UNCERTAINTY ESTIMATION VIA BAYESIAN DECISION LAYER WITH WASSERSTEIN GEOMETRIC MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty estimation should distinguish predictive confidence from the degree of compatibility between an input and the data distributions associated with different classes. Existing approaches typically characterize uncertainty through predictive probabilities, posterior variability, or distances to class structures, but do not jointly model these aspects. Probabilistic methods do not necessarily quantify compatibility between input- and class-level distributions, whereas distance-aware methods commonly operate on point representations. To address these challenges, we introduce the **W**asserstein **G**eneralized **B**ayesian **D**ecision **L**ayer (W-GBDL), which derives uncertainty from the compatibility between an input-conditioned distribution and learnable class reference distributions. W-GBDL represents each input and each class with variational distributions and measures their compatibility using the Wasserstein distance. This geometry provides a unified basis for training and minimum-risk prediction, while the entropy of the resulting class probabilities captures predictive ambiguity. Experiments across image and language tasks show improved uncertainty estimation and calibration while maintaining competitive classification performance.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable success in computer vision, natural language processing, and engineering applications (Deng et al., 2009; LeCun et al., 2015). However, the deterministic formulation can produce incorrect yet highly confident predictions(Guo et al., 2017; Kristiadi et al., 2020), which is particularly problematic in high-stakes applications (Bengio et al., 2013; Reichstein et al., 2019). Reliable uncertainty estimation is therefore essential for rejecting uncertain predictions or deferring them to human experts (Gal & Ghahramani, 2016). Existing approaches to uncertainty estimation broadly fall into three categories. Evidential deep learning places a Dirichlet distribution over class probabilities (Sensoy et al., 2018; Deng et al., 2023; Chen et al., 2025), but the statistical interpretation of evidence and uncertainty and the stability of its calibration remain under debate(Juergens et al., 2024). Deterministic methods, such as deep ensembles (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Gal & Ghahramani, 2016), typically require multiple models or repeated forward passes and therefore incur substantial computational cost(Fan et al., 2021). Bayesian neural networks (BNNs) instead place distributions over model parameters and use the induced predictive distribution to quantify confidence(Blundell et al., 2015; Hernández-Lobato & Adams, 2015; Tomczak et al., 2021). However, full BNNs are difficult to train and scale. Although a variety of approximate inference methods have been developed, inference over high-dimensional parameter posteriors remains computationally prohibitive for large models.

Bayesian last-layer (BLL) models provide a more scalable compromise. They retain a deterministic feature extractor and apply Bayesian inference only to the parameters of the final classification layer (Watson et al., 2021; Harrison et al., 2024; Hu et al., 2025). This substantially reduces the dimensionality of posterior inference and the overall computational burden while preserving some of the benefits of parameter uncertainty. However, BLL models generally express uncertainty through the posterior predictive distribution induced by the final-layer parameters. Because this predictive distribution is normalized over the known classes, it primarily describes their relative probabilities and does not explicitly compare an input distribution with the learned class structure. Consequently, posterior predictive confidence alone may be insufficient for distinguishing a difficult in-distribution sample from an unfamiliar input that lies beyond the learned class structure, especially when the

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.

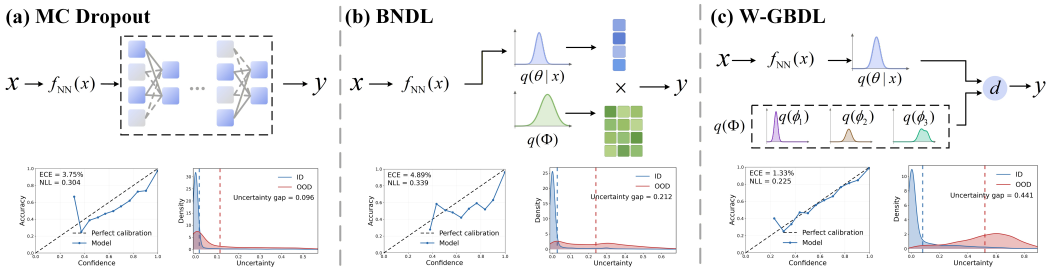


Figure 1: Motivation for distribution-aware class matching. Top: MC Dropout, BNDL as a representative BLL model and our methods. Bottom: the corresponding calibration and ID–OOD uncertainty behavior. W-GBDL extends Bayesian decision-layer modeling by comparing an input-conditioned distribution with class reference distributions through Wasserstein geometry.

deterministic representation extrapolates under distribution shift (Schwöbel et al., 2022; Hendrycks & Gimpel, 2016; Hein et al., 2019; Liu et al., 2020b).

This limitation points to a more fundamental distinction: uncertainty should reflect not only predictive confidence, but also the compatibility between an input and the learned class structure(Ovadia et al., 2019). In particular, reliable uncertainty estimation should account for both how strongly a classifier favors one class over the others and how well the input-conditioned distribution matches the corresponding class-reference distribution(Liu et al., 2020a). Based on distance and density classifiers address this issue by matching an input representation against learned class centers or class-conditional models (Lee et al., 2018; Van Amersfoort et al., 2020; Mukhoti et al., 2023). Their confidence therefore depends on the degree of sample-to-class matching. Most such approaches, however, still treat the input as a point representation, even when the class-conditional structure is modeled probabilistically, and therefore do not compare uncertainty-bearing distributions on both sides of the match. Fig. 1 illustrates the conceptual progression from conventional predictive uncertainty to distribution-aware class matching. MC Dropout represents the common strategy of estimating uncertainty from stochastic forward passes. BNDL serves as a representative BLL model, in which uncertainty is modeled through local and global latent factors while the final prediction is formed by a decision rule over the learned factors. W-GBDL takes a further step by representing both the input and each class reference as distributions and comparing them through Wasserstein geometry. The bottom row provides a qualitative motivation for this design. Compared with the two baselines, W-GBDL yields a calibration curve closer to the diagonal and a clearer separation between the ID and OOD uncertainty-score distributions in the shown example.

Based on these considerations, we propose the **W**asserstein **G**eneralized **B**ayesian **D**ecision **L**ayer (W-GBDL), a scalable framework for distribution-aware classification and uncertainty estimation. W-GBDL retains one-hot supervision for identifying the correct class, while replacing conventional point-wise class comparison with distribution-to-distribution matching. Specifically, it learns an input-conditioned stochastic distribution and a global reference distribution for each class, preserving distributional information on both sides of the classification process. The sample distribution captures input-conditioned latent variability, while each class reference encodes the learned location, shape, and uncertainty-relevant variability of the corresponding class. W-GBDL measures the compatibility between the sample distribution and each class reference using the second-order Wasserstein distance induced by optimal transport, and predicts the class with the smallest distance. The resulting Wasserstein distances are converted into class compatibilities, which in turn define a predictive distribution whose entropy provides an uncertainty measure. This formulation jointly captures distributional variability from both the sample and class-reference sides within a unified generalized Bayesian decision framework. Our main contributions are:

- We introduce W-GBDL, a distribution-aware bayesian decision layer that represents input-conditioned sample factors and learnable class references as probability distributions, replacing point-wise comparison with Wasserstein-based distributional matching.
- We propose a tractable generalized variational objective in which Weibull local and global posteriors yield an analytic Wasserstein discrepancy, enabling differentiable learning of sample- and class-side uncertainty without Monte Carlo transport estimation.

- We evaluate W-GBDL across multiple image and language datasets and model architectures. Experiments demonstrate competitive performance in uncertainty estimation and calibration. Further analyses examine the sources of aleatoric and epistemic uncertainty.

2 Related Work

2.1 Uncertainty Estimation via Probabilistic Modeling

Beyond point predictions, predictive uncertainty estimation has been widely studied as a means of improving the reliability of deep neural networks. Deep ensembles estimate epistemic uncertainty by aggregating predictions from independently trained models (Lakshminarayanan et al., 2017; Liu et al., 2021). Evidential deep learning instead models higher-order predictive distributions, such as Dirichlet distributions for classification, to characterize uncertainty without explicit model sampling (Sensoy et al., 2018). Related prior-network formulations extend this principle to regression settings (Malinin et al., 2020). Bayesian neural networks place prior distributions over model parameters, although full posterior inference is often computationally demanding (Blundell et al., 2015; Hernández-Lobato & Adams, 2015). Bayesian last-layer methods offer a more tractable alternative by restricting probabilistic inference to the classifier head (Weber et al., 2018; Harrison et al., 2024; Watson et al., 2021; Hu et al., 2025). BLL methods have been extended through noise marginalization (Ober & Rasmussen, 2019), Laplace approximations for probabilistic prediction (Daxberger et al., 2021), specialized retraining objectives (Weber et al., 2018), and improved variational inference (Harrison et al., 2024). Building on this line, Hu et al. (2025) cast the decision layer as a Bayesian non-negative factor model with stochastic local and global latent variables, jointly enhancing uncertainty estimation and interpretability.

