	AURC↓	OOD-AUROC↑	OOD-FPR95↓
Four shared CNNs	CSD	0.8446 ± 0.0463	0.5923 ± 0.0944	0.7030 ± 0.0306	0.8306 ± 0.0444	0.6073 ± 0.0864
	D-CUQ	0.8906 ± 0.0278	0.4967 ± 0.0875	0.6757 ± 0.0200	0.8621 ± 0.0350	0.5778 ± 0.0909
ResNet-18	Split-Ensemble (reprod.)	0.8934 ± 0.0070	0.3957 ± 0.0322	0.6759 ± 0.0128	0.8250 ± 0.0096	0.5177 ± 0.0330
	D-CUQ	0.8982 ± 0.0074	0.5070 ± 0.0256	0.6779 ± 0.0061	0.8691 ± 0.0087	0.5978 ± 0.0274

Table 20: Regression OOD AUROC (↑) for each dataset and baseline. Values are mean ± standard deviation over 5 seeds; Macro first averages the 5 datasets within each seed.

Baseline	Variant	Concrete	Energy	Kin8nm	Airfoil	Wine Red	Macro
Det. Gaussian	Base	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000
	M-only	0.7757 ± 0.0212	0.8005 ± 0.0582	0.6672 ± 0.0124	0.7365 ± 0.0245	0.7797 ± 0.0250	0.7519 ± 0.0193
	G-only	0.8242 ± 0.0328	0.7524 ± 0.1291	0.6894 ± 0.0124	0.7754 ± 0.0430	0.7868 ± 0.0135	0.7657 ± 0.0330
	D-CUQ (Dual)	0.8529 ± 0.0194	0.8409 ± 0.0723	0.7630 ± 0.0180	0.8393 ± 0.0184	0.8449 ± 0.0212	0.8282 ± 0.0214
MC Drop.	Base	0.6743 ± 0.0237	0.6402 ± 0.0293	0.5611 ± 0.0162	0.5744 ± 0.0195	0.6185 ± 0.0219	0.6137 ± 0.0070
	M-only	0.7759 ± 0.0185	0.7920 ± 0.0415	0.6564 ± 0.0134	0.7285 ± 0.0294	0.7814 ± 0.0233	0.7468 ± 0.0152
	G-only	0.8306 ± 0.0222	0.7802 ± 0.0788	0.6848 ± 0.0161	0.7610 ± 0.0368	0.7901 ± 0.0145	0.7694 ± 0.0247
	D-CUQ (Dual)	0.8524 ± 0.0172	0.8367 ± 0.0530	0.7421 ± 0.0086	0.8169 ± 0.0184	0.8431 ± 0.0229	0.8183 ± 0.0189
Deep Ens.	Base	0.8890 ± 0.0173	0.8781 ± 0.0136	0.7825 ± 0.0159	0.8924 ± 0.0275	0.8107 ± 0.0181	0.8505 ± 0.0064
	M-only	0.8608 ± 0.0130	0.8868 ± 0.0177	0.7758 ± 0.0136	0.8631 ± 0.0270	0.8400 ± 0.0190	0.8453 ± 0.0090
	G-only	0.8928 ± 0.0232	0.8885 ± 0.0301	0.7886 ± 0.0117	0.8908 ± 0.0073	0.8443 ± 0.0081	0.8610 ± 0.0112
	D-CUQ (Dual)	0.8887 ± 0.0197	0.9054 ± 0.0246	0.8138 ± 0.0162	0.8937 ± 0.0098	0.8639 ± 0.0171	0.8731 ± 0.0127
SWAG	Base	0.6880 ± 0.0326	0.7614 ± 0.0380	0.5966 ± 0.0331	0.6522 ± 0.0219	0.7007 ± 0.0279	0.6798 ± 0.0128
	M-only	0.7897 ± 0.0202	0.8180 ± 0.0573	0.6845 ± 0.0101	0.7668 ± 0.0279	0.7915 ± 0.0205	0.7701 ± 0.0170
	G-only	0.8285 ± 0.0279	0.7925 ± 0.0926	0.7030 ± 0.0086	0.7866 ± 0.0358	0.8050 ± 0.0183	0.7831 ± 0.0228
	D-CUQ (Dual)	0.8541 ± 0.0192	0.8471 ± 0.0672	0.7640 ± 0.0180	0.8417 ± 0.0178	0.8459 ± 0.0208	0.8306 ± 0.0205

Table 21: Regression Failure AUROC (↑) for each dataset and baseline. Values are mean ± standard deviation over 5 seeds; Macro first averages the 5 datasets within each seed.

