Uncertainty Estimation by Flexible Evidential Deep Learning

Taeseong Yoon **Heeyoung Kim**
Department of Industrial and Systems Engineering, KAIST
{bigstar0423, heeyoungkim}@kaist.ac.kr

Abstract

Uncertainty quantification (UQ) is crucial for deploying machine learning models in high-stakes applications, where overconfident predictions can lead to serious consequences. An effective UQ method must balance computational efficiency with the ability to generalize across diverse scenarios. Evidential deep learning (EDL) achieves efficiency by modeling uncertainty through the prediction of a Dirichlet distribution over class probabilities. However, the restrictive assumption of Dirichlet-distributed class probabilities limits EDL's robustness, particularly in complex or unforeseen situations. To address this, we propose *flexible evidential deep learning* ($\mathcal{F}$-EDL), which extends EDL by predicting a flexible Dirichlet distribution—a generalization of the Dirichlet distribution—over class probabilities. This approach provides a more expressive and adaptive representation of uncertainty, significantly enhancing UQ generalization and reliability under challenging scenarios. We theoretically establish several advantages of $\mathcal{F}$-EDL and empirically demonstrate its state-of-the-art UQ performance across diverse evaluation settings, including classical, long-tailed, and noisy in-distribution scenarios.

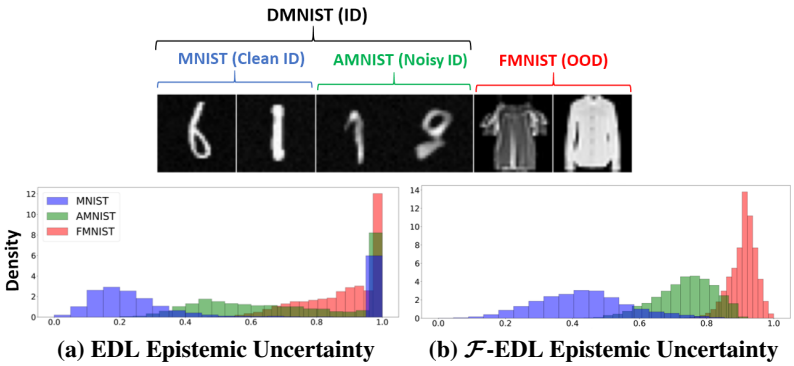
1 Introduction

Machine learning models have achieved remarkable predictive performance in diverse fields, including computer vision and natural language processing. However, their deployment in high-stakes applications—autonomous driving, medical diagnosis, and manufacturing—remains limited due to concerns about reliability and overconfident predictions [1]. Robust uncertainty quantification (UQ) is essential for ensuring safer, more trustworthy decision-making in such contexts.

Effective UQ methods must meet two fundamental requirements: (i) computational efficiency, enabling integration into real-time systems, and (ii) generalizability to diverse and unforeseen scenarios. Classical UQ methods—Bayesian neural networks [2], Monte Carlo dropout [3], and deep ensembles [4]—are well established but computationally intensive, as they require multiple forward passes. In response, single forward pass UQ methods [5, 6, 7] have emerged as efficient alternatives, among which evidential deep learning (EDL) [8, 9, 10, 11] stands out. EDL predicts a Dirichlet distribution over class probabilities to quantify uncertainty, leveraging the conjugate prior structure of the Dirichlet distribution and its closed-form uncertainty measures. This approach enables computationally efficient UQ and demonstrates strong performance in downstream tasks, such as out-of-distribution (OOD) detection [11].

Despite these advantages, EDL may struggle to provide robust uncertainty estimates in complex or unforeseen scenarios, leading to suboptimal UQ performance. This is shown in Figure 1(a), which illustrates the epistemic uncertainty distributions for EDL trained on the Dirty-MNIST (DMNIST) dataset [7]. DMNIST combines clean MNIST (blue), representing clean in-distribution (ID) samples, with Ambiguous-MNIST (AMNIST, green), a noisy ID dataset containing ambiguous samples. Additionally, Fashion-MNIST (FMNIST, red) serves as an OOD dataset. To ensure a fair compar-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) EDL Epistemic Uncertainty **(b) $\mathcal{F}$-EDL Epistemic Uncertainty**

Figure 1: Epistemic uncertainty distributions with DMNIST as the ID dataset. The top row presents sample images from three datasets: MNIST, AMNIST, and FMNIST. Panels (a) and (b) display histograms depicting the epistemic uncertainty distributions obtained by the EDL and $\mathcal{F}$-EDL (proposed) models, respectively, across these datasets.

ison, we adopt a standard uncertainty metric used in EDL [8]. An ideal UQ model should assign progressively higher epistemic uncertainty from clean ID to noisy ID and OOD samples, as intended in the DMNIST benchmark [7]. However, EDL exhibits substantial overlap between noisy ID and OOD distributions, and even assigns high uncertainty to some clean ID samples, indicating limited generalization in noisy ID scenarios. In contrast, our proposed model, described below, achieves the desired separation among these datasets, as shown in Figure 1(b), demonstrating robust UQ and improved generalization under noisy conditions. We hypothesize that EDL's limitation stems from its core assumption that the class probabilities for each data point follow a Dirichlet distribution. While this assumption facilitates computational efficiency, it restricts the model's ability to express complex or ambiguous forms of uncertainty. This motivates the need for more flexible yet efficient UQ methods.

Building upon this insight, we propose *flexible evidential deep learning* ($\mathcal{F}$-EDL), a novel UQ framework that extends EDL by predicting a flexible Dirichlet (FD) distribution [12] over class probabilities. The FD distribution generalizes the Dirichlet, enabling a more expressive representation of uncertainty while retaining the computational efficiency of EDL through its conjugate prior—allowing a single forward pass UQ suitable for real-time applications. This expressiveness improves the model's generalizability, enabling robust UQ across diverse and challenging scenarios such as noisy ID data, tail classes, and distribution shifts. Beyond the FD-based structure, $\mathcal{F}$-EDL introduces a tailored objective function for uncertainty-aware classification and label-wise variance-based uncertainty measures for robust, generalizable UQ.

$\mathcal{F}$-EDL introduces several theoretical advancements that strengthen its UQ capabilities. First, we prove that the FD distribution serves as a conjugate prior to the categorical likelihood. Building on this, we prove that $\mathcal{F}$-EDL predicts the posterior FD distribution for each input by employing an improper prior with parameters learned from the data, addressing the limitations of fixed priors in traditional EDL. Second, we prove that $\mathcal{F}$-EDL converges to standard EDL under specific parameter settings, establishing it as a generalization of EDL. Third, we prove that $\mathcal{F}$-EDL can generate multimodal class probability distributions on the simplex, allowing it to capture complex and nuanced uncertainty patterns. Fourth, we prove that $\mathcal{F}$-EDL functions as a mixture model that adaptively combines EDL and softmax predictions using input-dependent mixture weights, demonstrating how it leverages the strengths of both approaches to achieve optimal predictions and enhanced UQ. Finally, we prove that $\mathcal{F}$-EDL admits a *generalized subjective logic* interpretation, modeling uncertainty as a structured mixture of class-specific opinions that capture the model's hesitation among plausible hypotheses.

We empirically evaluate $\mathcal{F}$-EDL across a wide range of UQ-related downstream tasks, including classification, misclassification detection, OOD detection, and distribution shift detection. Notably, $\mathcal{F}$-EDL consistently achieves state-of-the-art performance across diverse settings, including classical, long-tailed, and noisy ID scenarios, highlighting its robustness and generalizability. In addition, qualitative analyses show that $\mathcal{F}$-EDL captures interpretable multimodal uncertainty reflecting ambiguity across plausible classes, while demonstrating faithful epistemic behavior that decreases with more data.

2 Preliminaries

2.1 Evidential Deep Learning

Evidential deep learning (EDL) quantifies uncertainty in classification tasks by predicting a Dirichlet distribution over class probabilities. Introduced by Sensoy et al. [8], EDL is grounded in Dempster-Shafer Theory of Evidence (DST) [13] and subjective logic (SL) [14]. DST represents subjective beliefs by assigning belief masses to classes, capturing both class likelihoods and uncertainty. SL extends DST by formalizing these subjective beliefs using a Dirichlet distribution—enabling uncertainty-aware classification within a principled framework.

The core innovation of Sensoy et al. [8] is using a neural network to predict Dirichlet parameters. Let D denote the input dimensionality and K denote the number of classes. For a given input $\mathbf{x} \in \mathbb{R}^D$, the class probability distribution is represented as $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where the concentration parameters $\boldsymbol{\alpha} \in \mathbb{R}^K$ are defined as $\boldsymbol{\alpha} = \mathbf{1} + \mathrm{ReLU}(f_{\boldsymbol{\theta}}(\mathbf{x}))$. Here, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^K$ is a vector of ones, and $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^K$ represents the neural network parameterized by $\boldsymbol{\theta}$. The objective function of EDL for a given data point $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ is the one-hot encoded label, consists of two components: the expected mean squared error (MSE) and a Kullback-Leibler (KL) divergence penalty. It is formulated as follows:

$$\mathcal{L} = \underbrace{\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\|\mathbf{y} - \boldsymbol{\pi}\|_2^2]}_{\text{Expected MSE over Dirichlet}} + \underbrace{\lambda \mathrm{D}_{\mathrm{KL}}[\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}) || \,\mathrm{Dir}(\mathbf{1})]}_{\text{KL divergence penalty}},$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \odot (\mathbf{1} - \mathbf{y}) + \mathbf{y}$, and λ is a regularization parameter.

2.2 Flexible Dirichlet Distribution

The flexible Dirichlet (FD) distribution [12] is a generalization of the Dirichlet distribution. The FD distribution is derived by normalizing a flexible Gamma (FG) basis. Specifically, the FG basis, denoted as $\mathbf{Y} = (Y_1, \ldots, Y_K)$, is constructed as follows:

$$Y_k = W_k + Z_k\, U,\ \forall k \in [K],$$

where $W_k \sim \mathrm{Gamma}(\alpha_k), \forall k \in [K]$, are independent Gamma random variables, and $U \sim \mathrm{Gamma}(\tau)$ is another independent Gamma random variable sharing the same scale parameter as each W_k. Here, the parameters satisfy $\alpha_k > 0, \forall k \in [K]$ and $\tau > 0$. The vector $\mathbf{Z} = (Z_1, \ldots, Z_K)$ follows a multinomial distribution, $\mathbf{Z} \sim \mathrm{Mu}(1, \mathbf{p})$, where $\mathbf{p} = (p_1, \ldots, p_K)$, $0 \le p_k \le 1, \forall k \in [K]$, and $\sum_{k=1}^K p_k = 1$. This FG basis is denoted as $\mathbf{Y} \sim \mathrm{FG}^K(\boldsymbol{\alpha}, \mathbf{p}, \tau)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. By combining independent Gamma random variables with a shared random component, the FG basis introduces flexibility, effectively modeling dependencies among components.

The FD distribution, denoted as $\mathbf{X} \sim \mathrm{FD}^K(\boldsymbol{\alpha}, \mathbf{p}, \tau)$ [1] , is defined as $\mathbf{X} = (X_1, \ldots, X_K)$ with $X_k = Y_k / \sum_{k=1}^K Y_k, \forall k \in [K]$. Due to its ability to capture complex dependencies and multimodal behavior, the FD distribution is well-suited for compositional data analysis [15, 12]. While previously unexplored for UQ, its flexibility makes it a compelling candidate for modeling uncertainty.

3 Flexible Evidential Deep Learning

The $\mathcal{F}$-EDL framework introduces an efficient and generalizable UQ methodology by leveraging the FD distribution to model a distribution over class probabilities. The adoption of the FD distribution in $\mathcal{F}$-EDL represents a principled generalization of EDL, rather than an ad-hoc extension of model flexibility. By introducing additional parameters $(\mathbf{p}, \tau)$, the model governs how evidence is allocated across classes (via $\mathbf{p}$) and concentrated in magnitude (via τ), thereby enabling structured and interpretable uncertainty modeling—particularly for ambiguous inputs where multiple class hypotheses may be simultaneously plausible. Furthermore, the FD-based formulation offers an input-adaptive mechanism for evidence extraction, capturing how uncertainty evolves dynamically in response to the characteristics of each input. A detailed theoretical discussion is provided in Appendix D.