2.2 Optimal Transport and Wasserstein Distance

Optimal transport (OT) compares probability distributions by finding a coupling that minimizes the expected transport cost (Villani, 2009; Peyré & Cuturi, 2019). For distributions P and Q with finite p-th moments, the p-Wasserstein distance is defined as

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int \|\boldsymbol{u} - \boldsymbol{v}\|^p \, d\pi(\boldsymbol{u}, \boldsymbol{v}) \right)^{1/p}, \tag{1}$$

where $\Pi(P,Q)$ denotes the set of couplings with marginals P and Q. Unlike pointwise divergences, Wasserstein distance accounts for the geometry of the underlying space and remains well defined for distributions with non-overlapping supports. It has been widely used in classification, domain adaptation, and generative modeling (Frogner et al., 2015; Courty et al., 2017; Arjovsky et al., 2017). W-GBDL adopts Wasserstein geometry at the decision layer, using the squared 2-Wasserstein distance to compare an input-conditioned sample posterior with each learned class reference posterior and the corresponding optimal coupling to construct the predictive uncertainty.

3 Method

3.1 Wasserstein-Based Hierarchical Generative Model

W-GBDL is formulated as a hierarchical generalized generative model. Given an observed input x_i, we introduce sample-specific non-negative latent variables $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^N \in \mathbb{R}_+^{N \times L}$ and global class-reference variables $\Phi = \{\boldsymbol{\phi}_c\}_{c=1}^C \in \mathbb{R}_+^{C \times L}$. A deterministic backbone first maps x_i to a feature representation, from which the input-conditioned latent variable $\boldsymbol{\theta}_i$ is inferred, while the class references Φ are globally shared across samples. The class-probability vector is introduced as an intermediate random variable, yielding a Dirichlet–Categorical hierarchy for label generation.

We place independent Gamma priors on the non-negative latent variables(Zhou et al., 2015; Duan et al., 2024):

$$\boldsymbol{\theta}_i \sim \text{Gamma}\big(a_\theta, b_\theta\big), \qquad \boldsymbol{\phi}_c \sim \text{Gamma}\big(a_\phi, b_\phi\big), \tag{2}$$

where a_θ and a_ϕ are shape parameters, and b_θ and b_ϕ are rate parameters. When $a_\theta < 1$ and $a_\phi < 1$, these priors place substantial mass near zero, encouraging shrinkage of unnecessary latent factors and promoting sparse representations.

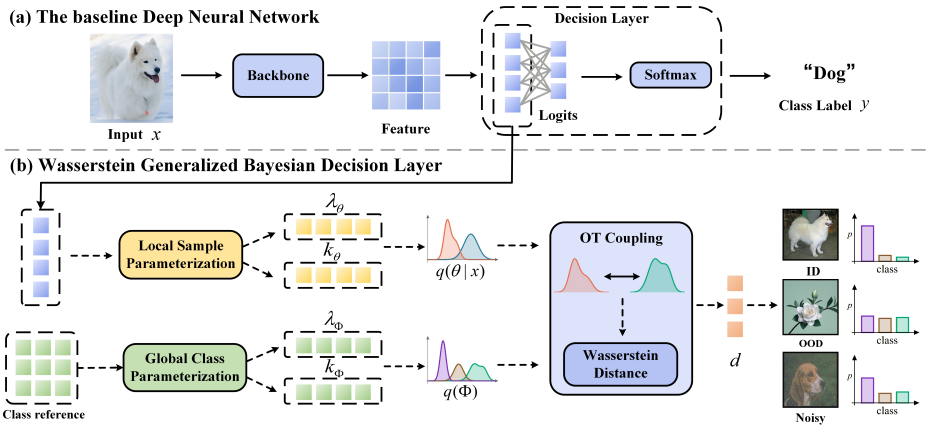


Figure 2: Framework of our method. (a)The predictive process of output label y for the baseline Deep Neural Network. (b) W-GBDL generation via non-negative latent factors, with the sample-level ($\boldsymbol{\theta}$) and global-level (Φ) distributions compared using the Wasserstein distance.

After observing $\mathcal{D}$, the ideal generalized posterior distributions are denoted by $p(\boldsymbol{\theta}_i \mid x_i)$ and $p(\Phi)$. Let $p(\boldsymbol{\phi}_c)$ denote the marginal posterior of the c-th class reference. We lift the squared distance loss $\ell_c(\boldsymbol{\theta}_i, \boldsymbol{\phi}_c) = \|\boldsymbol{\theta}_i - \boldsymbol{\phi}_c\|_2^2$ to a distribution-level risk through optimal transport. We define the transport discrepancy between the sample and class posteriors as

$$d_{ic} = \inf_{\pi \in \Pi_{ic}} \mathbb{E}_\pi \left[\ell_c(\boldsymbol{\theta}_i, \boldsymbol{\phi}_c)\right] = W_2^2(p(\boldsymbol{\theta}_i \mid x_i), p(\boldsymbol{\phi}_c)), \tag{3}$$

where Π_{ic} is the set of couplings with marginals $p(\boldsymbol{\theta}_i \mid x_i)$ and $p(\boldsymbol{\phi}_c)$. Collecting these discrepancies as $\boldsymbol{d}_i = [d_{i1}, \ldots, d_{iC}]^\top$, we convert the Wasserstein risks into Dirichlet concentration parameters: $\alpha_{ic} = \exp(-d_{ic}/\tau)$. Here, τ controls the scale of the distance-induced parameters.

The class-probability vector $\boldsymbol{\rho}_i = [\rho_{i1}, \ldots, \rho_{iC}]^\top$ belongs to the probability simplex Δ^{C-1} and follows a Dirichlet distribution:

$$\mathrm{Dir}(\boldsymbol{\rho}_i \mid \boldsymbol{\alpha}_i) = \frac{\Gamma(\alpha_{i0})}{\prod_{c'=1}^{C} \Gamma(\alpha_{ic'})} \prod_{c'=1}^{C} \rho_{ic'}^{\alpha_{ic'}-1}, \qquad \alpha_{i0} = \sum_{c'=1}^{C} \alpha_{ic'}. \tag{4}$$

Conditioned on this random probability vector, the observed label is generated according to

$$y_i \mid \boldsymbol{\rho}_i \sim \mathrm{Categorical}(\boldsymbol{\rho}_i), \tag{5}$$

The expected predictive class probability coincides with the Dirichlet mean:

$$p(y_i = c \mid \alpha_i) = \bar{\rho}_i = \int_{\Delta^{C-1}} p(y_i = c \mid \rho_i) \mathrm{Dir}(\rho_i \mid \alpha_i)\, d\rho_i = \frac{\alpha_{ic}}{\alpha_{i0}}. \tag{6}$$

The hierarchy captures two complementary sources of latent uncertainty. The input-conditioned posterior $p(\boldsymbol{\theta}_i \mid x_i)$ reflects input-dependent, aleatoric-related uncertainty, while the class-reference posterior $p(\boldsymbol{\phi}_c)$ provides an epistemic-related source. These two sources are jointly propagated through the Wasserstein discrepancy d_{ic} to determine class compatibility and, through the Dirichlet–Categorical layer, the predictive distribution and its entropy. Thus, the W-GBDL generalized bayes decision rule is $\widehat{y}_i = \arg\max_c \bar{\rho}_{ic} = \arg\min_c d_{ic}$.

3.2 Variational Inference Network

The generalized posterior targets induced by the likelihood and the Gamma priors are generally intractable. We adopt the mean-field approximation $q(\boldsymbol{\theta}, \Phi) = q(\Phi) \prod_{i=1}^{N} q(\boldsymbol{\theta}_i \mid x_i)$. Here, $q(\boldsymbol{\phi}_c)$ is a class-specific global variational factor shared across all samples, whereas $q(\boldsymbol{\theta} i \mid x_i)$ is an instance-specific factor conditioned on x_i. The global factor further decomposes as $q(\Phi) = \prod_{c=1}^{C} q(\boldsymbol{\phi}_c)$, so that Φ contains one globally learned reference distribution for each class.