Baseline	Variant	Concrete	Energy	Kin8nm	Airfoil	Wine Red	Macro
Det. Gaussian	Base	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000
	M-only	0.6739 ± 0.0487	0.7412 ± 0.0974	0.6250 ± 0.0178	0.7510 ± 0.0377	0.6901 ± 0.0109	0.6962 ± 0.0157
	G-only	0.7148 ± 0.0321	0.7227 ± 0.1307	0.4994 ± 0.0167	0.6873 ± 0.0409	0.7202 ± 0.0375	0.6689 ± 0.0397
	D-CUQ (Dual)	0.7361 ± 0.0313	0.7552 ± 0.1167	0.5995 ± 0.0190	0.8232 ± 0.0125	0.7482 ± 0.0189	0.7324 ± 0.0247
MC Drop.	Base	0.5763 ± 0.0483	0.5915 ± 0.0613	0.5503 ± 0.0231	0.5349 ± 0.0253	0.5408 ± 0.0263	0.5588 ± 0.0129
	M-only	0.6246 ± 0.0294	0.7473 ± 0.0801	0.6774 ± 0.0128	0.7517 ± 0.0296	0.5823 ± 0.0318	0.6767 ± 0.0143
	G-only	0.6792 ± 0.0773	0.7205 ± 0.1060	0.4853 ± 0.0087	0.6360 ± 0.0417	0.6015 ± 0.0364	0.6245 ± 0.0402
	D-CUQ (Dual)	0.6823 ± 0.0558	0.7610 ± 0.1092	0.6302 ± 0.0115	0.7924 ± 0.0254	0.6096 ± 0.0458	0.6951 ± 0.0266
Deep Ens.	Base	0.7376 ± 0.0288	0.8891 ± 0.0366	0.6386 ± 0.0170	0.7782 ± 0.0258	0.6488 ± 0.0259	0.7385 ± 0.0131
	M-only	0.7256 ± 0.0269	0.8512 ± 0.0573	0.6746 ± 0.0113	0.8041 ± 0.0275	0.6663 ± 0.0181	0.7444 ± 0.0100
	G-only	0.7433 ± 0.0456	0.8584 ± 0.0519	0.5642 ± 0.0156	0.7756 ± 0.0137	0.6648 ± 0.0313	0.7213 ± 0.0200
	D-CUQ (Dual)	0.7515 ± 0.0258	0.8473 ± 0.0660	0.6353 ± 0.0206	0.8241 ± 0.0161	0.6760 ± 0.0230	0.7468 ± 0.0193
SWAG	Base	0.6676 ± 0.0497	0.8083 ± 0.0221	0.5707 ± 0.0343	0.6087 ± 0.0329	0.6534 ± 0.0725	0.6618 ± 0.0178
	M-only	0.6811 ± 0.0460	0.7565 ± 0.0861	0.6253 ± 0.0183	0.7637 ± 0.0298	0.6887 ± 0.0127	0.7031 ± 0.0126
	G-only	0.7114 ± 0.0359	0.7551 ± 0.0932	0.5096 ± 0.0110	0.6956 ± 0.0398	0.7269 ± 0.0391	0.6797 ± 0.0319
	D-CUQ (Dual)	0.7342 ± 0.0271	0.7636 ± 0.1075	0.5978 ± 0.0186	0.8245 ± 0.0140	0.7499 ± 0.0266	0.7340 ± 0.0236

Table 22: Five-dataset macro regression UQ results. Bold values are best within each base-UQ block.

Baseline	Variant	NLL↓	CRPS↓	Cal. error↓	Error ρ ↑	AURC↓
Deterministic Gaussian	Base	6.6637 ± 2.9686	0.4403 ± 0.0248	0.1258 ± 0.0191	0.0000 ± 0.0000	0.5781 ± 0.0298
	M-only	2.9406 ± 1.1842	0.4225 ± 0.0217	0.0810 ± 0.0152	0.3200 ± 0.0235	0.4186 ± 0.0146
	G-only	3.2964 ± 1.7911	0.4248 ± 0.0237	0.0803 ± 0.0120	0.2747 ± 0.0640	0.4332 ± 0.0154
	D-CUQ (Dual)	2.6158 ± 1.2107	0.4186 ± 0.0214	0.0765 ± 0.0108	0.3714 ± 0.0447	0.3840 ± 0.0138
MC Dropout	Base	1.9020 ± 0.4317	0.3659 ± 0.0122	0.0990 ± 0.0097	0.0971 ± 0.0263	0.4529 ± 0.0113
	M-only	1.1100 ± 0.1605	0.3550 ± 0.0119	0.0588 ± 0.0117	0.2770 ± 0.0252	0.3928 ± 0.0090
	G-only	1.2532 ± 0.2596	0.3610 ± 0.0129	0.0675 ± 0.0163	0.2097 ± 0.0545	0.4181 ± 0.0237
	D-CUQ (Dual)	1.0339 ± 0.1572	0.3549 ± 0.0119	0.0644 ± 0.0140	0.3109 ± 0.0311	0.3788 ± 0.0101
Deep Ensemble	Base	3.1081 ± 0.6719	0.3810 ± 0.0176	0.1000 ± 0.0079	0.4000 ± 0.0255	0.3391 ± 0.0145
	M-only	1.9616 ± 0.4381	0.3771 ± 0.0199	0.0747 ± 0.0083	0.3962 ± 0.0137	0.3376 ± 0.0126
	G-only	2.0735 ± 0.7551	0.3867 ± 0.0203	0.0852 ± 0.0046	0.3742 ± 0.0390	0.3431 ± 0.0137
	D-CUQ (Dual)	1.7765 ± 0.5268	0.3826 ± 0.0200	0.0821 ± 0.0098	0.4058 ± 0.0293	0.3350 ± 0.0120
SWAG	Base	7.0083 ± 3.0929	0.4442 ± 0.0245	0.1265 ± 0.0190	0.2553 ± 0.0397	0.4567 ± 0.0252
	M-only	3.0871 ± 1.2399	0.4264 ± 0.0216	0.0820 ± 0.0153	0.3275 ± 0.0244	0.4087 ± 0.0142
	G-only	3.4616 ± 1.8721	0.4285 ± 0.0236	0.0810 ± 0.0121	0.2920 ± 0.0542	0.4194 ± 0.0115
	D-CUQ (Dual)	2.7491 ± 1.2684	0.4223 ± 0.0214	0.0773 ± 0.0110	0.3742 ± 0.0438	0.3855 ± 0.0130