To operationalize this formulation within a neural framework, $\mathcal{F}$-EDL consists of three key components: (i) a distinctive model structure (Section 3.1), (ii) a tailored objective function (Section 3.2), and (iii) label-wise variance-based uncertainty measures (Section 3.3).

[1] We refer to this distribution as FD, omitting the superscript when the dimensionality K is clear from the context.

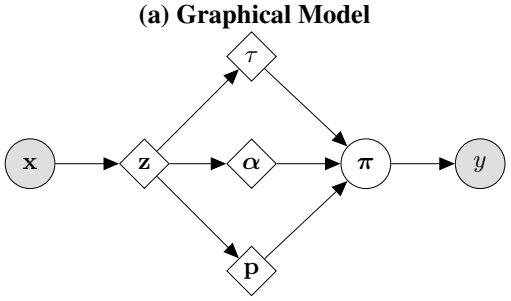


Figure 2: (a) Graphical model and (b) generative process of the proposed $\mathcal{F}$-EDL framework.

3.1 Model Structure

Let $\mathbf{x} \in \mathbb{R}^D$ denote an input with dimension D, and let $\mathbf{z} = f_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathbb{R}^H$ denote its feature representation, where $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^H$ is a feature extractor parameterized by $\boldsymbol{\theta}$ and H is the feature dimension. From $\mathbf{z}$, the model predicts the parameters of the FD distribution: concentration parameters $\boldsymbol{\alpha} \in \mathbb{R}_+^K$, allocation probabilities $\mathbf{p} \in \Delta^{K-1}$, and dispersion $\tau \in \mathbb{R}_+$, which together define a distribution over class probabilities $\boldsymbol{\pi} \in \Delta^{K-1}$, from which the label y is sampled. These parameters are computed by three neural heads: $g_{\boldsymbol{\phi}_1} : \mathbb{R}^H \to \mathbb{R}^K$, $g_{\boldsymbol{\phi}_2} : \mathbb{R}^H \to \mathbb{R}^K$, and $g_{\boldsymbol{\phi}_3} : \mathbb{R}^H \to \mathbb{R}$, parameterized by $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$, and $\boldsymbol{\phi}_3$, respectively.

To ensure the non-negativity and stability of $\boldsymbol{\alpha}$, we use $\exp$ activation, following prior work [16, 11]. To map $\mathbf{p}$ onto the probability simplex, we use the $\mathrm{softmax}$ function. For τ, we use the $\mathrm{softplus}$ activation for smoother gradients and improved empirical performance. We also apply spectral normalization [17] to $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\phi}_1}$ to improve the robustness of $\boldsymbol{\alpha}$ and enhance the quality of UQ [5, 11]. This normalization further enforces Lipschitz continuity and constrains the magnitudes of network outputs, thereby implicitly regularizing $\boldsymbol{\alpha}$ without requiring explicit penalty terms.

The overall structure is illustrated in Figure 2, with the graphical model in panel (a) and the generative process in panel (b).

3.2 Objective Function

The $\mathcal{F}$-EDL framework is optimized using an objective function consisting of two components: (i) the expected MSE with respect to the FD distribution, following standard EDL practice for uncertainty-aware classification, and (ii) a Brier-score-based regularization term for $\mathbf{p}$, to promote input-dependent calibration and prevent degenerate solutions.

For a given data point $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ is the one-hot encoded representation of the label y, the objective function is defined as follows:

$$\mathcal{L} = \underbrace{\mathbb{E}_{\boldsymbol{\pi} \sim \text{FD}(\boldsymbol{\alpha}, \mathbf{p}, \tau)}[\|\mathbf{y} - \boldsymbol{\pi}\|_2^2]}_{\text{Expected MSE over FD}} + \underbrace{\|\mathbf{y} - \mathbf{p}\|_2^2}_{\text{Regularization term}} .$$

$\mathcal{F}$-EDL enables analytic training due to the closed-form moments of the FD distribution, eliminating the need for sampling. It also simplifies optimization by reducing sensitivity to hyperparameters, unlike recent EDL methods that require careful tuning [9, 10].

3.3 Label-Wise Variance-Based Uncertainty Measures

After training, the predicted label for a testing example $\mathbf{x}^\star$ is determined by assigning the class with the highest expected class probability under the FD distribution: $\hat{y}(\mathbf{x}^\star) = \mathrm{argmax}_{k \in [K]} \, \mathbb{E}_{\boldsymbol{\pi} \sim \text{FD}(\boldsymbol{\alpha}, \mathbf{p}, \tau)}[\pi_k]$, where $\boldsymbol{\alpha}$, $\mathbf{p}$, and τ are predicted parameters for $\mathbf{x}^\star$.

To quantify uncertainty, we adopt a label-wise variance-based approach [18], replacing the Dirichlet with the FD distribution, enabling robust and interpretable UQ consistent with established axioms [19]. For each class $k \in [K]$, the predictive uncertainty for $\mathbf{x}^\star$, denoted by $\text{TU}_k(\mathbf{x}^\star)$, is defined as the variance of the class label: $\text{TU}_k(\mathbf{x}^\star) = \text{Var}(y_k | \mathbf{x}^\star)$. This quantity is decomposed via the law of total variance into aleatoric and epistemic components, denoted by $\text{AU}_k(\mathbf{x}^\star)$ and $\text{EU}_k(\mathbf{x}^\star)$,

respectively:

$$\underbrace{\mathrm{Var}(y_k|\mathbf{x}^\star)}_{\mathrm{TU}_k(\mathbf{x}^\star)} = \underbrace{\mathbb{E}_{\boldsymbol{\pi}}[\mathrm{Var}(y_k|\boldsymbol{\pi},\mathbf{x}^\star)]}_{\mathrm{AU}_k(\mathbf{x}^\star)} + \underbrace{\mathrm{Var}_{\boldsymbol{\pi}}(\mathbb{E}[y_k|\boldsymbol{\pi},\mathbf{x}^\star])}_{\mathrm{EU}_k(\mathbf{x}^\star)}.$$

The total, aleatoric, and epistemic uncertainties for $\mathbf{x}^\star$ are computed by summing the respective class-wise uncertainties and substituting the moments of the FD distribution.

$$\mathrm{EU}(\mathbf{x}^\star) = \sum_{k=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\pi}}[\pi_k](1-\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])}{(\alpha_0+\tau+1)} + \frac{\tau^2 p_k(1-p_k)}{(\alpha_0+\tau)(\alpha_0+\tau+1)},$$

$$\mathrm{TU}(\mathbf{x}^\star) = 1 - \sum_{k=1}^{K}(\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2, \quad \mathrm{AU}(\mathbf{x}^\star) = \mathrm{TU}(\mathbf{x}^\star) - \mathrm{EU}(\mathbf{x}^\star),$$

where $\alpha_0 = \sum_{k=1}^{K}\alpha_k$ and $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = \frac{\alpha_k+\tau p_k}{\alpha_0+\tau}, \forall k \in [K]$.

Detailed derivations of the closed-form expressions for the objective function and uncertainty measures are presented in Appendix A, while the algorithms for $\mathcal{F}$-EDL are provided in Appendix B.

4 Theoretical Analysis

$\mathcal{F}$-EDL offers several theoretical advancements that enhance its UQ generalizability. First, we prove that the FD distribution is a conjugate prior to the categorical likelihood (Lemma 4.1). Building on this, we prove that the class probability distribution for a given input $\mathbf{x}$, derived from $\mathcal{F}$-EDL, corresponds to the posterior FD distribution for $\boldsymbol{\pi}$, obtained with an improper prior $\boldsymbol{\pi}(\mathbf{x}) \sim \mathrm{FD}(\mathbf{0}, \mathbf{p}_{\mathrm{prior}}(\mathbf{x}), \tau_{\mathrm{prior}}(\mathbf{x}))$, where the prior parameters $\mathbf{p}_{\mathrm{prior}}(\mathbf{x})$ and $\tau_{\mathrm{prior}}(\mathbf{x})$ are learned from the data (Theorem 4.2).[2] By introducing improper priors with input-dependent parameters, $\mathcal{F}$-EDL eliminates the need for manually specified priors required in traditional EDL.

Second, we prove that the class probability distribution for a given input derived from $\mathcal{F}$-EDL is equivalent to that of standard EDL under specific parameter settings: $\tau = 1$ and $p_k = \alpha_k / \sum_{k=1}^{K}\alpha_k$ for all $k \in [K]$ (Theorem 4.3). This establishes EDL as a special case of $\mathcal{F}$-EDL, demonstrating that our framework retains EDL's strengths while enabling greater expressiveness when needed.

Third, we prove that the class probability distribution for a testing example derived from $\mathcal{F}$-EDL forms a multimodal distribution over the simplex, represented as a mixture of Dirichlet distributions, with the number of distinct modes determined by the cardinality of the allocation probability vector $\mathbf{p}$, denoted as $\|\mathbf{p}\|_0$ (Theorem 4.4). This allows $\mathcal{F}$-EDL to capture complex, multimodal uncertainty—particularly useful for handling ambiguous inputs.

Fourth, we prove that the predictive distribution for a testing example derived from $\mathcal{F}$-EDL can be decomposed into contributions from EDL and softmax, with input-dependent mixture weights (Theorem 4.5). This demonstrates that $\mathcal{F}$-EDL functions as a mixture model of EDL and softmax, adaptively integrating their complementary strengths to achieve optimal performance in both prediction and UQ. Specifically, the EDL component tends to dominate for clean ID data, whereas for ambiguous or OOD inputs, the model interpolates between the two, maintaining robust performance across varying uncertainty levels.

Finally, we prove that $\mathcal{F}$-EDL admits a *generalized subjective logic* interpretation, modeling a mixture of K class-specific subjective opinions (i.e., hypotheses), each defined by a shared base evidence $\boldsymbol{\alpha}$ and dispersion parameter τ, and weighted by the hypothesis selection probabilities $\mathbf{p}$. [3] This perspective reveals that $\mathcal{F}$-EDL performs UQ via principled evidence aggregation grounded in subjective logic [14], rather than simply adding flexibility.

Lemma 4.1. ***(Conjugacy of FD Prior)*** *Let $y_i \sim \mathrm{Cat}(\boldsymbol{\pi}), \forall i \in [N]$, where $\boldsymbol{\pi} \sim \mathrm{FD}(\boldsymbol{\alpha}, \mathbf{p}, \tau)$. The posterior distribution of $\boldsymbol{\pi}$ given $\{y_i\}_{i=1}^N$ is also an FD distribution, $\boldsymbol{\pi}|\{y_i\}_{i=1}^N \sim \mathrm{FD}(\boldsymbol{\alpha}', \mathbf{p}, \tau)$, where $\boldsymbol{\alpha}' = (\alpha'_1, \ldots, \alpha'_K)$ and $\alpha'_k = \alpha_k + \sum_{i=1}^{N}\mathbf{1}_{\{y_i=k\}}, \forall k \in [K]$.*

Theorem 4.2. ***(Bayesian Interpretation of $\mathcal{F}$-EDL)*** *The class probability distribution for a given input $\mathbf{x}$, derived from $\mathcal{F}$-EDL, is equivalent to the posterior FD distribution of $\boldsymbol{\pi}$, with*

[2] See Appendix C.2 for a more detailed explanation of the Bayesian interpretation of $\mathcal{F}$-EDL.