We use Weibull distributions (Zhang et al., 2018) to approximate the Gamma posterior for scalable variational inference. This choice is motivated by two properties: (i) the Weibull family closely matches the shape of the Gamma family and is suitable for sparse, skewed and positive distributions; (ii) the KL divergence from the Weibull to the Gamma distributions has an analytic expression. For $q(z) = \text{Weibull}(z; k, \lambda)$ and a Gamma prior $p(z) = \text{Gamma}(z; a, b)$ with rate b,

$$\text{KL}(q\|p) = \log k - a\log\lambda - a\log b + \log\Gamma(a) + b\lambda\Gamma\left(1+\frac{1}{k}\right) + \frac{a\gamma_{\text{E}}}{k} - \gamma_{\text{E}} - 1, \tag{7}$$

where γ_{E} is the Euler–Mascheroni constant. We define the generalized posteriors of $\boldsymbol{\theta}$ and Φ as

$$q(\boldsymbol{\theta}_i \mid x_i) = \text{Weibull}(\boldsymbol{\theta}_i; \boldsymbol{k}_\theta, \boldsymbol{\lambda}_\theta), \qquad q(\Phi) = \text{Weibull}(\Phi; \boldsymbol{k}_\Phi, \boldsymbol{\lambda}_\Phi), \tag{8}$$

where $\boldsymbol{k}$ and $\boldsymbol{\lambda}$ are the Weibull shape and scale parameters, and $q(\boldsymbol{\phi}_c)$ is the marginal variational distribution of the c-th class reference. Both factors are regularized by the Gamma priors in Eq. 2.

The local and global Weibull parameters are generated differently: local parameters depend on the input representation, whereas global parameters are determined by learnable class-specific codes. Let $W = [\boldsymbol{w}_1^\top, \cdots, \boldsymbol{w}_C^\top] \in \mathbb{R}^{C\times L}$ collect these codes. The inference maps produce positive Weibull means and shapes, with the scales determined by the Weibull mean identity:

$$\begin{aligned} \boldsymbol{k}_\theta &= f_k(\boldsymbol{h}), & \boldsymbol{\lambda}_\theta &= \frac{f_\lambda(\boldsymbol{h})}{\Gamma(1+1/\boldsymbol{k}_\theta)}, \\ \boldsymbol{k}_\Phi &= f_k(\text{Softplus}(W)), & \boldsymbol{\lambda}_\Phi &= \frac{f_\lambda(\text{Softplus}(W))}{\Gamma(1+1/\boldsymbol{k}_\Phi)} \end{aligned} \tag{9}$$

where $\boldsymbol{h} = \text{Softplus}(f_{\text{NN}}(\boldsymbol{x})) \in \mathbb{R}^L$ is the hidden representation produced by a deterministic feature extractor, and L is the feature dimension. The networks f_k and f_λ infer the Weibull parameters, while the elementwise Softplus function enforces positivity. Consequently, $q(\boldsymbol{\theta}_i \mid x_i)$ varies with the input through h, whereas each $q(\boldsymbol{\phi}_c)$ is learned from the corresponding class code in W.

Replacing the latent distributions in Eq. 3 with their Weibull variational approximations in Eq. 8 yields a variational approximation to the Wasserstein discrepancy. These discrepancies then determine the categorical probabilities $\bar{\rho}_{ic}$ through Eq. 6. To obtain a differentiable stochastic realization of this categorical distribution, we apply the Gumbel–Softmax relaxation (Jang et al., 2016).Specifically, for independent noise variables $g_{ic} \sim \text{Gumbel}(0, 1)$ and temperature τ_{g}, we define

$$\widetilde{p}_{ic}(\boldsymbol{g}_i) = \frac{\exp\left[(\log\bar{\rho}_{ic} + g_{ic})/\tau_{\text{g}}\right]}{\sum_{c'=1}^{C}\exp\left[(\log\bar{\rho}_{ic'} + g_{ic'})/\tau_{\text{g}}\right]}. \tag{10}$$

3.3 Generalized Variational Inference

The posterior transport risk in Eq. 3 is generally intractable. In the variational implementation, we replace the posterior distributions in that definition with the corresponding variational factors from Eq. 8; for notational simplicity, we continue to denote the resulting variational transport cost by d_{ic}. Under the diagonal Weibull parameterization, coordinate-wise quantile matching yields the following closed-form transport cost (Villani, 2009; Peyré & Cuturi, 2019):

$$d_{ic} = \sum_{\ell=1}^{L}\left[\lambda_{\theta,i\ell}^2\Gamma\left(1+\frac{2}{k_{\theta,i\ell}}\right) + \lambda_{\phi,c\ell}^2\Gamma\left(1+\frac{2}{k_{\phi,c\ell}}\right) - 2\lambda_{\theta,i\ell}\lambda_{\phi,c\ell}\Gamma\left(1+\frac{1}{k_{\theta,i\ell}}+\frac{1}{k_{\phi,c\ell}}\right)\right]. \tag{11}$$

The derivation is provided in Appendix B.1. It makes the Wasserstein discrepancy differentiable with respect to the variational parameters. The resulting discrepancies determine the categorical probabilities through Eq. 6, while Eq. 11 provides a differentiable stochastic class assignment.

We define the generalized negative log-likelihood induced by the Wasserstein discrepancy as

$$\ell_{\mathrm{W}}(x_i, y_i) = \mathbb{E}_{\boldsymbol{g}_i}\left[-\log \widetilde{p}_{i,y_i}(\boldsymbol{g}_i)\right]. \tag{12}$$

Then, we impose a Dirichlet regularization term on the distance-induced class distribution, encouraging less concentrated class probabilities and improving the reliability of the predictive distribution. The resulting Wasserstein-based generalized variational objective is

$$\begin{aligned}\mathcal{J}(\mathcal{D}) = \sum_{i=1}^{N} \Big[& \ell_{\mathrm{W}}(x_i, y_i) + \beta_\rho \mathrm{KL}(\mathrm{Dir}(\boldsymbol{\alpha}_i) \,\|\, \mathrm{Dir}(\mathbf{1}_C)) \\ & + \beta_\theta \mathrm{KL}(q(\boldsymbol{\theta}_i \mid x_i) \,\|\, p(\boldsymbol{\theta}_i)) \Big] + \beta_\Phi \mathrm{KL}(q(\Phi) \,\|\, p(\Phi)).\end{aligned} \tag{13}$$

The Wasserstein-induced categorical term performs supervised label fitting through the decision rule defined by the transport costs, while the KL terms regularize the local and global variational distributions toward their respective Gamma priors. The Dirichlet KL term additionally regularizes the induced class distribution toward the uniform Dirichlet prior, discouraging overly concentrated predictive probabilities. Since the categorical probabilities are induced by the Wasserstein discrepancies, Eq. 13 constitutes a loss-based generalized variational objective. The coefficients β_θ, β_Φ, and β_ρ control the corresponding regularization strengths. The closed-form transport risk and analytic KL terms together enable efficient stochastic optimization.

3.4 OT-coupled Predictive Uncertainty

To clarify how the local and global variational distributions contribute to these costs, let $\mu_{\boldsymbol{\theta},i}$ and $\Sigma_{\boldsymbol{\theta},i}$ denote the mean and covariance of $q(\boldsymbol{\theta}_i \mid x_i)$, and let $\mu_{\boldsymbol{\phi},c}$ and $\Sigma_{\boldsymbol{\phi},c}$ denote those of $q(\boldsymbol{\phi}_c)$. The Wasserstein transport risk can be decomposed as

$$d_{ic} = \|\boldsymbol{\mu}_{\theta,i} - \boldsymbol{\mu}_{\phi,c}\|_2^2 + \mathrm{tr}(\Sigma_{\theta,i}) + \mathrm{tr}(\Sigma_{\phi,c}) - 2\mathrm{tr}(C_{ic}), \tag{14}$$

where $C_{ic} = \mathrm{Cov}_{\pi_{i,c}}(\boldsymbol{\theta}_i, \boldsymbol{\phi}_c)$ is the cross-covariance induced by the optimal coupling.

Eq. 14 shows how the mean and dispersion of the two distributions, together with their cross-covariance under the optimal coupling, jointly determine class-wise compatibility. The mean-discrepancy term measures the mismatch between the sample representation and the class reference. The dispersion of $q(\boldsymbol{\theta}_i \mid x_i)$ reflects input-conditioned latent variability and is associated with aleatoric uncertainty, while the dispersion of $q(\boldsymbol{\phi}_c)$ reflects uncertainty in the learned class reference and is associated with epistemic uncertainty. The cross-covariance term captures how variations in the sample and class-reference distributions are coupled under the optimal transport plan. These terms are coupled through the optimal transport structure, so the resulting uncertainty is not explicitly decomposed into independent aleatoric and epistemic components; rather, the two distributions provide distinct sources of uncertainty that jointly shape the transport-based predictive uncertainty.