Table 23: Five-dataset macro interval and point-prediction results. Coverage values closest to their nominal levels are bold; lower is better for width, RMSE, and MAE. RMSE and MAE tie because D-CUQ preserves the predictive mean.

Baseline	Variant	Cov.90	Cov.95	Width95↓	RMSE↓	MAE↓
Deterministic Gaussian	Base	0.6733 ± 0.0189	0.7312 ± 0.0142	1.5855 ± 0.1214	0.7816 ± 0.0477	0.5581 ± 0.0307
	M-only	0.7534 ± 0.0233	0.8096 ± 0.0209	2.1313 ± 0.2845	0.7816 ± 0.0477	0.5581 ± 0.0307
	G-only	0.7631 ± 0.0193	0.8099 ± 0.0229	2.2396 ± 0.2269	0.7816 ± 0.0477	0.5581 ± 0.0307
	D-CUQ (Dual)	0.7789 ± 0.0206	0.8278 ± 0.0197	2.3219 ± 0.2593	0.7816 ± 0.0477	0.5581 ± 0.0307
MC Dropout	Base	0.7249 ± 0.0149	0.7763 ± 0.0114	1.6348 ± 0.0516	0.6469 ± 0.0229	0.4833 ± 0.0157
	M-only	0.8208 ± 0.0200	0.8730 ± 0.0190	2.2176 ± 0.1753	0.6469 ± 0.0229	0.4833 ± 0.0157
	G-only	0.8139 ± 0.0279	0.8617 ± 0.0242	2.3047 ± 0.1220	0.6469 ± 0.0229	0.4833 ± 0.0157
	D-CUQ (Dual)	0.8448 ± 0.0224	0.8914 ± 0.0177	2.3964 ± 0.1380	0.6469 ± 0.0229	0.4833 ± 0.0157
Deep Ensemble	Base	0.7290 ± 0.0070	0.7826 ± 0.0073	1.7681 ± 0.0730	0.6954 ± 0.0312	0.4986 ± 0.0223
	M-only	0.8228 ± 0.0083	0.8675 ± 0.0097	2.5103 ± 0.2489	0.6954 ± 0.0312	0.4986 ± 0.0223
	G-only	0.8154 ± 0.0206	0.8597 ± 0.0163	2.6584 ± 0.2031	0.6954 ± 0.0312	0.4986 ± 0.0223
	D-CUQ (Dual)	0.8390 ± 0.0117	0.8795 ± 0.0125	2.7869 ± 0.2317	0.6954 ± 0.0312	0.4986 ± 0.0223
SWAG	Base	0.6723 ± 0.0193	0.7307 ± 0.0145	1.5966 ± 0.1241	0.7890 ± 0.0472	0.5629 ± 0.0305
	M-only	0.7533 ± 0.0240	0.8078 ± 0.0206	2.1463 ± 0.2906	0.7890 ± 0.0472	0.5629 ± 0.0305
	G-only	0.7615 ± 0.0197	0.8091 ± 0.0229	2.2559 ± 0.2306	0.7890 ± 0.0472	0.5629 ± 0.0305
	D-CUQ (Dual)	0.7793 ± 0.0211	0.8271 ± 0.0193	2.3386 ± 0.2645	0.7890 ± 0.0472	0.5629 ± 0.0305