[3] See Appendix C.3 for a more detailed explanation of the subjective logic interpretation of $\mathcal{F}$-EDL.

likelihood $\sum_{i=1}^{N} \mathbf{1}_{\{y_i=k\}} = (\exp(g_{\boldsymbol{\phi}_1}(f_{\boldsymbol{\theta}}(\mathbf{x}))))_k, \forall k \in [K]$*, and an improper prior* $\boldsymbol{\pi}(\mathbf{x}) \sim \mathrm{FD}(\mathbf{0}, \mathbf{p}_{\mathrm{prior}}(\mathbf{x}), \tau_{\mathrm{prior}}(\mathbf{x}))$*, where* $\mathbf{p}_{\mathrm{prior}}(\mathbf{x})$ *and* $\tau_{\mathrm{prior}}(\mathbf{x})$ *are input-dependent parameters learned from the data.*

Theorem 4.3. ***(Generalization of EDL)*** *The class probability distribution for a testing example* $\mathbf{x}^\star$*, derived from* $\mathcal{F}$*-EDL, is equivalent to the class probability distribution of EDL when* $\tau = 1$ *and* $p_k = \alpha_k / \sum_{k=1}^{K} \alpha_k, \forall k \in [K]$*.*

Theorem 4.4. ***(Multimodality of $\mathcal{F}$-EDL)*** [4] *The class probability distribution for a testing example* $\mathbf{x}^\star$*, derived from* $\mathcal{F}$*-EDL, is expressed as a mixture of Dirichlet distributions:*

$$p_{\mathcal{F}\text{-EDL}}(\boldsymbol{\pi}|\mathbf{x}^\star) = \sum_{k=1}^{K} p_k \operatorname{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \tau \mathbf{e}_k),$$

where $\mathbf{e}_k$ *represents the* k*-th standard basis vector in* $\mathbb{R}^K$*. The number of distinct modes is determined by the cardinality of* $\mathbf{p}$*, i.e.,* $\|\mathbf{p}\|_0$*.*

Theorem 4.5. ***(Predictive Distribution Decomposition)*** *Let* $p_{\mathrm{EDL}}(y|\mathbf{x}^\star)$ *and* $p_{\mathrm{SM}}(y|\mathbf{x}^\star)$ *denote the predictive distributions for a testing example* $\mathbf{x}^\star$*, derived from the EDL and softmax models, respectively. The predictive distribution for* $\mathbf{x}^\star$*, obtained by the* $\mathcal{F}$*-EDL framework, is decomposed into the contributions of the EDL and softmax models as follows:*

$$p_{\mathcal{F}\text{-EDL}}(y|\mathbf{x}^\star) = w_{\mathrm{EDL}}(\mathbf{x}^\star) \times p_{\mathrm{EDL}}(y|\mathbf{x}^\star) + w_{\mathrm{SM}}(\mathbf{x}^\star) \times p_{\mathrm{SM}}(y|\mathbf{x}^\star),$$

where $w_{\mathrm{EDL}}(\mathbf{x}^\star) = \alpha_0/(\alpha_0 + \tau)$ *and* $w_{\mathrm{SM}}(\mathbf{x}^\star) = \tau/(\alpha_0 + \tau)$ *are input-dependent mixture weights.*

Proposition 4.6. ***(Subjective Logic Interpretation of $\mathcal{F}$-EDL)*** *$\mathcal{F}$-EDL admits a generalized subjective logic interpretation, representing K class-specific subjective opinions (i.e., hypotheses) defined by shared base evidence $\boldsymbol{\alpha}$, dispersion parameter τ, and hypothesis selection probabilities $\boldsymbol{p}$.*

For the proofs and additional insights about the theorems, refer to Appendix C.

5 Related Works

EDL methods. EDL [20] refers to a family of UQ methods that predict a Dirichlet distribution over class probabilities to estimate uncertainty in classification. KL-PN [21] encourages concentrated Dirichlets for ID data and uniform Dirichlets for OOD data via KL divergence, while RKL-PN [22] uses reverse KL to mitigate issues with unintended multimodal distributions. However, both rely on OOD data, which is often unavailable in practice. To mitigate the reliance on OOD data, PostNet [23] and NatPN [24] leverage feature space density estimated using Normalizing Flow [25] to predict the posterior Dirichlet distribution. However, their performance heavily depends on accurate density estimation, which remains challenging in high-dimensional settings.

Building on EDL [8], several methods have attempted to address its limitations. $\mathcal{I}$-EDL [9] incorporates Fisher information to account for varying levels of data uncertainty. R-EDL [10] treats a prior parameter as a tunable hyperparameter and simplifies the loss by removing the variance term, thereby relaxing EDL's restrictive assumption—a motivation shared with our approach. DAEDL [11] uses feature space density during prediction with an alternative parameterization. However, all these methods remain constrained by the Dirichlet assumption, limiting their generalization to complex scenarios. Separately, distillation-based approaches [26, 27, 28] improve epistemic uncertainty estimation but typically require multiple forward passes, reducing their practicality. Beyond standard classification, EDL methods have also been extended to diverse tasks, including regression [29], domain adaptation [30], semantic segmentation [31, 32], calibration of large language models [33], and multi-view learning [34, 35, 36, 37]

On the other hand, recent studies have questioned the ability of EDL to represent epistemic uncertainty. Bengs et al. [38, 39] showed that its second-order loss is not a strictly proper scoring rule, leading to non-vanishing uncertainty. Jürgens et al. [40] found that regularized EDL enforces a fixed uncertainty budget, yielding relative rather than absolute uncertainty measures. Shen et al. [28] further unified these critiques, revealing that the EDL objective collapses to a sample-size-independent

[4] Adapted from Proposition 4.1 of [12].

Table 1: UQ-related downstream task results are reported using the CIFAR-10 and CIFAR-100 datasets as the ID dataset. "Test.Acc." denotes the test accuracy, "Conf." denotes the AUPR scores for misclassification detection based on aleatoric uncertainty, and "SVHN / C-100" and "SVHN / TIN" indicate AUPR scores for OOD detection based on epistemic uncertainty, where the former corresponds to using CIFAR-10 as the ID dataset with SVHN and CIFAR-100 as the OOD datasets, and the latter corresponds to using CIFAR-100 as the ID dataset with SVHN and TinyImageNet as the OOD datasets. Baseline results for CIFAR-10 are sourced from existing literature [9, 10, 11]. Bold numbers indicate the best performance for each metric.

	ID: CIFAR-10			ID: CIFAR-100		
Method	Test.Acc.	Conf.	SVHN / C-100	Test.Acc.	Conf.	SVHN / TIN
Dropout	82.84 ±0.1	97.15 ±0.0	51.39 ±0.1 / 45.57 ±1.0	65.94 ±0.6	92.00 ±0.3	71.83 ±2.0 / 74.93 ±0.6
EDL	83.55 ±0.6	97.86 ±0.2	79.12 ±3.7 / 84.18 ±0.7	45.91 ±5.6	91.28 ±0.8	56.21 ±3.1 / 70.13 ±2.0
$\mathcal{I}$-EDL	89.20 ±0.3	98.72 ±0.1	82.96 ±2.2 / 84.84 ±0.6	66.38 ±0.5	92.84 ±0.1	67.51 ±2.9 / 75.86 ±0.3
R-EDL	90.09 ±0.3	98.98 ±0.1	85.00 ±1.2 / 87.73 ±0.3	63.53 ±0.5	92.69 ±0.2	61.80 ±3.4 / 69.78 ±1.3
DAEDL	91.11 ±0.2	99.08 ±0.0	85.54 ±1.4 / 88.19 ±0.1	66.01 ±2.6	86.00 ±0.3	72.07 ±4.1 / 77.40 ±1.6
$\mathcal{F}$-EDL	**91.19** ±0.2	**99.10** ±0.0	**91.20** ±1.3 / **88.37** ±0.3	**69.40** ±0.2	**94.01** ±0.1	**75.35** ±2.3 / **80.58** ±0.2

Table 2: AUPR scores for distribution shift detection from CIFAR-10 to CIFAR-10-C using aleatoric uncertainty estimates. $\mathcal{C} \in \{1, 2, 3, 4, 5\}$ denotes severity levels of corruption in CIFAR-10-C, averaged across 19 corruption types. Baseline results are sourced from [11].

Method	$\mathcal{C} = 1$	$\mathcal{C} = 2$	$\mathcal{C} = 3$	$\mathcal{C} = 4$	$\mathcal{C} = 5$
MSP	56.39 ±0.7	61.88 ±1.1	65.86 ±1.3	69.91 ±1.5	75.01 ±1.8
EDL	54.76 ±0.3	59.01 ±0.4	62.46 ±0.5	65.87 ±0.6	70.21 ±0.8
$\mathcal{I}$-EDL	56.33 ±0.2	61.52 ±0.5	65.44 ±0.5	69.45 ±0.5	74.56 ±0.5
R-EDL	57.37 ±0.5	62.20 ±1.0	65.74 ±1.4	69.33 ±1.9	73.58 ±2.6
DAEDL	57.89 ±0.3	63.23 ±0.4	67.53 ±0.4	72.21 ±0.4	77.74 ±0.4
$\mathcal{F}$-EDL	**59.01** ±0.8	**65.11** ±0.7	**69.48** ±0.5	**73.88** ±0.3	**78.72** ±0.4

Dirichlet target, behaving more like energy-based OOD detectors than true uncertainty estimators. However, $\mathcal{F}$-EDL empirically mitigates these issues through a more expressive FD-based formulation and distinct loss design, yielding better-calibrated epistemic uncertainty in practice.

Deterministic uncertainty methods. Beyond EDL, deterministic uncertainty methods (DUMs) [31] offer computationally efficient UQ in a single forward pass. DUMs are designed with two primary objectives: (i) applying regularization techniques to learn feature representations that distinguish between ID and OOD data, and (ii) quantifying uncertainty based on these regularized representations. Regularization methods include spectral normalization [5, 41, 42, 7] and gradient penalties [6]. For uncertainty estimation, DUMs employ diverse strategies, including Gaussian processes [5, 41], radial basis function networks [6], Gaussian discriminant analysis [7], and kernel density estimation [42]. However, these methods often require significant modifications to the neural network architecture, hindering their practical applicability.

6 Experiments

We conducted comprehensive experiments to evaluate the generalizability and effectiveness of $\mathcal{F}$-EDL in UQ. Our evaluation consists of two parts. First, we performed a quantitative study (Section 6.2) across diverse ID scenarios—including classical, long-tailed, and noisy settings—focusing on UQ-related downstream tasks. We further conducted an ablation study to assess the contribution of each FD parameter. Second, we performed a qualitative analysis (Section 6.3) examining whether $\mathcal{F}$-EDL produces semantically meaningful uncertainty representations for ambiguous samples and whether its epistemic uncertainty decreases with increasing training data, as theoretically expected. The code for our model is publicly available at https://github.com/TaeseongYoon/F-EDL.

6.1 Overview of Experiments

Tasks. For each scenario, we conducted three primary UQ-related downstream tasks: (i) classification, (ii) misclassification detection, and (iii) OOD detection. In the classical setting, we additionally included distribution shift detection. Classification performance was assessed using test accuracy. Misclassification detection was evaluated using the area under the precision-recall curve (AUPR)

Table 3: UQ-related downstream task results are reported using the CIFAR-10-LT dataset as the ID dataset under both heavily ($\rho = 0.01$) and mildly ($\rho = 0.1$) imbalanced settings.

	Heavy Imbalance ($\rho = 0.01$)			Mild Imbalance ($\rho = 0.1$)		
Method	Test.Acc.	Conf.	SVHN / C-100	Test.Acc.	Conf.	SVHN / C-100
Dropout	39.22 ±3.1	63.62 ±2.7	33.33 ±1.7 / 54.17 ±1.1	70.87 ±3.0	89.82 ±2.3	37.37 ±1.4 / 61.18 ±1.3
EDL	42.62 ±2.7	82.63 ±1.7	51.99 ±3.8 / 66.86 ±0.9	79.09 ±0.4	95.36 ±0.1	72.18 ±2.1 / 80.09 ±0.7
$\mathcal{I}$-EDL	57.88 ±1.3	84.10 ±1.3	52.85 ±6.8 / 69.19 ±1.3	84.86 ±0.1	97.31 ±0.2	79.83 ±3.9 / 83.50 ±0.4
R-EDL	63.36 ±1.0	78.34 ±1.0	48.71 ±7.1 / 64.20 ±1.4	85.35 ±0.2	94.35 ±0.2	60.58 ±5.0 / 69.53 ±1.6
DAEDL	63.36 ±1.4	82.15 ±1.0	51.03 ±5.6 / 65.31 ±1.2	84.95 ±0.4	95.22 ±0.4	69.40 ±4.5 / 74.56 ±1.7
$\mathcal{F}$-EDL	**63.73** ±1.4	**85.99** ±1.7	**62.56** ±2.8 / **70.18** ±2.0	**85.46** ±0.2	**97.60** ±0.1	**85.36** ±1.5 / **83.64** ±0.7

Table 4: UQ-related downstream task results are reported using the DMNIST dataset as the ID dataset. "FMNIST" denotes AUPR scores for OOD detection based on epistemic uncertainty, with FMNIST as the OOD dataset.