We define the predictive uncertainty as the normalized entropy of the categorical distribution:

$$U_W(x_i) = \frac{1}{\log C}\left[\log \alpha_{i0} + \frac{1}{\tau \alpha_{i0}} \sum_{c=1}^{C} \alpha_{ic} \mathbb{E}_{\pi_{ic}^\star}\left[\|\boldsymbol{\theta}_i - \boldsymbol{\phi}_c\|_2^2\right]\right]. \tag{15}$$

We interpret $U_W(x_i)$ as an OT-coupled predictive uncertainty. Its underlying transport costs are influenced by the input-conditioned variability represented by $q(\boldsymbol{\theta}_i \mid x_i)$ and the uncertainty in the learned class references represented by $q(\boldsymbol{\phi}_c)$. Thus, their uncertainty contributions are jointly reflected in the resulting predictive uncertainty.

4 Experiments

Experiment Setup: We evaluate W-GBDL on conventional deep neural networks and fine-tuned large language models (LLMs). For image classification, we use a ResNet-50 backbone (He et al., 2016), while the LM experiments are conducted with LLaMA3-8B (Grattafiori et al., 2024).

Uncertainty Evaluation Metrics: A reliable uncertainty estimator should assign higher uncertainty to OOD samples than to ID samples and to misclassified samples than to correctly classified

Table 1: Performance comparison evaluated by classification accuracy and AUPR scores for OOD detection on CIFAR-10. →X indicates using X as OOD data. ResNet-50 serves as the backbone. Results are averaged over 5 runs, with mean and standard deviation.

Methods	**CIFAR10**		**→ CIFAR100**	**→ GTSRB**	**→ Places365**	**→ Food101**	**Mean**
	Cls Acc	Mis Detect	OOD Detect	OOD Detect	OOD Detect	OOD Detect	OOD Detect
MC Dropout	92.86 ± 0.71	$\mathbf{99.39 \pm 0.59}$	88.48 ± 3.28	89.71 ± 1.03	73.00 ± 4.89	81.44 ± 5.52	83.16 ± 2.41
BM	93.00 ± 0.38	97.40 ± 0.46	77.04 ± 0.88	82.21 ± 2.11	50.78 ± 2.52	58.22 ± 5.30	67.06 ± 1.87
CARD	91.06 ± 0.19	98.99 ± 0.07	85.02 ± 0.48	86.69 ± 1.93	66.08 ± 1.55	48.29 ± 1.53	71.52 ± 1.47
EDL	88.48 ± 0.32	98.75 ± 0.09	87.28 ± 0.35	87.45 ± 1.21	70.30 ± 1.17	80.07 ± 1.17	81.28 ± 0.98
I-EDL	88.90 ± 0.34	98.41 ± 0.14	84.32 ± 0.65	86.73 ± 2.01	65.25 ± 1.22	69.21 ± 1.10	76.38 ± 1.00
Re-EDL	91.70 ± 0.67	98.96 ± 0.14	88.42 ± 0.91	$\mathbf{93.43 \pm 1.92}$	73.33 ± 3.32	81.80 ± 3.50	84.25 ± 2.43
VBLL	92.69 ± 0.44	99.27 ± 0.21	86.33 ± 0.68	87.86 ± 1.71	69.94 ± 2.30	74.56 ± 2.58	79.67 ± 1.83
BNDL	92.95 ± 0.48	99.38 ± 0.07	87.36 ± 0.74	91.18 ± 2.16	73.26 ± 2.18	79.47 ± 7.70	81.34 ± 4.27
W-GBDL	$\mathbf{93.11 \pm 0.21}$	99.34 ± 0.03	$\mathbf{90.05 \pm 0.51}$	92.68 ± 1.78	$\mathbf{76.70 \pm 1.28}$	$\mathbf{85.54 \pm 1.67}$	$\mathbf{86.24 \pm 0.52}$
Deep Ensemble	94.80	**99.65**	90.76	92.96	75.64	84.33	85.92
W-GBDL+Deep Ensemble	**94.84**	99.51	**92.36**	**95.47**	**79.89**	**89.29**	**89.25**

Table 2: Calibration performance on CIFAR-10 and CIFAR-100. ResNet-50 serves as the backbone.

Method	**CIFAR-10**			**CIFAR-100**		
	ACC ↑	ECE ↓	NLL ↓	ACC ↑	ECE ↓	NLL ↓
MC Dropout	92.86 ± 0.71	3.56 ± 0.15	0.30 ± 0.01	70.22 ± 0.83	15.49 ± 0.24	1.77 ± 0.01
BM	93.00 ± 0.38	2.09 ± 0.27	0.28 ± 0.01	72.00 ± 0.31	12.13 ± 0.16	1.41 ± 0.00
Re-EDL	91.70 ± 0.67	8.24 ± 0.19	0.63 ± 0.01	69.97 ± 0.16	$\mathbf{10.21 \pm 0.17}$	1.73 ± 0.01
VBLL	92.69 ± 0.44	4.65 ± 0.17	0.30 ± 0.01	72.39 ± 0.21	13.13 ± 0.19	1.41 ± 0.01
BNDL	92.95 ± 0.48	4.62 ± 0.26	0.32 ± 0.01	70.87 ± 0.34	20.71 ± 0.25	2.05 ± 0.01
W-GBDL	$\mathbf{93.11 \pm 0.21}$	$\mathbf{1.31 \pm 0.38}$	$\mathbf{0.23 \pm 0.01}$	$\mathbf{74.44 \pm 0.25}$	11.88 ± 0.28	$\mathbf{1.12 \pm 0.01}$

ones. We evaluate both OOD and misclassification detection using the area under the precision–recall curve (AUPR), with ID samples treated as the positive class and OOD samples as the negative class(Deng et al., 2023; Yoon & Kim, 2024; Chen et al., 2025). Each method is evaluated using its corresponding uncertainty measure, while OT-coupling predictive entropy is used for W-GBDL. We additionally report classification accuracy (ACC), expected calibration error (ECE), and negative log-likelihood (NLL). Details for evaluation metrics are provided in Appendix A.1.

4.1 Image Classification Performance

Baselines: We compare W-GBDL with MC Dropout(Gal & Ghahramani, 2016), Deep Ensemble(Lakshminarayanan et al., 2017), BM(Joo et al., 2020), CARD(Han et al., 2022), EDL(Sensoy et al., 2018), I-EDL(Deng et al., 2023), Re-EDL(Chen et al., 2025), VBLL(Harrison et al., 2024) and BNDL(Hu et al., 2025). All methods are trained on CIFAR-10 as the ID dataset and evaluated on CIFAR-100(Krizhevsky et al., 2009), GTSRB, Places365(Zhou et al., 2017), and Food-101(Bossard et al., 2014) as OOD datasets. To ensure a fair comparison, all methods use ResNet-50 backbone and follow the same evaluation protocol. See Appendix A.3 for experimental details.

Experiment Results: Table 1 reports the closed-set classification accuracy, the misclassification-detection score on CIFAR-10, and the OOD detection scores for the external datasets. The OOD

Table 3: Results of the noisy setting on CIFAR-10. Noisy samples are generated by adding zero-mean Gaussian noises with standard deviations of [0.025:0.025:0.200] to the testing samples.

SD of Noise	0.025			0.050			0.075			0.100		
Method	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg
MC Dropout	**90.23**	53.29	71.76	**76.48**	64.14	70.31	**55.20**	76.95	66.08	35.97	86.16	61.06
BM	88.33	52.03	70.18	70.77	57.47	64.12	50.04	62.37	56.20	35.98	66.59	51.29
EDL	81.01	54.80	67.90	63.82	67.20	65.51	45.70	76.90	61.30	33.55	81.97	57.76
Re-EDL	89.87	54.71	**72.29**	73.70	68.03	70.87	50.35	79.44	64.89	33.65	86.51	60.08
W-GBDL	89.00	**55.37**	72.19	74.55	**68.73**	**71.64**	53.86	**83.32**	**68.59**	**37.76**	**91.83**	**64.79**
SD of Noise	0.125			0.150			0.175			0.200		
Method	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg	Cls Acc	Noisy Detec	Avg
MC Dropout	24.87	90.90	57.89	19.08	92.54	55.81	15.40	93.53	54.63	13.89	94.05	53.97
BM	27.41	71.56	49.48	22.42	75.31	48.86	19.34	78.21	48.77	17.28	80.75	49.01
EDL	26.10	84.75	55.42	22.11	86.28	54.19	19.88	87.33	53.60	19.58	88.11	53.85
Re-EDL	25.74	90.56	58.15	21.42	92.62	57.02	18.42	94.01	56.22	15.85	94.76	55.31
W-GBDL	**28.97**	**96.00**	**62.49**	**24.26**	**97.86**	**61.06**	**21.41**	**98.76**	**60.08**	**19.24**	**99.30**	**59.27**

Table 4: Calibration performance of uncertainty-aware methods on fine-tuned LLaMA3-8B.