Method	Test.Acc.	Conf.	FMNIST
MSP	83.90 ±0.1	96.01 ±0.0	98.10 ±0.3
Dropout	84.13 ±0.1	96.09 ±0.0	94.68 ±0.6
DDU	84.05 ±0.1	82.73 ±0.1	98.49 ±0.4
EDL	77.37 ±5.8	95.19 ±0.3	92.23 ±1.7
$\mathcal{I}$-EDL	83.46 ±0.2	95.58 ±0.0	94.11 ±0.9
R-EDL	83.41 ±0.1	95.58 ±0.1	90.91 ±5.6
DAEDL	84.12 ±0.1	95.93 ±0.0	99.44 ±0.2
$\mathcal{F}$-EDL	**84.28** ±0.1	**96.17** ±0.1	**99.76** ±0.1

Table 5: Ablation study on $\mathcal{F}$-EDL using the DMNIST dataset as the ID dataset. "Fix-$\mathbf{p}$ (U), τ" and "Fix-$\mathbf{p}$ (N), τ" fix $\mathbf{p}$ to a uniform vector or normalized concentration parameters, respectively, and fix $\tau = 1$. "Fix-$\mathbf{p}$ (U)" and "Fix-$\mathbf{p}$ (N)" fix only $\mathbf{p}$; "Fix-τ" fixes only τ. The full model learns both.

Variant	Test.Acc.	Conf.	FMNIST
Fix-$\mathbf{p}$ (U), τ	83.34 ±0.2	95.62 ±0.1	97.22 ±1.0
Fix-$\mathbf{p}$ (N), τ	83.27 ±0.1	95.59 ±0.4	97.91 ±1.3
Fix-$\mathbf{p}$ (U)	83.39 ±0.1	95.60 ±0.1	96.94 ±1.1
Fix-$\mathbf{p}$ (N)	83.32 ±0.1	95.61 ±0.3	97.48 ±1.8
Fix-τ	83.39 ±0.1	95.60 ±0.1	98.46 ±0.1
$\mathcal{F}$-EDL	**84.28** ±0.1	**96.17** ±0.1	**99.76** ±0.1

score, with confidence defined as negative aleatoric uncertainty for ID samples (label: correct = 1, incorrect = 0). OOD detection was also evaluated using AUPR, based on negative epistemic uncertainty (labels: ID = 1, OOD = 0), assessing whether the model assigns higher uncertainty to OOD examples. Distribution shift detection was formulated as OOD detection by applying varying levels of corruption to a base dataset to simulate shifts. All AUPR scores were normalized to a 0-100 scale.

Datasets. In the classical setting, CIFAR-10 and CIFAR-100 [43] were used as the primary ID datasets. For the long-tailed setting, we used CIFAR-10-LT [44], an artificially imbalanced version of CIFAR-10, as the ID dataset. For the noisy setting, DMNIST, a variant of MNIST containing ambiguous data points, was used as the ID dataset. For OOD detection, SVHN [45] and CIFAR-100 served as OOD datasets when CIFAR-10 or CIFAR-10-LT was used as ID, while SVHN and TinyImageNet (TIN) were used as the OOD datasets for CIFAR-100. FMNIST was used as the OOD dataset for DMNIST. For distribution shift detection, we utilized CIFAR-10-C [46], a dataset created by applying continuous distribution shifts to CIFAR-10, as the OOD dataset.

Baselines. We compared $\mathcal{F}$-EDL against classical and state-of-the-art EDL methods, including EDL [8], $\mathcal{I}$-EDL [9], R-EDL [10], and DAEDL [11]. To highlight task difficulty, we also included Dropout [3] as a traditional UQ baseline. For distribution shift detection, we compared with MSP [46], a widely used benchmark. For the noisy setting, we evaluated both MSP and DDU [7], with DDU specifically introduced for DMNIST. All baselines were evaluated following the uncertainty estimation protocols described in their original papers.

Implementation. We adopted VGG-16 [47] for CIFAR-10 and CIFAR-10-LT, and ResNet-18 [48] for CIFAR-100. For DMNIST, we used a lightweight CNN with three convolutional and three dense layers. To predict the FD parameters $\mathbf{p}$ and τ, we added two shallow MLPs with 1-2 layers depending on the dataset, introducing minimal overhead (e.g., 1.8% for VGG-16). At inference time, $\mathcal{F}$-EDL remains highly efficient without any post-hoc processing, running only 1.3% slower than EDL yet over 50% faster than DAEDL on CIFAR-10.

For clarity and brevity, we defer additional experimental explanations (Appendix E), extended results (Appendix F), and supplementary qualitative analyses (Appendix G) to the appendix.

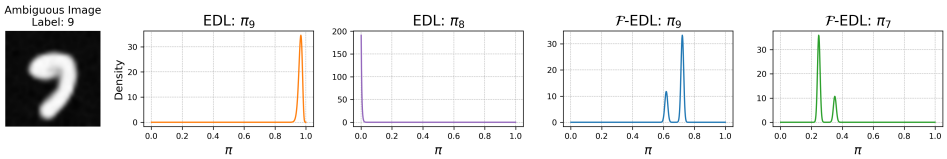


Figure 3: Posterior class probability distributions for a DMNIST input (true label: 9), which is visually ambiguous and resembles class 7. The leftmost panel shows the input image. The second and third panels display the marginal distributions of the two most probable classes predicted by EDL (denoted as π_k for class k), while the fourth and fifth panels show those predicted by $\mathcal{F}$-EDL.

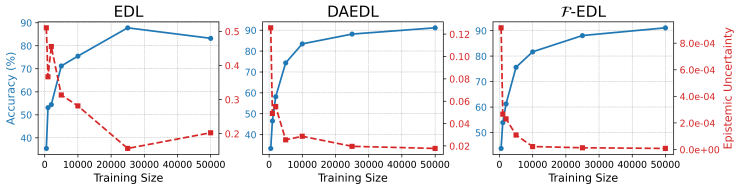


Figure 4: Test accuracy (blue, solid line) and epistemic uncertainty (red, dashed line) of EDL (left), DAEDL (middle), and $\mathcal{F}$-EDL (right) across increasing training set sizes: {500, 1000, 2000, 5000, 10000, 25000, 50000}. Accuracy is plotted on the left y-axis, and uncertainty is plotted on the right y-axis in each subplot.

6.2 Quantitative Study: UQ-Related Downstream Tasks across Diverse Settings

UQ in Classical Setting. We first evaluate $\mathcal{F}$-EDL in the classical setting, a standard benchmark for UQ models, using CIFAR-10 and CIFAR-100 across four UQ-related downstream tasks described in Section 6.1. As shown in Table 1 and Table 2, $\mathcal{F}$-EDL consistently outperforms competitive baselines across all tasks and datasets. These improvements in UQ performance are broadly attributed to two key innovations: (i) enhanced expressiveness enabled by generalizing beyond the Dirichlet family (Theorem 4.3), and (ii) the use of input-dependent improper priors that address the prior specification issue (Theorem 4.2). In addition, its strong performance in distribution shift detection reflects the model's ability to effectively combine the complementary strengths of softmax-based (MSP) and evidential (EDL) approaches to achieve optimal UQ (Theorem 4.5).

UQ in Long-Tailed Setting. To evaluate robustness under class imbalance, we further test $\mathcal{F}$-EDL on CIFAR-10-LT with two levels of imbalance: (i) mild ($\rho = 0.1$) and (ii) heavy ($\rho = 0.01$), across the three tasks described in Section 6.1. Effective UQ in long-tailed settings requires producing calibrated uncertainty estimates under class imbalance to detect misclassified examples and to distinguish rare ID inputs from OOD data. These capabilities were assessed through misclassification detection and OOD detection tasks, respectively. As shown in Table 3, $\mathcal{F}$-EDL demonstrates strong and consistent performance across all tasks and imbalance levels. This performance highlights $\mathcal{F}$-EDL's ability to provide reliable uncertainty estimates under complex distributional structures—a benefit attributed to its flexibility beyond the Dirichlet assumption (Theorem 4.3).

UQ in Noisy Setting. To evaluate robustness in noisy ID scenarios, we apply $\mathcal{F}$-EDL to the DMNIST dataset across the three tasks described in Section 6.1. Effective UQ in such settings requires leveraging aleatoric uncertainty to detect ambiguous inputs that are prone to misclassification and epistemic uncertainty to distinguish ID from OOD data. These capabilities were assessed via misclassification detection and OOD detection, respectively. As shown in Table 4, $\mathcal{F}$-EDL consistently achieves strong performance across all tasks. This performance highlights $\mathcal{F}$-EDL's improved ability to handle ambiguous inputs more effectively by representing multimodal opinions over plausible classes in a structured manner (Theorem 4.4 and Proposition 4.6).

Ablation Study: Contributions of $\mathbf{p}$ and τ. To assess the contribution of additional parameters introduced in the FD distribution, we conduct an ablation study in the noisy ID setting using the DMNIST dataset. Specifically, we compare the full $\mathcal{F}$-EDL model to several simplified variants: two configurations that fix $\mathbf{p}$—one to a uniform distribution ($\mathbf{p} = \mathbf{1}/K$) and another to the normalized concentration parameters ($\mathbf{p} = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_1$)—and another configuration that fixes $\tau = 1$. Combinations where both $\mathbf{p}$ and τ are fixed are also evaluated. As shown in Table 5, all fixed variants degrade in UQ performance, with "Fix-τ" showing favorable trade-offs, but no single variant consistently outperforms the others. The full model, which learns both $\mathbf{p}$ and τ, achieves the best results,

suggesting that the performance gains arise from the synergistic flexibility of the FD distribution, rather than merely increased capacity.

6.3 Qualitative Study: Multimodality and Faithful Epistemic Behavior

Multimodal Uncertainty Representations for Ambiguous Inputs. To assess whether the flexibility of $\mathcal{F}$-EDL leads to semantically meaningful uncertainty representations, we visualize class probability distributions for ambiguous inputs from the DMNIST dataset, on which the model was trained. For each test input, we extract the marginal distributions of the two most probable classes under both EDL and $\mathcal{F}$-EDL. As shown in Figure 3, EDL collapses the prediction into a single, overconfident mode, skewed toward one class, failing to capture the underlying uncertainty, with the second most probable class often misaligned with any visually plausible alternative. In contrast, $\mathcal{F}$-EDL produces multimodal distributions with distinct peaks near plausible classes—such as "7" and "9"—reflecting the model's hesitation in the presence of ambiguity. These qualitative improvements highlight $\mathcal{F}$-EDL's ability to express structured, multimodal representations, in line with its theoretical grounding (Theorem 4.4, Proposition 4.6).

Faithfulness of Epistemic Uncertainty Estimates. To evaluate whether $\mathcal{F}$-EDL provides faithful estimates of epistemic uncertainty, we examine how uncertainty evolves as the training dataset grows. A key characteristic of epistemic uncertainty is that it should decrease with more data, reflecting increased model confidence as knowledge improves [28]. Following the evaluation protocol proposed in [28], we train models on progressively larger subsets of CIFAR-10 and compute the average epistemic uncertainty over a fixed test set. As shown in Figure 4, $\mathcal{F}$-EDL demonstrates a clear, monotonic decrease in uncertainty with more data, unlike EDL and DAEDL, which show inconsistent trends. This suggests that $\mathcal{F}$-EDL successfully captures the theoretically desirable behavior wherein epistemic uncertainty consistently decreases as more data become available—empirically addressing a key limitation of standard EDL, whose epistemic uncertainty often fails to exhibit a reliable monotonic decline [28].

7 Conclusion

Summary. We propose $\mathcal{F}$-EDL, a novel UQ framework that maintains the efficiency of EDL while enhancing its generalizability by predicting a more expressive FD distribution over class probabilities. Theoretically, $\mathcal{F}$-EDL possesses several desirable properties that support reliable and robust UQ under complex conditions. Empirically, it achieves state-of-the-art performance across a variety of UQ-related downstream tasks in diverse and challenging ID scenarios, supported by qualitative evidence of interpretable multimodal predictions and faithful epistemic uncertainty.

Limitations & Future Directions. Despite its improved flexibility, $\mathcal{F}$-EDL faces several open challenges. First, it is currently limited to classification; extending it to regression, for instance, by building on evidential regression models [29], is a natural next step. Second, although $\mathcal{F}$-EDL provides a variance-based decomposition of uncertainty, it does not fully disentangle aleatoric and epistemic components—highlighting the need for further work on structured disentanglement, a longstanding challenge in UQ [49, 50]. Third, while $\mathcal{F}$-EDL empirically alleviates several theoretical limitations of EDL, it still relies on external regularization to control epistemic uncertainty [38], suggesting the need for an intrinsically stable training objective.

Beyond these aspects, $\mathcal{F}$-EDL opens multiple promising research avenues. First, it can serve as a foundation for developing UQ methods tailored to specific ID scenarios—such as long-tailed classification, where combining $\mathcal{F}$-EDL with logit adjustment [51] could improve calibration under imbalance. Second, its modular architecture supports plug-and-play integration into existing EDL-based frameworks, allowing the Dirichlet components to be replaced with the more expressive FD formulation in downstream applications—such as trusted multi-view learning [34]—through the design of compatible evidence-fusion mechanisms.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C2005453, RS-2023-00218913).

References

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

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

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

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

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

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

[7] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

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

[9] 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*, pages 7596–7616. PMLR, 2023.

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

[11] Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep learning. In *International Conference on Machine Learning*, pages 57217–57243. PMLR, 2024.

[12] Andrea Ongaro and Sonia Migliorati. A generalization of the dirichlet distribution. *Journal of Multivariate Analysis*, 114:412–426, 2013.

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

[14] Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

[15] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

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

[17] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[18] Ruxiao Duan, Brian Caffo, Harrison X Bai, Haris I Sair, and Craig Jones. Evidential uncertainty quantification: A variance-based perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2132–2141, 2024.

[19] Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[20] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023.

[21] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[22] Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[24] Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. In *International Conference on Learning Representations*, 2021.

[25] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[26] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

[27] Hanjing Wang and Qiang Ji. Beyond dirichlet-based models: when bayesian neural networks meet evidential deep learning. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[28] Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? *Advances in Neural Information Processing Systems*, 37:107830–107864, 2024.

[29] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[30] Liang Chen, Yihang Lou, Jianzhong He, Tao Bai, and Minghua Deng. Evidential neighborhood contrastive learning for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6258–6267, 2022.

[31] Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. In *International Conference on Machine Learning*, pages 17870–17909. PMLR, 2022.

[32] Jishnu Mukhoti, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.

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

[34] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

[35] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16129–16137, 2024.

[36] Xinyan Liang, Pinhan Fu, Yuhua Qian, Qian Guo, and Guoqing Liu. Trusted multi-view classification via evolutionary multi-view fusion. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Jueqing Lu, Wray Buntine, YUANYUAN QI, Joanna Dipnall, Belinda Gabbe, and Lan Du. Navigating conflicting views: Harnessing trust for learning. In *Forty-second International Conference on Machine Learning*, 2025.

[38] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35:29205–29216, 2022.

[39] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pages 2078–2091. PMLR, 2023.

[40] Mira Juergens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *International Conference on Machine Learning*, pages 22624–22642. PMLR, 2024.

[41] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

[42] Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323, 2022.

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

[44] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[45] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[46] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] 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*, pages 770–778, 2016.

[49] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 2024.

[50] Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Reexamining the aleatoric and epistemic uncertainty dichotomy. In *The Fourth Blogpost Track at ICLR 2025*, 2025.

[51] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[52] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

[54] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. `https://tiny-imagenet.herokuapp.com/`, 2015.

[55] 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*, pages 248–255. IEEE, 2009.

[56] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[57] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[58] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, page 6. San Diego, California;, 2015.

[59] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The complete proofs of all theorems are presented in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental settings and implementation details necessary to reproduce the main results are provided in Appendix E.2. Additionally, we include the full codebase and instructions in the supplementary material to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the full code, including scripts and detailed instructions, in the supplementary material. All datasets used in our experiments are publicly available. Upon acceptance, we will release the code as open-source on GitHub to facilitate broader accessibility and reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental settings and implementation details necessary for reproduction are provided in Appendix E.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean and standard deviation across five random trials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the relevant details about the computer resources at Appendix E.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and the research conducted in the paper conforms with it in every aspect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The primary goal of this paper is to develop a robust UQ method to enhance the reliability of machine learning systems, which could have positive societal impacts. We do not anticipate any negative societal impacts from our work.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We only use publicly available and standard image classification datasets, which doesn't have any risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used in the paper are all publicly available datasets. We cited the relevant papers for credit.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new asssets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were used solely for writing, editing, and formatting assistance. They did not contribute to the core methodology, experimental design, or scientific content of the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix for the Paper

"Uncertainty Estimation by Flexible Evidential Deep Learning"

A Additional Explanations about Objective Function and Uncertainty Measures

In this section, we present the closed-form representations for the formulas discussed in the main text. First, we derive the closed-form representation of the objective function for $\mathcal{F}$-EDL, as provided in Section 3.2. Next, we introduce the formulas for total, aleatoric, and epistemic uncertainty using variance-based uncertainty measures and derive the corresponding closed-form representations for $\mathcal{F}$-EDL, as outlined in Section 3.3.

A.1 Objective Function

Let $\mathbf{y} = (y_1, \ldots, y_K)$ denote the one-hot encoded labels, and let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ represent the class probabilities. Furthermore, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\mathbf{p} = (p_1, \ldots, p_K)$, and τ denote the parameters of the FD distribution. The objective function is defined as:

$$\mathcal{L} = \underbrace{\mathbb{E}_{\boldsymbol{\pi} \sim \text{FD}(\boldsymbol{\alpha}, \mathbf{p}, \tau)}[\|\mathbf{y} - \boldsymbol{\pi}\|_2^2]}_{\mathcal{L}^{\text{MSE}}} + \underbrace{\|\mathbf{y} - \mathbf{p}\|_2^2}_{\mathcal{L}^{\text{reg}}},$$

where the first term, $\mathcal{L}^{\text{MSE}}$, represents the expected MSE over the FD distribution, and the second term, $\mathcal{L}^{\text{reg}}$, is a Brier-score-based regularization term for $\mathbf{p}$.

Expected MSE over FD. In the first term, the expected MSE can be expressed as:

$$\mathcal{L}^{\text{MSE}} = \mathbb{E}_{\boldsymbol{\pi}}[\|\mathbf{y} - \boldsymbol{\pi}\|_2^2].$$

Expanding the squared difference gives:

$$\mathcal{L}^{\text{MSE}} = \|\mathbf{y}\|_2^2 - 2(\mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{\pi}])^T \mathbf{y} + \mathbb{E}_{\boldsymbol{\pi}}[\|\boldsymbol{\pi}\|_2^2].$$

Rewriting this as a summation over k:

$$\mathcal{L}^{\text{MSE}} = \sum_{k=1}^{K} \left(y_k^2 - 2\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] y_k + \mathbb{E}_{\boldsymbol{\pi}}[\pi_k^2] \right).$$

Using the relationship $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k^2] = (\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2 + \text{Var}_{\boldsymbol{\pi}}(\pi_k)$, this simplifies to:

$$\mathcal{L}^{\text{MSE}} = \sum_{k=1}^{K} (y_k - \mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2 + \sum_{k=1}^{K} \text{Var}_{\boldsymbol{\pi}}(\pi_k).$$

The mean of the FD distribution, $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k]$, is given by:

$$\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = \frac{\alpha_k + \tau p_k}{\alpha_0 + \tau}, \forall k \in [K],$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

The variance of the FD distribution, $\text{Var}_{\boldsymbol{\pi}}[\pi_k]$, is given by:

$$\text{Var}_{\boldsymbol{\pi}}(\pi_k) = \frac{(\alpha_k + \tau p_k)(\alpha_0 - \alpha_k + \tau(1 - p_k))}{(\alpha_0 + \tau)^2(\alpha_0 + \tau + 1)} + \frac{\tau^2 p_k (1 - p_k)}{(\alpha_0 + \tau)(\alpha_0 + \tau + 1)}, \forall k \in [K].$$

By substituting the mean and variance of the FD distribution, $\mathcal{L}^{\text{MSE}}$ becomes:

$$\mathcal{L}^{\text{MSE}} = \sum_{k=1}^{K} \left(y_k - \frac{\alpha_k + p_k \tau}{\alpha_0 + \tau} \right)^2 + \sum_{k=1}^{K} \frac{(\alpha_k + \tau p_k)(\alpha_0 - \alpha_k + \tau(1 - p_k))}{(\alpha_0 + \tau)^2(\alpha_0 + \tau + 1)} + \frac{\tau^2 p_k (1 - p_k)}{(\alpha_0 + \tau)(\alpha_0 + \tau + 1)}.$$

Regularization Term. The second term, representing the regularization term for $\mathbf{p}$, is given by:

$$\mathcal{L}^{\text{reg}} = \|\mathbf{y} - \mathbf{p}\|_2^2 = \sum_{k=1}^{K} (y_k - p_k)^2.$$

Combined Objective Function. Combining the two terms, the complete objective function becomes:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}^{\text{MSE}} + \mathcal{L}^{\text{reg}} \\ &= \sum_{k=1}^{K}\left(y_k - \frac{\alpha_k + p_k\tau}{\alpha_0+\tau}\right)^2 + \sum_{k=1}^{K}\frac{(\alpha_k+\tau p_k)(\alpha_0-\alpha_k+\tau(1-p_k))}{(\alpha_0+\tau)^2(\alpha_0+\tau+1)} + \frac{\tau^2 p_k(1-p_k)}{(\alpha_0+\tau)(\alpha_0+\tau+1)} \\ &+ \sum_{k=1}^{K}(y_k-p_k)^2.\end{aligned}$$

A.2 Uncertainty Measures

The closed-form expressions for the total, aleatoric, and epistemic uncertainties for a testing example $\mathbf{x}^\star$ can be derived using the variance-based uncertainty measures and the moments of the FD distribution. Below, we present the derivations for these uncertainty measures.

Total Uncertainty. The predictive uncertainty for each class $k \in [K]$, for a testing example $\mathbf{x}^\star$, denoted as $\text{TU}_k(\mathbf{x}^\star)$, is defined as the variance of the label for that class:

$$\text{TU}_k(\mathbf{x}^\star) = \text{Var}(y_k|\mathbf{x}^\star).$$

The total uncertainty for $\mathbf{x}^\star$, denoted as $\text{TU}(\mathbf{x}^\star)$, is obtained by summing the class-wise uncertainties:

$$\text{TU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\text{Var}(y_k|\mathbf{x}^\star).$$

Using the law of total variance, this can be decomposed as:

$$\text{TU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}[\text{Var}(y_k|\boldsymbol{\pi},\mathbf{x}^\star)] + \sum_{k=1}^{K}\text{Var}_{\boldsymbol{\pi}}(\mathbb{E}[y_k|\boldsymbol{\pi},\mathbf{x}^\star]).$$

Since $y_k \sim \text{Ber}(\pi_k), \forall k \in [K]$, we have:

$$\text{TU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}[\pi_k(1-\pi_k)] + \sum_{k=1}^{K}\text{Var}_{\boldsymbol{\pi}}(\pi_k).$$

Expanding the terms:

$$\text{TU}(\mathbf{x}^\star) = \sum_{k=1}^{K}(\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] - \mathbb{E}_{\boldsymbol{\pi}}[\pi_k^2]) + \sum_{k=1}^{K}(\mathbb{E}_{\boldsymbol{\pi}}[\pi_k^2] - (\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2).$$

Simplifying:

$$\text{TU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] - (\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2.$$

Since $\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = 1$, the total uncertainty simplifies to:

$$\text{TU}(\mathbf{x}^\star) = 1 - \sum_{k=1}^{K}(\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2,$$

where $\mathbb{E}_{\boldsymbol{\pi}}(\pi_k) = \frac{\alpha_k+\tau p_k}{\alpha_0+\tau}, \forall k \in [K]$.