Metrics	Method	ARC-C	ARC-E	OBQA	CSQA	SciQ	RACE	Mean
Acc ↑	MAP	79.74 ± 0.27	92.27 ± 0.17	88.60 ± 0.87	81.51 ± 0.60	93.37 ± 0.12	88.15 ± 0.31	87.27 ± 0.42
	MCD	79.55 ± 0.21	92.25 ± 0.13	88.63 ± 0.84	81.52 ± 0.60	93.30 ± 0.20	88.12 ± 0.25	87.23 ± 0.37
	Ens	79.50 ± 0.10	92.26 ± 0.41	88.53 ± 0.59	81.39 ± 0.63	93.27 ± 0.15	88.09 ± 0.09	87.17 ± 0.33
	LA	77.79 ± 0.39	92.18 ± 0.24	88.34 ± 0.62	81.30 ± 0.44	93.37 ± 0.15	88.09 ± 0.18	86.85 ± 0.34
	EDL	79.35 ± 1.11	92.31 ± 0.76	87.67 ± 0.31	80.67 ± 0.29	93.13 ± 0.36	87.76 ± 0.21	86.82 ± 0.51
	VID	79.99 ± 0.13	92.30 ± 0.25	87.57 ± 0.29	80.82 ± 0.94	92.93 ± 0.23	88.23 ± 0.25	86.98 ± 0.35
	I-EDL	80.37 ± 0.48	92.76 ± 0.47	88.52 ± 0.50	81.13 ± 0.39	93.47 ± 0.06	85.91 ± 0.40	87.03 ± 0.42
	IB-EDL	**81.14 ± 0.09**	**92.55 ± 0.15**	89.00 ± 0.40	81.71 ± 0.38	93.57 ± 0.15	88.03 ± 0.21	87.67 ± 0.23
	W-GBDL	80.93 ± 0.12	92.38 ± 0.18	**89.60 ± 0.34**	**82.55 ± 0.27**	**93.70 ± 0.12**	**88.37 ± 0.13**	**87.92 ± 0.27**
ECE ↓	MAP	19.68 ± 0.43	7.18 ± 0.14	10.52 ± 0.87	17.29 ± 0.57	5.74 ± 0.08	7.95 ± 0.35	11.39 ± 0.41
	MCD	19.91 ± 0.39	7.10 ± 0.02	10.48 ± 0.86	16.98 ± 0.10	5.74 ± 0.09	7.93 ± 0.35	11.36 ± 0.30
	Ens	18.20 ± 0.17	3.81 ± 1.60	10.08 ± 0.90	16.11 ± 1.63	5.72 ± 0.24	7.80 ± 0.19	10.29 ± 0.79
	LA	18.49 ± 0.44	3.33 ± 0.66	5.26 ± 1.30	6.62 ± 0.10	**2.47 ± 0.08**	3.61 ± 0.27	6.63 ± 0.48
	EDL	6.52 ± 0.12	5.94 ± 0.87	8.28 ± 1.62	7.43 ± 1.48	11.13 ± 1.11	6.51 ± 0.80	7.64 ± 1.00
	VID	10.96 ± 0.40	3.33 ± 0.40	5.99 ± 1.41	8.38 ± 0.93	2.60 ± 0.13	4.58 ± 0.14	5.97 ± 0.57
	I-EDL	5.08 ± 1.94	9.69 ± 1.58	7.57 ± 0.52	6.95 ± 0.59	13.07 ± 0.24	14.52 ± 1.15	9.48 ± 1.00
	IB-EDL	**2.78 ± 0.87**	2.70 ± 0.58	**2.34 ± 0.61**	4.34 ± 0.20	3.86 ± 0.86	4.47 ± 0.31	3.42 ± 0.57
	W-GBDL	3.77 ± 0.28	**2.42 ± 0.41**	4.62 ± 0.52	**3.79 ± 0.27**	2.51 ± 0.38	**2.16 ± 0.21**	**3.21 ± 0.42**
NLL ↓	MAP	2.25 ± 0.08	0.61 ± 0.02	0.78 ± 0.05	1.25 ± 0.04	0.36 ± 0.00	0.47 ± 0.01	0.95 ± 0.03
	MCD	2.27 ± 0.09	0.59 ± 0.00	0.77 ± 0.05	1.22 ± 0.03	0.36 ± 0.01	0.45 ± 0.03	0.94 ± 0.04
	Ens	2.02 ± 0.07	0.43 ± 0.09	0.74 ± 0.05	1.22 ± 0.12	0.35 ± 0.01	0.46 ± 0.05	0.87 ± 0.07
	LA	0.80 ± 0.01	0.42 ± 0.01	0.42 ± 0.01	0.62 ± 0.04	0.22 ± 0.00	0.37 ± 0.01	0.46 ± 0.02
	EDL	0.74 ± 0.02	0.45 ± 0.02	0.45 ± 0.02	0.68 ± 0.01	0.31 ± 0.01	0.43 ± 0.01	0.49 ± 0.01
	VID	0.78 ± 0.01	0.46 ± 0.02	0.46 ± 0.02	0.72 ± 0.03	0.28 ± 0.01	0.36 ± 0.00	0.49 ± 0.01
	I-EDL	0.72 ± 0.02	0.43 ± 0.00	0.43 ± 0.00	0.68 ± 0.01	0.32 ± 0.01	0.52 ± 0.02	0.53 ± 0.01
	IB-EDL	0.69 ± 0.01	0.40 ± 0.03	0.40 ± 0.03	0.66 ± 0.01	0.25 ± 0.01	0.35 ± 0.00	0.45 ± 0.02
	W-GBDL	**0.67 ± 0.02**	**0.34 ± 0.01**	**0.39 ± 0.02**	**0.63 ± 0.01**	**0.21 ± 0.01**	**0.32 ± 0.00**	**0.43 ± 0.01**

score is computed from the uncertainty ranking produced by each method; higher values indicate better separation between ID and OOD samples. W-GBDL achieves the highest average OOD detection scores across all four OOD datasets and the highest classification accuracy among the single-model methods. Table 2 further shows that W-GBDL achieves lower ECE and NLL on both CIFAR-10 and CIFAR-100, indicating improved calibration and likelihood quality.

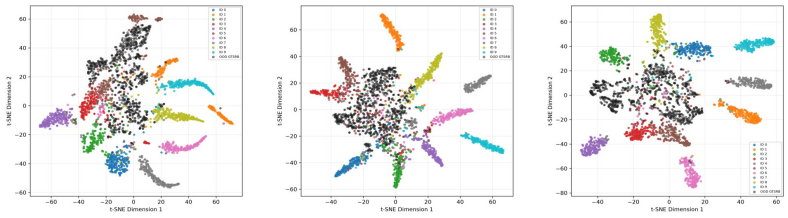
Noisy Setting of Classification: Real-world inputs can be corrupted by noise, which may affect both prediction accuracy and uncertainty estimation. We therefore evaluate W-GBDL under increasing levels of input noise. Specifically, zero-mean isotropic Gaussian noise is added to the CIFAR-10 ID test set to construct corrupted samples, and we assess both classification accuracy and the quality of noisy-sample uncertainty estimates. Table 3 reports the classification accuracy and AUPR for noisy-sample detection. W-GBDL consistently achieves higher AUPR than the baseline methods across almost all noise levels, with the performance gap generally increasing as the noise intensity grows. These results indicate that W-GBDL can effectively reflect the increasing uncertainty associated with progressively corrupted inputs while maintaining competitive classification performance.

4.2 Large Language Model Calibration

Since fine-tuned LLMs are prone to overconfidence under limited supervision, leading to miscalibrated predictions, we further extend our evaluation beyond image tasks to LLM fine-tuning, thereby assessing the effectiveness of our method across different data modalities and model architectures.

Baselines: We compare W-GBDL with MAP fine-tuning, Deep Ensembles (Ens), MC Dropout (MCD), Laplace-LoRA (LA)(Yang et al., 2024), and EDL-family methods including vanilla EDL, VID, I-EDL, and IB-EDL(Li et al., 2025). We evaluate on six multiple-choice benchmarks: ARC-C and ARC-E (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2019), SciQ (Welbl et al., 2017), and RACE (Lai et al., 2017). For each dataset the label space corresponds to answer options. See Appendix A.4 for experimental details.

Experiment Results: Table 4 summarizes results on LLaMA3-8B. MAP fine-tuning exhibits severe overconfidence, as reflected by its relatively large average ECE and NLL. Across the evaluated datasets, W-GBDL achieves the best average ACC while also attaining the lowest average ECE and NLL among the compared methods. These results demonstrate the effectiveness of W-GBDL in improving uncertainty estimation and mitigating overconfidence in low-data fine-tuning.



(a) MC Dropout (b) EDL (c) W-GBDL

Figure 4: t-SNE visualization of feature distributions for ID (CIFAR-10) and OOD (GTSRB).

Table 5: Component ablation. DUQ is the Euclidean reference; Δ denotes the detect gain over DUQ.

Method	Variational posterior		Geometry	Performance (%)		
	Local q_θ	Global q_ϕ	Distance	Cls Acc ↑	OOD Detect ↑	Δ
DUQ	–	–	ℓ_2	92.20±0.23	77.84±0.47	–
Global only	–	✓	W_2	92.32±0.31	81.90±0.28	+4.06
Local only	✓	–	W_2	92.44±0.19	80.03±0.42	+2.19
W-GBDL	✓	✓	W_2	**93.11±0.21**	**86.24±0.52**	**+8.40**

Table 6: Ablation of local and global KL regularization.

KL terms	Performance (%)	
	Cls Acc ↑	OOD Detect ↑
None	92.37±0.32	82.56±0.44
Global only	92.96±0.26	83.59±0.35
Local only	92.66±0.23	83.26±0.49
Both	**93.11±0.21**	**86.24±0.52**

4.3 Qualitative and Mechanistic Analysis

Uncertainty Behavior and Representation Geometry: We examine whether OT-coupled predictive uncertainty responds to both model knowledge and data uncertainty. As shown in Fig. 3, uncertainty decreases monotonically with training-set size, consistent with the epistemic-related interpretation of the globally shared class-reference component. Table 3 shows that W-GBDL increasingly separates corrupted from clean inputs as perturbations grow, consistent with the aleatoric-related interpretation of the input-conditioned component.

Fig. 4 visualizes the representations of CIFAR-10 ID samples and GTSRB OOD samples by t-SNE (Van der Maaten & Hinton, 2008). W-GBDL produces more compact ID clusters, while most OOD samples lie between or outside the ID regions. This pattern is consistent with its distributional class reference formulation, which encourages ID samples to remain close to their corresponding class references while assigning lower compatibility to poorly supported samples.

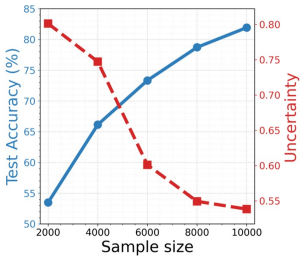


Figure 3: Test accuracy (blue, solid) and predictive uncertainty (red, dashed) across training-set sizes.

Component Attribution: Tables 5 and 6 examine the roles of distributional representation, transport geometry, and posterior regularization. The one-sided variants, which combine Wasserstein geometry with either a local sample distribution or a global class reference distribution, improve detection over the Euclidean DUQ (Van Amersfoort et al., 2020); jointly modeling both sides performs best. The KL ablation shows the same ordering: regularizing either posterior provides a partial benefit, whereas regularizing both the local and global factors gives the strongest performance. These results support the benefit of jointly modeling and regularizing the sample–class comparison. Additional visualizations, inference costs, and hyperparameter analyses are reported in Appendix C.

5 Conclusion

We presented the Wasserstein Generalized Bayesian Decision Layer (W-GBDL), a probabilistic framework that represents both inputs and class references as non-negative distributions and performs classification through Wasserstein geometry. By modeling prediction as distribution-to-distribution matching, W-GBDL derives uncertainty from OT coupling within a unified classification framework. This distributional formulation improves predictive accuracy and uncertainty estimation quality, with consistent performance across different data modalities and backbone architectures, demonstrating its general applicability to uncertainty-aware classification. More broadly, our findings highlight the potential of explicit distributional representations and Wasserstein geometry as a general foundation for reliable uncertainty-aware deep learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing the manuscript to improve its readability, clarity, grammar, and overall presentation. We also used generative AI tools to assist with editing and refining software code. We have reviewed all AI-assisted work, and all AI-assisted code was manually inspected and tested by the authors to verify its correctness. We take full responsibility for the final content of this work, including all text, claims, code, experimental results, and conclusions.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Revisiting essential and nonessential settings of evidential deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

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

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

Zhibin Duan, Tiansheng Wen, Muyao Wang, Bo Chen, and Mingyuan Zhou. A non-negative vae: the generalized gamma belief network. *arXiv preprint arXiv:2408.03388*, 2024.

Xinjie Fan, Shujian Zhang, Korawat Tanwisuth, Xiaoning Qian, and Mingyuan Zhou. Contextual dropout: An efficient sample-dependent dropout module. *arXiv preprint arXiv:2103.04181*, 2021.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2053–2061, 2015.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *International Conference on Learning Representations*, volume 2024, pp. 54251–54280, 2024.

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

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

Xinyue Hu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. Enhancing uncertainty estimation and interpretability via bayesian non-negative decision layer. *arXiv preprint arXiv:2505.22199*, 2025.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. doi: 10.48550/arXiv.1611.01144.

Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. In *International conference on machine learning*, pp. 4950–4961. PMLR, 2020.

Mira Juergens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Yawei Li, David Rügamer, Bernd Bischl, and Mina Rezaei. Calibrating llms with information-theoretic evidential deep learning. *arXiv preprint arXiv:2502.06351*, 2025.

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

Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity. *arXiv preprint arXiv:2106.14568*, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020b.

Andrey Malinin, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. Regression prior networks. *arXiv preprint arXiv:2006.11590*, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

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

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Sebastian W Ober and Carl Edward Rasmussen. Benchmarking the neural linear model for regression. *arXiv preprint arXiv:1912.08416*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

Pola Schwöbel, Martin Jørgensen, Sebastian W Ober, and Mark Van Der Wilk. Last layer marginal likelihood for invariance learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3542–3555. PMLR, 2022.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Marcin Tomczak, Siddharth Swaroop, Andrew Foong, and Richard Turner. Collapsed variational bounds for bayesian neural networks. *Advances in Neural Information Processing Systems*, 34: 25412–25426, 2021.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in neural information processing systems*, 37:67758–67794, 2024.

Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent derivative bayesian last layer networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1198–1206. PMLR, 2021.

Noah Weber, Janez Starc, Arpit Mittal, Roi Blanco, and Lluís Màrquez. Optimizing over a bayesian last layer. In *NeurIPS workshop on Bayesian Deep Learning*, 2018.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Adam Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *international conference on learning representations*, volume 2024, pp. 1812–1842, 2024.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep learning. *arXiv preprint arXiv:2409.08754*, 2024.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. Whai: Weibull hybrid autoencoding inference for deep topic modeling. *arXiv preprint arXiv:1803.01328*, 2018.

Bolei Zhou, Agata 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, 2017.

Mingyuan Zhou, Yulai Cong, and Bo Chen. The poisson gamma belief network. *Advances in Neural Information Processing Systems*, 28, 2015.

A Experimental Setup

This appendix provides the implementation details, analytic derivation, and additional experiments that complement the main paper. Unless stated otherwise, tables report means with standard deviations; all experiments use five independent seeds (1000–1004).

All experiments were conducted on a workstation with a 13th Gen Intel Core i7-13700 CPU (16 physical cores and 24 threads), 128 GB of RAM, and one NVIDIA GeForce RTX 4090 GPU. The implementation uses Python 3.12 and PyTorch 2.12. We use the same preprocessing, optimization, and evaluation protocol for all methods within a given experiment so that differences are attributable to the decision layer rather than to data handling.

A.1 Details for Evaluation Metrics

Classification and Calibration Metrics: We report classification accuracy (ACC), expected calibration error (ECE)(Naeini et al., 2015), and negative log-likelihood (NLL). ACC measures predictive performance, whereas ECE and NLL jointly assess the quality of the estimated uncertainties. NLL is defined as:

$$\text{NLL} = \frac{1}{N}\sum_{n=1}^{N} -\log p(y_n \mid x_n) \tag{16}$$

That is, it is the expected negative log probability of the correct class under the model.

ECE measures how a model's confidence aligns with the accuracy of its predictions. It can be computed by binning the predictions by their confidence. We then compute a weighted average of the difference between the accuracy and confidence with in each bin:

$$\text{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} |\text{acc}(B_k) - \text{conf}(B_k)| \tag{17}$$

where K is the number of bins and B_k is the set of samples in the k-th bin. Following Wang et al. (2024), we use $K = 15$ in all experiments.

Direction of AUPR for uncertainty evaluation: For OOD and noisy-sample evaluation, our primary objective is to assess the quality of the uncertainty ranking rather than to deploy a binary OOD detector at a selected threshold. Accordingly, we use the negative uncertainty as a confidence score and define the in-distribution (ID), or clean, samples as the positive class when computing AUPR. This is the ID-positive (confidence-ranking) convention: a reliable uncertainty estimator should assign larger scores to supported ID samples and smaller scores to OOD or corrupted samples (Deng et al., 2023; Yoon & Kim, 2024; Chen et al., 2025). The corresponding OOD-positive curve can also be computed by reversing the labels and using the uncertainty score, but under the present ID/OOD pairings several baselines are close to precision–recall saturation, which provides little resolution for comparing uncertainty quality. We therefore use ID-positive AUPR consistently in the main tables and interpret it as a ranking metric, not as evidence that the model itself performs thresholded OOD detection.