Epistemic Uncertainty. The epistemic uncertainty for a testing example $\mathbf{x}^\star$, denoted as $\text{EU}(\mathbf{x}^\star)$, can be expressed as:

$$\text{EU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\text{Var}_{\boldsymbol{\pi}}(\pi_k).$$

Substituting the expression for the variance of the FD distribution, we have:

$$\text{EU}(\mathbf{x}^\star) = \sum_{k=1}^{K}\frac{\mathbb{E}_{\boldsymbol{\pi}}[\pi_k](1-\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])}{\alpha_0+\tau+1} + \frac{\tau^2 p_k(1-p_k)}{(\alpha_0+\tau)(\alpha_0+\tau+1)},$$

where $\alpha_0 = \sum_{k=1}^{K}\alpha_k$ and $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = \frac{\alpha_k+\tau p_k}{\alpha_0+\tau}, \forall k \in [K]$.

Aleatoric Uncertainty. The aleatoric uncertainty for a testing example $\mathbf{x}^\star$, denoted as $\mathrm{AU}(\mathbf{x}^\star)$, is obtained by subtracting the epistemic uncertainty from the total uncertainty:

$$\mathrm{AU}(\mathbf{x}^\star) = \mathrm{TU}(\mathbf{x}^\star) - \mathrm{EU}(\mathbf{x}^\star).$$

B Algorithm

We present the training algorithm and the prediction and UQ procedure for $\mathcal{F}$-EDL in Algorithm 1 and Algorithm 2, respectively. These algorithms are straightforward to implement, requiring no sensitive hyperparameter tuning or reliance on additional techniques.

Algorithm 1 $\mathcal{F}$-EDL Training

Input: Training data $\mathcal{D}_{\mathrm{tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, initial model parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3\}$, maximum epoch $T_{\max}$, learning rate η
for $i = 1$ **to** $T_{\max}$ **do**
Update the parameters using the Adam optimizer and $\mathcal{F}$-EDL objective function:

$$\big\{\boldsymbol{\theta} = \{W_{\boldsymbol{\theta}}^{(l)}, b_{\boldsymbol{\theta}}^{(l)}\}_{l=1}^{L_1},\ \boldsymbol{\phi}_1 = \{W_{\boldsymbol{\phi}_1}^{(l)}, b_{\boldsymbol{\phi}_1}^{(l)}\}_{l=1}^{L_2}, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3\big\} \leftarrow \mathrm{Adam}(\nabla \mathcal{L}^{\mathcal{F}\text{-}\mathrm{EDL}}, \eta).$$

Apply spectral normalization to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_1$:

$$\{W_{\boldsymbol{\theta}}^{(l)}\}_{l=1}^{L_1}, \{W_{\boldsymbol{\phi}_1}^{(l)}\}_{l=1}^{L_2} \leftarrow \mathrm{SpectNorm}(\{W_{\boldsymbol{\theta}}^{(l)}\}_{l=1}^{L_1}, \{W_{\boldsymbol{\phi}_1}^{(l)}\}_{l=1}^{L_2}).$$

end for
Output: Trained model parameters $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, \hat{\boldsymbol{\phi}}_3\}$

Algorithm 2 $\mathcal{F}$-EDL Prediction and Uncertainty Quantification

Input: Testing example $\mathbf{x}^\star$ and trained model parameters $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, \hat{\boldsymbol{\phi}}_3\}$
Step 1: Compute FD Parameters:

$$\boldsymbol{\alpha}(\mathbf{x}^\star) = \exp\left(g_{\hat{\boldsymbol{\phi}}_1}(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}^\star))\right),$$
$$\mathbf{p}(\mathbf{x}^\star) = \mathrm{softmax}\left(g_{\hat{\boldsymbol{\phi}}_2}(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}^\star))\right), \quad \tau(\mathbf{x}^\star) = \mathrm{softplus}\left(g_{\hat{\boldsymbol{\phi}}_3}(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}^\star))\right).$$

Step 2: Prediction:

$$\hat{y}(\mathbf{x}^\star) = \underset{k\in[K]}{\mathrm{argmax}}\, \mathbb{E}_{\boldsymbol{\pi}}[\pi_k], \text{where } \mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = \frac{\alpha_k(\mathbf{x}^\star) + \tau(\mathbf{x}^\star) p_k(\mathbf{x}^\star)}{\alpha_0(\mathbf{x}^\star) + \tau(\mathbf{x}^\star)}, \forall k \in [K].$$

Step 3: Uncertainty Quantification:

$$\text{Total Uncertainty:} \quad \mathrm{TU}(\mathbf{x}^\star) = 1 - \sum_{k=1}^K (\mathbb{E}_{\boldsymbol{\pi}}[\pi_k])^2,$$
$$\text{Epistemic Uncertainty:} \quad \mathrm{EU}(\mathbf{x}^\star) = \sum_{k=1}^K \left\{ \frac{\mathbb{E}_{\boldsymbol{\pi}}[\pi_k]\big(1 - \mathbb{E}_{\boldsymbol{\pi}}[\pi_k]\big)}{\alpha_0(\mathbf{x}^\star) + \tau(\mathbf{x}^\star) + 1} + \frac{\tau^2(\mathbf{x}^\star) p_k(\mathbf{x}^\star)\big(1 - p_k(\mathbf{x}^\star)\big)}{\big(\alpha_0(\mathbf{x}^\star) + \tau(\mathbf{x}^\star)\big)\big(\alpha_0(\mathbf{x}^\star) + \tau(\mathbf{x}^\star) + 1\big)} \right\},$$
$$\text{Aleatoric Uncertainty:} \quad \mathrm{AU}(\mathbf{x}^\star) = \mathrm{TU}(\mathbf{x}^\star) - \mathrm{EU}(\mathbf{x}^\star).$$

Output: Predicted label $\hat{y}(\mathbf{x}^\star)$ and uncertainty estimates $\{\mathrm{TU}(\mathbf{x}^\star), \mathrm{EU}(\mathbf{x}^\star), \mathrm{AU}(\mathbf{x}^\star)\}$

C Additional Explanations about Theorems

In this section, we expand on the theoretical foundations of $\mathcal{F}$-EDL (Section 4). First, we present detailed proofs for Lemma 4.1 through Theorem 4.5.[5] Second, we provide further insights into the Bayesian perspective underlying $\mathcal{F}$-EDL and elaborate on Theorem 4.2. Finally, we offer an extended explanation of the subjective logic (SL) interpretation associated with Proposition 4.6.

C.1 Proofs for Theorems

We prove Lemma 4.1, Theorem 4.2, Theorem 4.3, Theorem 4.4, and Theorem 4.5 sequentially.

Proof of Lemma 4.1 The posterior distribution of $\boldsymbol{\pi}$ given $\{y_i\}_{i=1}^N$ can be expressed as:

$$p(\boldsymbol{\pi}|\{y_i\}_{i=1}^N) \propto p_{\text{FD}}(\boldsymbol{\pi}; \boldsymbol{\alpha}, \mathbf{p}, \tau) \times p(\{y_i\}_{i=1}^N|\boldsymbol{\pi}).$$

Expanding each term, we have:

$$p(\boldsymbol{\pi}|\{y_i\}_{i=1}^N) \propto \prod_{k=1}^K \pi_k^{\alpha_k-1} \sum_{k=1}^K p_k \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k+\tau)} \pi_k^{\tau} \times \prod_{i=1}^N \prod_{k=1}^K \pi_k^{\mathbf{1}_{\{y_i=k\}}}.$$

Simplifying the terms yields:

$$p(\boldsymbol{\pi}|\{y_i\}_{i=1}^N) \propto \prod_{k=1}^K \pi_k^{\alpha_k+\sum_{i=1}^N \mathbf{1}_{\{y_i=k\}}-1} \sum_{k=1}^K p_k \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k+\tau)} \pi_k^{\tau}.$$

Recognizing the resulting form, we see that the posterior distribution of $\boldsymbol{\pi}$ retains the form of an FD distribution, i.e.,

$$p(\boldsymbol{\pi}|\{y_i\}_{i=1}^N) = p_{\text{FD}}(\boldsymbol{\pi}|\boldsymbol{\alpha}', \mathbf{p}, \tau),$$

where $\boldsymbol{\alpha}' = (\alpha_1', \ldots, \alpha_K')$ and each α_k' is given by:

$$\alpha_k' = \alpha_k + \sum_{i=1}^N \mathbf{1}_{\{y_i=k\}}, \forall k \in [K].$$

Thus, the FD distribution is conjugate to the categorical likelihood.

Proof of Theorem 4.2. Assume that the prior distribution for a given input $\mathbf{x}$ is specified as:

$$\boldsymbol{\pi}(\mathbf{x}) \sim \text{FD}\left(\boldsymbol{\alpha}_{\text{prior}}(\mathbf{x}), \mathbf{p}_{\text{prior}}(\mathbf{x}), \tau_{\text{prior}}(\mathbf{x})\right),$$

where the parameters are defined as follows:

$$\boldsymbol{\alpha}_{\text{prior}}(\mathbf{x}) = \mathbf{0}, \;\; \mathbf{p}_{\text{prior}}(\mathbf{x}) = \text{softmax}(g_{\boldsymbol{\phi}_2}(f_{\boldsymbol{\theta}}(\mathbf{x}))), \;\; \tau_{\text{prior}}(\mathbf{x}) = \text{softplus}(g_{\boldsymbol{\phi}_3}(f_{\boldsymbol{\theta}}(\mathbf{x}))).$$

Here, $\boldsymbol{\alpha}_{\text{prior}}(\mathbf{x}) = \mathbf{0}$ represents an improper prior as the prior parameter $\boldsymbol{\alpha}$ is zero. The prior parameters for $\mathbf{p}$ and τ are learned for each input $\mathbf{x}$ using a neural network.

The likelihood is estimated using a neural network as:

$$\sum_{i=1}^N \mathbf{1}_{\{y_i=k\}} = \left(\exp(g_{\boldsymbol{\phi}_1}(f_{\boldsymbol{\theta}}(\mathbf{x})))\right)_k, \forall k \in [K].$$

Using Lemma 4.1, the posterior distribution of $\boldsymbol{\pi}$ for a given input $\mathbf{x}$ is obtained as:

$$\boldsymbol{\pi}(\mathbf{x})|\mathcal{D} \sim \text{FD}\left(\exp(g_{\boldsymbol{\phi}_1}(f_{\boldsymbol{\theta}}(\mathbf{x}))), \text{softmax}(g_{\boldsymbol{\phi}_2}(f_{\boldsymbol{\theta}}(\mathbf{x}))), \text{softplus}(g_{\boldsymbol{\phi}_3}(f_{\boldsymbol{\theta}}(\mathbf{x})))\right).$$

This posterior distribution corresponds to the predicted FD distribution of $\boldsymbol{\pi}$ for a given input $\mathbf{x}$, as described in Section 3.1. Consequently, we conclude that the class probability distribution for a given input $\mathbf{x}$, derived from $\mathcal{F}$-EDL, is equivalent to the posterior FD distribution of $\boldsymbol{\pi}$, obtained with an improper prior with input-dependent parameters learned from the data.

[5] We omit the proof for Proposition 4.6, as it provides a conceptual interpretation rather than a formal mathematical statement. A detailed discussion is included in Appendix C.3.

Proof of Theorem 4.3 The class probability distribution for a testing example $\mathbf{x}^\star$, derived from $\mathcal{F}$-EDL, is expressed as:

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha},\mathbf{p},\tau)=\frac{\Gamma(\alpha_0+\tau)}{\prod_{k=1}^K\Gamma(\alpha_k)}\prod_{k=1}^K\pi_k^{\alpha_k-1}\sum_{k=1}^K p_k\frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k+\tau)}\pi_k^{\tau},$$

where $\alpha_0=\sum_{k=1}^K\alpha_k$. By substituting $\tau=1$ and $p_k=\alpha_k/\alpha_0,\forall k\in[K]$, the expression simplifies as follows:

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0+1)}{\prod_{k=1}^K\Gamma(\alpha_k)}\prod_{k=1}^K\pi_k^{\alpha_k-1}\sum_{k=1}^K(\frac{\alpha_k}{\alpha_0})\frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k+1)}\pi_k.$$

Since $\Gamma(\alpha_k+1)=\alpha_k\Gamma(\alpha_k)$, this reduces to:

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0+1)}{\prod_{k=1}^K\Gamma(\alpha_k)}\prod_{k=1}^K\pi_k^{\alpha_k-1}\sum_{k=1}^K(\frac{\pi_k}{\alpha_0}).$$

Because $\sum_{k=1}^K\pi_k=1$, the expression becomes:

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0+1)}{\prod_{k=1}^K\Gamma(\alpha_k)}\prod_{k=1}^K\pi_k^{\alpha_k-1}\times\frac{1}{\alpha_0}.$$

Rewriting using $\Gamma(\alpha_0+1)=\alpha_0\Gamma(\alpha_0)$:

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0)}{\prod_{k=1}^K\Gamma(\alpha_k)}\prod_{k=1}^K\pi_k^{\alpha_k-1}.$$

This is precisely the probability density function of the Dirichlet distribution, i.e.,

$$p_{\text{FD}}(\boldsymbol{\pi};\boldsymbol{\alpha})=p_{\text{Dir}}(\boldsymbol{\pi};\boldsymbol{\alpha}).$$

Therefore, we conclude that the class probability distribution for $\mathbf{x}^\star$, derived from $\mathcal{F}$-EDL, is equivalent to the class probability distribution of EDL when $\tau=1$ and $p_k=\alpha_k/\alpha_0,\forall k\in[K]$.