A.2 Datasets and Preprocessing

For image classification, CIFAR-10 is the in-distribution (ID) dataset. CIFAR-100, GTSRB, Places365, and Food-101 are used as OOD datasets under the protocol described in the main paper. CIFAR-10 and CIFAR-100 training images are augmented with random horizontal flips and random crops with four pixels of padding. GTSRB, Places365, and Food-101 images are resized to 32×32 pixels before evaluation. All datasets use their standard class labels and publicly available train/test splits.

For language-model experiments, we use ARC-Challenge (ARC-C), ARC-Easy (ARC-E), OpenBookQA (OBQA), CommonsenseQA (CSQA), SciQ, and the "middle" subset of RACE. Each example is converted to a multiple-choice prompt that contains the question, answer choices, and an answer prefix. The target label is the position of the correct answer choice.

A.3 Image Classification Training

The image experiments use ResNet-50 as the deterministic backbone trained from scratch. Images are processed at 32×32 resolution. We split the ID training data into 95% training and 5% validation, use a batch size of 64, and optimize with Adam at learning rate 10^{-4} for at most 200 epochs. The validation split is used for model selection and temperature-related analyses; the test split is kept separate for the final metrics. In the noisy-setting experiment, models are trained only on clean CIFAR-10 images. At test time, zero-mean isotropic Gaussian noise with standard deviation in $\{0.025, 0.050, \ldots, 0.200\}$ is added to the test images.

A.4 LLaMA Fine-tuning

We use LLaMA3-8B as the base causal language model. The model is loaded in bfloat16 precision and fine-tuned with LoRA. Unless otherwise noted, LoRA uses rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.1; the W-GBDL head is optimized jointly with the LoRA parameters. The default maximum sequence length is 300 tokens (512 for RACE), and the per-device training and evaluation batch sizes are both 4. We use learning rate 5×10^{-5}, weight decay 10^{-3}, cosine scheduling, 20 warm-up steps, and at most 10,080 training steps. CSQA uses 30,000 steps. Gradients are clipped to norm 20. Evaluation and checkpointing are performed at the end of every epoch, with only the most recent checkpoint retained. The hidden-state variant feeds the final non-padding LLaMA hidden state to the W-GBDL head.

B Theoretical Derivations for W-GBDL

B.1 Theoretical Guarantees for Closed-form Wasserstein Distance

The Wasserstein decision rule in Eq. 11 is available in closed form for the diagonal Weibull variational factors used by W-GBDL. We state explicitly the structural assumption and the standard one-dimensional optimal-transport theorem used in the derivation.

Assumption B.1 (Coordinate independence)**.** *For each sample–class pair, the local and class-reference variational factors have independent coordinates, so their joint distributions factorize as products of their one-dimensional marginals. The transport cost is the squared Euclidean cost.*

Theorem B.1 (One-dimensional quantile formula)**.** *Let P and Q be probability measures on $\mathbb{R}$ with finite second moments, and let F_P^{-1} and F_Q^{-1} denote their generalized quantile functions. Then*

$$W_2^2(P,Q) = \int_0^1 \left| F_P^{-1}(u) - F_Q^{-1}(u) \right|^2 du. \tag{18}$$

Equivalently, the coupling $(F_P^{-1}(U), F_Q^{-1}(U))$, for $U \sim \mathrm{Uniform}(0,1)$, is optimal for the squared Euclidean cost (Villani, 2009; Peyré & Cuturi, 2019).

Theorem B.1 is the standard monotone rearrangement theorem for one-dimensional optimal transport; it is a mathematical result, not an additional modeling assumption.

Weibull specialization. For one feature coordinate, let $Z \sim \mathrm{Weibull}(k_\theta, \lambda_\theta)$ and $V \sim \mathrm{Weibull}(k_\phi, \lambda_\phi)$, where all shape and scale parameters are strictly positive. Their quantile functions are

$$F_\theta^{-1}(u) = \lambda_\theta[-\log(1-u)]^{1/k_\theta}, \qquad F_\phi^{-1}(u) = \lambda_\phi[-\log(1-u)]^{1/k_\phi}. \tag{19}$$

Applying Theorem B.1 gives

$$\begin{aligned} W_2^2(Z,V) &= \int_0^1 \left(F_\theta^{-1}(u) - F_\phi^{-1}(u) \right)^2 du \\ &= \lambda_\theta^2 \int_0^1 [-\log(1-u)]^{2/k_\theta}\, du + \lambda_\phi^2 \int_0^1 [-\log(1-u)]^{2/k_\phi}\, du \\ &\quad - 2\lambda_\theta\lambda_\phi \int_0^1 [-\log(1-u)]^{1/k_\theta + 1/k_\phi}\, du. \end{aligned} \tag{20}$$

With $t = -\log(1-u)$, we have $du = \exp(-t)\,dt$. Thus, for $a > -1$, the integral is

$$\int_0^\infty t^a \exp(-t)\,dt = \Gamma(a+1), \tag{21}$$

and substitution of the corresponding exponents yields

$$W_2^2(Z,V) = \lambda_\theta^2 \Gamma\left(1+\frac{2}{k_\theta}\right) + \lambda_\phi^2 \Gamma\left(1+\frac{2}{k_\phi}\right) - 2\lambda_\theta\lambda_\phi \Gamma\left(1+\frac{1}{k_\theta}+\frac{1}{k_\phi}\right). \tag{22}$$

Corollary B.1 (Product-distribution decomposition)**.** *Let $P = \bigotimes_{\ell=1}^L P_\ell$ and $Q = \bigotimes_{\ell=1}^L Q_\ell$ be the local and class-reference variational distributions, respectively, with independent coordinates. Under the squared Euclidean cost $\|z-v\|_2^2 = \sum_{\ell=1}^L |z_\ell - v_\ell|^2$, the multivariate squared 2-Wasserstein distance separates as*

$$W_2^2(P,Q) = \sum_{\ell=1}^L W_2^2(P_\ell, Q_\ell). \tag{23}$$

Proof. Indeed, every coupling induces a coupling of each coordinate pair, giving the lower bound by the sum of the marginal optimal costs; the product of the coordinate-wise optimal couplings attains this bound. □

Under Assumption B.1, Corollary B.1 and Theorem B.1 apply coordinate-wise. Hence, summing Eq. 22 over the L coordinates gives the transport cost d_{ic} in Eq. 11. Strictly positive Weibull shape and scale parameters ensure finite second moments. The resulting expression is differentiable with respect to both local and class-level Weibull parameters, so it can be evaluated directly during optimization without Monte Carlo samples at the decision layer.

B.2 Wasserstein Risk Bounds and Point-Representation Limit

The Wasserstein discrepancy has a direct interpretation in terms of the location and dispersion of the two variational distributions. Let P and Q be probability measures on $\mathbb{R}^L$ with finite second moments, and write their means and covariance matrices as (μ_P, Σ_P) and (μ_Q, Σ_Q), respectively.

Theorem B.2 (Mean and dispersion bounds)**.** *For the squared Euclidean transport cost,*

$$\|\mu_P - \mu_Q\|_2^2 \;\le\; W_2^2(P,Q) \;\le\; \|\mu_P - \mu_Q\|_2^2 + \operatorname{tr}(\Sigma_P) + \operatorname{tr}(\Sigma_Q). \tag{24}$$

Moreover, for an optimal coupling $\pi^\star$,

$$\begin{aligned} W_2^2(P,Q) = \|\mu_P - \mu_Q\|_2^2 + \operatorname{tr}(\Sigma_P) + \operatorname{tr}(\Sigma_Q) \\ - 2\operatorname{tr}(\operatorname{Cov}_{\pi^\star}(X,Y))\,, \end{aligned} \tag{25}$$

where $(X,Y) \sim \pi^\star$.

For any coupling π of P and Q, expanding the squared distance gives

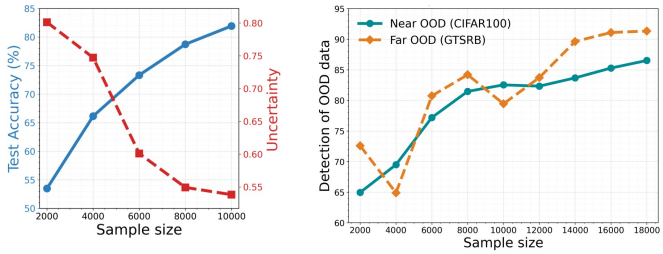
$$\mathbb{E}_\pi \|X-Y\|_2^2 = \|\mu_P - \mu_Q\|_2^2 + \operatorname{tr}(\Sigma_P) + \operatorname{tr}(\Sigma_Q) - 2\operatorname{tr}(\operatorname{Cov}_\pi(X,Y))\,.$$

The lower bound follows from Jensen's inequality applied to $X - Y$. For the upper bound, use the independent coupling $P \otimes Q$, whose cross-covariance is zero. Taking the infimum over couplings gives Eq. 25 for an optimal coupling and Eq. 24.

The first term in Eq. 25 is the mismatch between the distribution means, while the remaining terms quantify the effect of distributional dispersion and coupling alignment. In particular, the Wasserstein risk reduces to the usual pointwise squared distance when both distributions are degenerate:

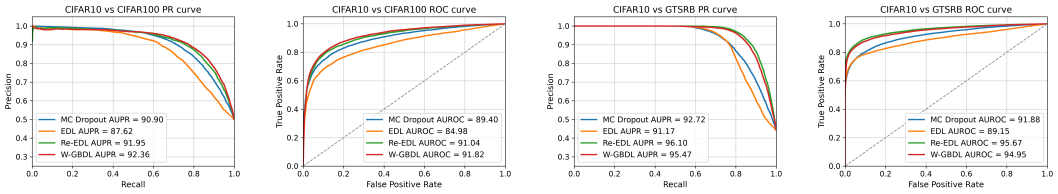
$$W_2^2(\delta_a, \delta_b) = \|a-b\|_2^2. \tag{26}$$

Thus, the distributional decision rule can be viewed as a generalization of point-based distance, while additionally capturing the uncertainty-bearing spread of the local and class-reference factors.



(a) Acc and uncertainty (b) Near and far OOD AUPR

Figure 5: Sample-size analysis. Curves show mean performance as the number of labeled ID training examples changes. In (a), accuracy and uncertainty use the left and right axes, respectively; in (b), the two curves correspond to near-OOD and far-OOD evaluation.



(a) CIFAR-100 PR (b) CIFAR-100 ROC (c) GTSRB PR (d) GTSRB ROC

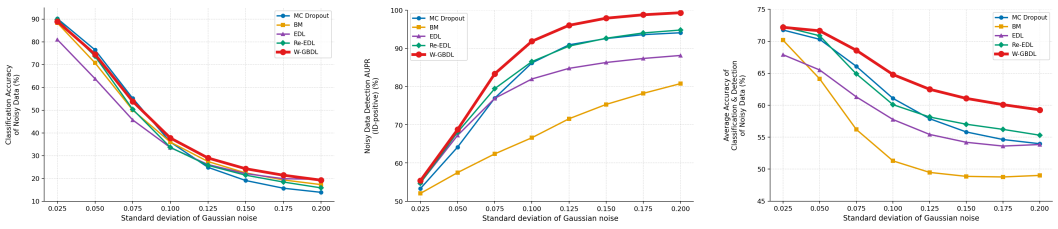
Figure 6: Precision–recall (PR) and receiver operating characteristic (ROC) curves for OOD detection with CIFAR-10 as ID. The two OOD datasets are CIFAR-100 and GTSRB.

C Additional Experimental Results

This section reports complementary results for calibration, sample efficiency, noise robustness, OOD ranking, and the two principal design choices of W-GBDL. The figures use the same ID/OOD definitions and uncertainty score as the main experiments.

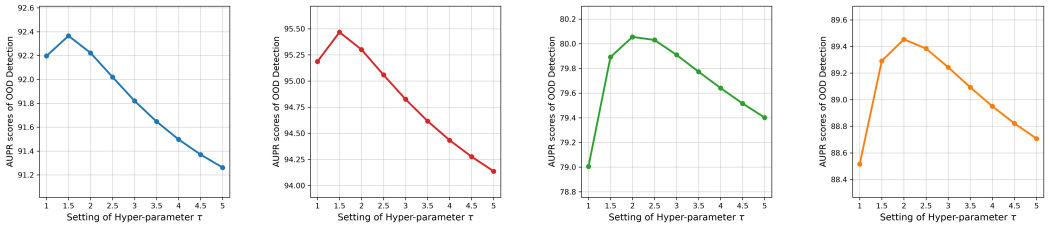
Sample efficiency and OOD ranking: Figure 5 examines how uncertainty changes as more labeled ID examples are available. Predictive uncertainty decreases as the training set grows, while test accuracy increases and OOD AUPR improves for both near-OOD and far-OOD datasets. The joint trend is consistent with the intended interpretation of the local posterior: more training evidence makes the sample distribution better supported by the learned class references.

OOD precision–recall and ROC curves: The curves in Figure 6 provide a threshold-level view of the OOD results. W-GBDL maintains a favorable precision–recall trade-off and a large ROC area on both CIFAR-100 and GTSRB, complementing the summary AUPR values in the main paper. The GTSRB panels also illustrate that the method remains effective when the OOD images differ substantially in semantic content from CIFAR-10.



(a) Classification accuracy (b) Noisy-sample AUPR (c) Average score

Figure 7: Performance under increasing Gaussian noise on CIFAR-10. The average score is the arithmetic mean of classification accuracy and noisy-sample AUPR.



(a) CIFAR-100 (b) GTSRB (c) Places365 (d) Food-101

Figure 8: Sensitivity of OOD-detection AUPR to the Gibbs temperature τ. CIFAR-10 is the ID dataset and each panel corresponds to one OOD dataset.

Table 7: Overall inference cost. We report per-image GFLOPs, median and 90th percentile latency (P50/P90), and the peak GPU memory usage. Lower values indicate better efficiency.

Methods	GFLOPs	P50 (ms)	P90 (ms)	Peak GPU Memory (MiB)
ResNet-50	2.596	4.02	4.53	110.38
MC Dropout	77.978	146.23	149.84	108.14
BM	2.597	4.61	5.17	108.13
CARD	89.745	846.26	852.78	1292.62
Deep Ensemble	12.987	21.80	23.54	478.60
EDL	2.596	4.03	4.55	108.13
I-EDL	2.598	4.92	5.51	199.62
Re-EDL	2.596	3.87	4.30	110.38
W-GBDL	2.604	4.61	5.24	124.47

Robustness to Gaussian noise: Figure 7 evaluate detection of corrupted CIFAR-10 test images. W-GBDL has the strongest combined accuracy–detection score at all but the smallest noise level and its noisy-sample AUPR grows to 99.30% at standard deviation 0.200. This monotonic improvement in detection, despite the expected decrease in classification accuracy, indicates that the transport-based uncertainty responds to progressively stronger input corruption.

Sensitivity to the temperature parameter: The temperature τ controls how sharply transport costs are converted into class compatibilities. We use $\tau = 1$ by default and, when tuning it, select the value with the lowest validation NLL without using any OOD information. Figure 8 reports AUPR for four OOD datasets as τ ranges from 1 to 5.The maximum performance variation is about 1.5 percentage points across this range, indicating limited sensitivity to τ. The best value differs across datasets. Nevertheless, $\tau = 1$ remains competitive without dataset-specific tuning.

Inference cost. We compare the computational overhead of different uncertainty estimation methods in terms of per-image GFLOPs, inference latency, and peak GPU memory usage. As shown in Table 7, MC Dropout and CARD incur substantially higher computational costs due to repeated stochastic inference, while Deep Ensemble requires multiple network evaluations, resulting in increased inference latency and memory consumption. In contrast, W-GBDL has an inference cost comparable to methods based on a single forward pass, indicating that it introduces only limited additional computational overhead while enabling distribution-based uncertainty estimation.

Table 8: Comparison of different prior–posterior families in W-GBDL. Results are reported as mean $\pm$ standard deviation over five runs. The best result in each column is highlighted in bold.

Prior–posterior family	Cls Acc ↑	OOD Detect ↑	ECE ↓	NLL ↓
Gaussian–Gaussian	92.70 ± 0.17	80.58 ± 0.49	2.33 ± 0.39	$\mathbf{0.23 \pm 0.01}$
Gamma–Weibull (W-GBDL)	$\mathbf{93.11 \pm 0.21}$	$\mathbf{86.24 \pm 0.52}$	$\mathbf{1.31 \pm 0.38}$	$\mathbf{0.23 \pm 0.01}$

Effect of Prior–Posterior Parameterization. As shown in Table 8, the Gamma–Weibull parameterization consistently outperforms the Gaussian–Gaussian alternative across all reported metrics. It achieves higher classification accuracy, OOD detection performance, while obtaining competitive ECE and NLL. This consistent improvement suggests that the combination of a sparse Gamma prior and a positive, skewed Weibull variational posterior is better suited to the non-negative latent factors used by W-GBDL. The benefit is therefore not limited to a single evaluation criterion, but appears simultaneously in predictive performance, uncertainty estimation, and calibration.