Proof of Theorem 4.4 For this proof, we rely on the definition and notation of the FG basis and FD distribution, as detailed in Section 2.2.

To establish that the class probability distribution for a testing example $\mathbf{x}^\star$, derived from $\mathcal{F}$-EDL, is represented as a mixture of Dirichlet distributions and exhibits multimodality, we derive the probability density function of the FD distribution, starting from its FG basis representation.

Since $\mathbf{Z}\sim\text{Mu}(1,\mathbf{p})$, the random variable Z_k follows the distribution:

$$Z_k=\begin{cases}1 & \text{with probability } p_k,\\ 0 & \text{with probability } 1-p_k.\end{cases}$$

The conditional distribution of Y_k given Z_k is:

$$Y_k|Z_k=\begin{cases}W_k+U & \text{if } Z_k=1,\\ W_k & \text{if } Z_k=0,\end{cases}$$

where $W_k\sim\text{Gamma}(\alpha_k),\forall k\in[K]$, and $U\sim\text{Gamma}(\tau)$ are independent Gamma random variables sharing the same scale parameter. Consequently, W_k+U is also Gamma-distributed: $W_k+U\sim\text{Gamma}(\alpha_k+\tau)$.

Let $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)$ denote the class probability vector, where

$$\pi_k=\frac{Y_k}{\sum_{k=1}^K Y_k},\forall k\in[K].$$

By the definition of the FD distribution, it follows that:

$$\boldsymbol{\pi}\sim\text{FD}(\boldsymbol{\alpha},\mathbf{p},\tau),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\mathbf{p} = (p_1, \ldots, p_K)$.

Using the conditional distribution of Y_k, the relationship between $\boldsymbol{\pi}$ and $Y_k, \forall k \in [K]$, and the property that normalizing a set of independent Gamma random variables yields a Dirichlet distribution, the conditional distribution of $\boldsymbol{\pi}|\mathbf{Z}$ is given by:

$$\boldsymbol{\pi}|\mathbf{Z} = \begin{cases} \mathrm{Dir}(\alpha_1, \ldots, \alpha_k + \tau, \ldots, \alpha_K) & \text{if } Z_k = 1 \text{ for some } k, \\ \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) & \text{if } Z_k = 0 \text{ for all } k. \end{cases}$$

Since $\mathbf{Z} \sim \mathrm{Mu}(1, \mathbf{p})$, the case $Z_k = 0$ for all k is impossible. Consequently, the distribution of $\boldsymbol{\pi}$ is derived as:

$$\begin{aligned} p(\boldsymbol{\pi}) &= \sum_{k=1}^{K} P(Z_k = 1 \text{ for } k\text{th class}) \times \mathrm{Dir}(\alpha_1, \ldots, \alpha_k + \tau, \ldots, \alpha_K) \\ &= \sum_{k=1}^{K} p_k \, \mathrm{Dir}(\boldsymbol{\alpha} + \tau \mathbf{e}_k), \end{aligned}$$

where $\mathbf{e}_k$ represents the k-th standard basis vector in $\mathbb{R}^K$.

Thus, the class probability distribution for a testing example $\mathbf{x}^\star$, derived from $\mathcal{F}$-EDL, is expressed as a mixture of Dirichlet distributions. The number of distinct modes in this distribution is determined by the cardinality of $\mathbf{p}$, i.e., $\|\mathbf{p}\|_0$.

Proof of Theorem 4.5 The predictive class probability for a testing example $\mathbf{x}^\star$ belonging to each class k, $\forall k \in [K]$, derived from $\mathcal{F}$-EDL, can be represented as follows:

$$p_{\mathcal{F}\text{-EDL}}(y = k|\mathbf{x}^\star) = \int p(y = k|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \mathbf{p}, \tau, \mathbf{x}^\star) d\boldsymbol{\pi}.$$

Using the fact that $p(y = k|\boldsymbol{\pi}) = \pi_k, \forall k \in [K]$, this integral simplifies to:

$$p_{\mathcal{F}\text{-EDL}}(y = k|\mathbf{x}^\star) = \mathbb{E}_{\boldsymbol{\pi} \sim \mathrm{FD}(\boldsymbol{\alpha}, \mathbf{p}, \tau)}[\pi_k].$$

Substituting the expression for the expectation of π_k and expanding the terms, we obtain:

$$p_{\mathcal{F}\text{-EDL}}(y = k|\mathbf{x}^\star) = \frac{\alpha_0}{\alpha_0 + \tau} \times \frac{\alpha_k}{\alpha_0} + \frac{\tau}{\alpha_0 + \tau} \times p_k.$$

On the other hand, for the EDL framework, the predictive distribution for $\mathbf{x}^\star$ is given by:

$$p_{\mathrm{EDL}}(y = k|\mathbf{x}^\star) = \mathbb{E}_{\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})}[\pi_k] = \frac{\alpha_k}{\alpha_0}.$$

For the softmax-based model, the predictive distribution for $\mathbf{x}^\star$ is given by:

$$p_{\mathrm{SM}}(y = k|\mathbf{x}^\star) = p_k.$$

Define the input-dependent mixture weights as:

$$w_{\mathrm{EDL}}(\mathbf{x}^\star) = \frac{\alpha_0}{\alpha_0 + \tau}, \;\; w_{\mathrm{SM}}(\mathbf{x}^\star) = \frac{\tau}{\alpha_0 + \tau}.$$

Using the predictive distributions of EDL and softmax models and substituting these weights, the predictive distribution of $\mathcal{F}$-EDL can be expressed as:

$$p_{\mathcal{F}\text{-EDL}}(y|\mathbf{x}^\star) = w_{\mathrm{EDL}}(\mathbf{x}^\star) \times p_{\mathrm{EDL}}(y|\mathbf{x}^\star) + w_{\mathrm{SM}}(\mathbf{x}^\star) \times p_{\mathrm{SM}}(y|\mathbf{x}^\star).$$

Thus, the predictive distribution for $\mathbf{x}^\star$, obtained by the $\mathcal{F}$-EDL framework, is decomposed into the contributions of the EDL and softmax models.

C.2 Bayesian Interpretation of $\mathcal{F}$-EDL (Theorem 4.2)

In three stages, we provide an additional explanation about the Bayesian interpretation of $\mathcal{F}$-EDL, outlined in Theorem 4.2. First, we establish a Bayesian framework termed the *input-dependent FD-Categorical model*, extending the *input-dependent Dirichlet-Categorical model* commonly used to interpret EDL methods [23, 11]. Second, we introduce a *prior specification* problem, which has been a significant challenge in traditional EDL methods. Third, we introduce an additional theorem (Theorem C.1) that bridges the gap between the prior specification in $\mathcal{F}$-EDL and traditional EDL methods. By leveraging this theorem together with Theorem 4.2 from the main text, we clarify how $\mathcal{F}$-EDL resolves the *prior specification* problem using an improper prior and input-dependent prior parameters.

Input-Dependent FD-Categorical Model. Let $\mathbf{x}$ represent a given input, $\boldsymbol{\pi}(\mathbf{x})$ the corresponding class probabilities, and $\{\tilde{y}_i\}_{i=1}^N$ the set of *pseudo-observations*. Then, the *input-dependent FD-Categorical model* can be described as a Bayesian model as follows:

$$\begin{aligned}
&\textbf{Prior} && \boldsymbol{\pi}(\mathbf{x}) \sim \text{FD}(\boldsymbol{\alpha}_{\text{prior}}(\mathbf{x}), \mathbf{p}_{\text{prior}}(\mathbf{x}), \tau_{\text{prior}}(\mathbf{x})),\\
&\textbf{Likelihood} && \{\tilde{y}_i\}_{i=1}^N \sim \text{Cat}(\boldsymbol{\pi}(\mathbf{x})),\\
&\textbf{Posterior} && \boldsymbol{\pi}(\mathbf{x})|\{\tilde{y}_i\}_{i=1}^N \sim \text{FD}(\boldsymbol{\alpha}_{\text{post}(\mathbf{x})}, \mathbf{p}_{\text{post}}(\mathbf{x}), \tau_{\text{post}}(\mathbf{x})),
\end{aligned}$$

where $\boldsymbol{\alpha}_{\text{prior}}(\mathbf{x})$, $\mathbf{p}_{\text{prior}}(\mathbf{x})$, and $\tau_{\text{prior}}(\mathbf{x})$ are the prior parameters, and $\boldsymbol{\alpha}_{\text{post}(\mathbf{x})}$, $\mathbf{p}_{\text{post}}(\mathbf{x})$, and $\tau_{\text{post}}(\mathbf{x})$ are the posterior parameters.

Leveraging the results in Lemma 4.1, the relationship between the prior and posterior parameters can be expressed as follows:

$$\begin{aligned}
\alpha_{\text{post}}^{(k)}(\mathbf{x}) &= \alpha_{\text{prior}}^{(k)}(\mathbf{x}) + \sum_{i=1}^{N} \mathbf{1}_{\{\tilde{y}_i=k\}}, \forall k \in [K],\\
\mathbf{p}_{\text{post}}(\mathbf{x}) &= \mathbf{p}_{\text{prior}}(\mathbf{x}), \quad \tau_{\text{post}}(\mathbf{x}) = \tau_{\text{prior}}(\mathbf{x}).
\end{aligned}$$

This formulation reveals the intrinsic connection between $\mathcal{F}$-EDL and the input-dependent FD-Categorical model. The fundamental concept of $\mathcal{F}$-EDL, predicting an FD distribution over class probabilities for a given input $\mathbf{x}$, is inherently equivalent to predicting the posterior distribution over class probabilities, given the pre-specified prior distributions. In particular, the mechanism of $\mathcal{F}$-EDL is equivalent to predicting *pseudo-observations* $\{\tilde{y}_i^{(j)}\}_{j=1}^N$ for each data point $\mathbf{x}$ using a neural network. Specifically, $\mathcal{F}$-EDL predicts the counts of the pseudo-observations for each class, i.e., *pseudo-counts*, utilizing a neural network $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\phi}_1}$ as follows:

$$\sum_{i=1}^{N} \mathbf{1}_{\{\tilde{y}_i=k\}} \approx \exp(g_{\boldsymbol{\phi}_1}(f_{\boldsymbol{\theta}}(\mathbf{x})))_k.$$

Using this, the posterior parameter, $\boldsymbol{\alpha}_{\text{post}}$, is expressed as:

$$\boldsymbol{\alpha}_{\text{post}}(\mathbf{x}) = \boldsymbol{\alpha}_{\text{prior}}(\mathbf{x}) + \exp(g_{\boldsymbol{\phi}_1}(f_{\boldsymbol{\theta}}(\mathbf{x}))).$$

Prior Specification Problem of EDL Methods. Standard EDL methods face the *prior specification* problem, as highlighted in recent works, including R-EDL [10] and DAEDL [11]. This issue arises because EDL methods employ a uniform prior, $\boldsymbol{\pi}_{\text{prior}}(\mathbf{x}) \sim \text{Dir}(\mathbf{1})$, when predicting the posterior distribution under the *input-dependent Dirichlet-Categorical model* framework [23, 11]. Notably, the *input-dependent Dirichlet-Categorical model* is analogous to the *input-dependent FD-Categorical model* discussed earlier, differing only in its use of the Dirichlet distribution in place of the FD distribution. To address the *prior specification* problem under this framework, R-EDL [10] treats the prior parameter as an adjustable hyperparameter, adopting a prior of the form $\boldsymbol{\pi}_{\text{prior}}(\mathbf{x}) \sim \text{Dir}(\lambda\mathbf{1})$, where λ is a tunable hyperparameter. DAEDL [11] eliminates the prior parameter, effectively utilizing an improper prior $\boldsymbol{\pi}_{\text{prior}}(\mathbf{x}) \sim \text{Dir}(\mathbf{0})$. For a more detailed discussion of the prior specification problem in EDL methods and their proposed solutions, we refer the reader to the respective papers [10, 11].

Table 6: Comparison of prior specification scheme across EDL [8], $\mathcal{I}$-EDL [9], R-EDL [10], DAEDL [11], and $\mathcal{F}$-EDL. Here, $\lambda \in \mathbb{R}$ denotes a tunable hyperparameter, $\mathbf{1} \in \mathbb{R}^K$ is a vector of ones, $\mathbf{0} \in \mathbb{R}^K$ is a vector of zeros, and $\mathbf{p}_{\text{prior}}(\mathbf{x}) \in \mathbb{R}^K$ and $\tau_{\text{prior}}(\mathbf{x}) \in \mathbb{R}$ denote the prior parameters dynamically learned for a given input $\mathbf{x}$.

Method	EDL, $\mathcal{I}$-EDL	R-EDL	DAEDL	$\mathcal{F}$-EDL (Ours)
Prior	$\boldsymbol{\pi}(\mathbf{x}) \sim \text{Dir}(\mathbf{1})$	$\boldsymbol{\pi}(\mathbf{x}) \sim \text{Dir}(\lambda\mathbf{1})$	$\boldsymbol{\pi}(\mathbf{x}) \sim \text{Dir}(\mathbf{0})$	$\boldsymbol{\pi}(\mathbf{x}) \sim \text{FD}(\mathbf{0}, \mathbf{p}_{\text{prior}}(\mathbf{x}), \tau_{\text{prior}}(\mathbf{x}))$

While these methods partially address the *prior specification* problem, both approaches have notable limitations. First, R-EDL requires manual tuning of the additional hyperparameter λ, which can be both challenging and time-consuming, particularly when applied to new tasks or datasets. Since the

choice of λ has a non-negligible impact on UQ performance, precise tuning is critical for achieving optimal performance. Second, while DAEDL effectively mitigates the prior specification issue through its parameterization scheme, its success depends heavily on accurate density estimation to represent uncertainty. This dependence can become problematic in complex ID scenarios, where obtaining high-quality density estimates is particularly challenging. Thus, there remains a need for an EDL model that addresses the *prior specification* problem without introducing additional computational overhead or relying on precise density estimation. Such a model would enhance the robustness and generalizability of EDL methods across complex and unforeseen scenarios.

Table 6 provides a comprehensive comparison of the prior specification scheme used by representative EDL methods and $\mathcal{F}$-EDL.

Resolving the Prior Specification Problem with $\mathcal{F}$-EDL. $\mathcal{F}$-EDL presents a novel approach to the *prior specification* problem by utilizing an improper prior with learned parameters. Before exploring this further, we introduce an additional theorem that bridges the gap between $\mathcal{F}$-EDL and EDL.

Theorem C.1. ***(Variant of Theorem 4.2)*** *The predictive distribution for a given input* $\mathbf{x}$*, derived from* $\mathcal{F}$*-EDL, is equivalent to the predictive distribution derived from EDL, utilizing the prior* $\boldsymbol{\pi}(\mathbf{x}) \sim \mathrm{Dir}(\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x}))$*, where* $\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x})$ *are the input-dependent prior parameters learned from the data.*

Proof of Theorem C.1 The predictive distribution for a given input $\mathbf{x}$, derived from $\mathcal{F}$-EDL, is expressed as:

$$p_{\mathcal{F}\text{- EDL}}(y = k|\mathbf{x}) = \frac{\alpha_k + \tau p_k}{\sum_{k=1}^{K} \alpha_k + \tau}, \forall k \in [K].$$

Let us define $\boldsymbol{\alpha}' = (\alpha'_1, \ldots, \alpha'_K)$, where

$$\alpha'_k = \alpha_k + \tau p_k, \forall k \in [K].$$

Using this definition, the predictive distribution for $\mathcal{F}$-EDL can be rewritten as:

$$p_{\mathcal{F}\text{- EDL}}(y = k|\mathbf{x}) = \frac{\alpha'_k}{\sum_{k=1}^{K} \alpha'_k}.$$

This expression is equivalent to the predictive distribution derived from the standard EDL model, where the concentration parameters $\boldsymbol{\alpha}'$ are parameterized as follows:

$$\boldsymbol{\alpha}' = \boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x}) + \mathbf{e}(\mathbf{x}),$$

with the components defined as:

$$\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x}) = \mathrm{softplus}(g_{\boldsymbol{\phi_3}}(f_{\boldsymbol{\theta}}(\mathbf{x}))) \times \mathrm{softmax}(g_{\boldsymbol{\phi_2}}(f_{\boldsymbol{\theta}}(\mathbf{x}))), \quad \mathbf{e}(\mathbf{x}) = \exp(g_{\boldsymbol{\phi_1}}(f_{\boldsymbol{\theta}}(\mathbf{x}))).$$

Within the *input-dependent Dirichlet-Categorical model* framework, EDL with this parameterization can be interpreted as predicting the posterior Dirichlet distribution for $\boldsymbol{\pi}$, where the likelihood (evidence vector) $\mathbf{e}(\mathbf{x})$ is defined as described earlier, and the prior distribution is specified as:

$$\boldsymbol{\pi}(\mathbf{x}) \sim \mathrm{Dir}(\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x})),$$

where $\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x})$ is the input-dependent prior parameters learned from the data.

In essence, Theorem 4.2 demonstrates that $\mathcal{F}$-EDL can be interpreted as predicting the posterior FD distribution for a given input, under the *input-dependent FD-Categorical model* framework. Furthermore, Theorem C.1 establishes that the predictive distribution of $\mathcal{F}$-EDL is equivalent to that of EDL, with an input-dependent prior, i.e., $\boldsymbol{\pi}(\mathbf{x}) \sim \mathrm{Dir}(\boldsymbol{\lambda}_{\mathrm{prior}}(\mathbf{x}))$.

From these theorems, we can infer that the prior specification scheme of $\mathcal{F}$-EDL demonstrates distinct advantages in two different aspects compared to the conventional approaches. First, $\mathcal{F}$-EDL utilizes an improper prior by setting $\boldsymbol{\alpha}_{\mathrm{prior}}(\mathbf{x}) = \mathbf{0}$. This addresses the limitation of uniform priors,

Table 7: Subjective logic interpretation of EDL [8], R-EDL [10], and $\mathcal{F}$-EDL. For EDL and R-EDL, a single SL triplet $(\mathbf{b}, u, \mathbf{a})$ represents a single opinion, where $\mathbf{b} \in \mathbb{R}^K$, $u \in \mathbb{R}$, and $\mathbf{a} \in \Delta^{K-1}$ represent belief mass, uncertainty, and base rate, respectively. In contrast, $\mathcal{F}$-EDL yields K distinct SL triplets, each encoding a class-specific opinion.

Property	**EDL**	**R-EDL**	**$\mathcal{F}$-EDL (Ours)**
Distribution	Dirichlet	Dirichlet	Flexible Dirichlet
SL Triplet $(\mathbf{b}, u, \mathbf{a})$	$\left(\frac{\boldsymbol{\alpha}-\mathbf{1}}{\|\boldsymbol{\alpha}\|_1}, \frac{K}{\|\boldsymbol{\alpha}\|_1}, \frac{\mathbf{1}}{K}\right)$	$\left(\frac{\boldsymbol{\alpha}-\lambda\mathbf{1}}{\|\boldsymbol{\alpha}\|_1}, \frac{\lambda K}{\|\boldsymbol{\alpha}\|_1}, \frac{\mathbf{1}}{K}\right)$	$\left(\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_1+\tau}, \frac{\tau}{\|\boldsymbol{\alpha}\|_1+\tau}, \mathbf{e}_j\right), \forall j \in [K]$
Hypothesis assignment	–	–	$\mathbf{p}$ over K class-specific opinions
Projected Probability $\mathbf{P}$	$\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_1}$	Same	$\frac{\boldsymbol{\alpha}+\tau\mathbf{p}}{\|\boldsymbol{\alpha}\|_1+\tau}$
Multimodality	No	No	Yes (Dirichlet mixture)
Epistemic source	magnitude of $\boldsymbol{\alpha}$	Same	$\mathbf{p}$ (inter-opinion) + $\boldsymbol{\alpha}, \tau$ (intra-opinion)

which often degrade classification performance due to the difficulty in balancing the prior parameters with the *pseudo-likelihood* [11]. By adopting this approach, $\mathcal{F}$-EDL avoids the challenges associated with using fixed uniform priors. Second, $\mathcal{F}$-EDL employs learnable prior parameters for $\mathbf{p}$ and τ. Specifically, the optimal prior parameters for each data point $\mathbf{x}$ are estimated using neural networks. Compared to R-EDL, which requires manual tuning of prior parameters, $\mathcal{F}$-EDL offers a significant advantage by adaptively determining these parameters, enabling both optimal predictive performance and robust UQ. Additionally, compared to DAEDL, $\mathcal{F}$-EDL reduces the dependence on density estimation to achieve optimal UQ, enabling effective generalization to the challenging scenarios where obtaining high-quality density estimates is infeasible.

C.3 Subjective Logic Interpretation of $\mathcal{F}$-EDL (Proposition 4.6)

In this section, we elaborate on Proposition 4.6 to clarify the subjective logic (SL) perspective of $\mathcal{F}$-EDL. We begin by revisiting the motivation of $\mathcal{F}$-EDL through the SL lens. Next, we formalize its multimodal predictive structure as a mixture of class-specific Dirichlet-distributed opinions. We then introduce a *generalized subjective logic* interpretation, highlighting how it extends traditional EDL formulations. Finally, we explain how the proposed variance-based uncertainty measures naturally align with this interpretation, enabling faithful UQ.

Motivation: Limitations of Single-Opinion SL in EDL Most EDL methods [8, 9, 10, 11] can be interpreted through the lens of SL [14], where each prediction corresponds to a single subjective opinion—a Dirichlet distribution encoding belief and uncertainty over class probabilities. However, this one-opinion-per-input framework restricts the model's ability to express the ambiguity between multiple plausible hypotheses. Such limitations are especially evident for inputs with overlapping features (e.g., ambiguous digits such as "1" vs "7" in MNIST)—which commonly arise in high-risk, real-world ML applications.

$\mathcal{F}$-EDL as a Mixture of Opinions. To address the limitation of representing only a single opinion, $\mathcal{F}$-EDL models uncertainty as a mixture of class-specific subjective opinions. Each component corresponds to a class-specific hypothesis, represented by a Dirichlet distribution biased toward that class, and weighted by a hypothesis selection probability. For a test input $\mathbf{x}^\star$, the resulting predictive distribution is:

$$p(\boldsymbol{\pi}|\mathbf{x}^\star) = \sum_{k=1}^{K} p_k \operatorname{Dir}(\boldsymbol{\alpha} + \tau\mathbf{e}_k),$$

where $\boldsymbol{\alpha}$ is the shared base concentration vector, τ controls the strength of class-specific bias in each opinion, $\mathbf{p} = (p_1, \ldots, p_K)$ denotes the mixture weights, and $\mathbf{e}_k$ is the k-th standard basis vector. This defines an FD distribution—a structured Dirichlet mixture capable of expressing multimodal beliefs over class probabilities.

Generalized Subjective Logic Interpretation of $\mathcal{F}$-EDL. This formulation aligns $\mathcal{F}$-EDL with a generalized SL framework that accommodates multiple competing opinions, rather than collapsing

them into a single fused belief. Each Dirichlet component $\mathrm{Dir}(\boldsymbol{\alpha} + \tau \mathbf{e}_k)$ encodes a class-specific opinion, while the hypothesis selection probabilities $\mathbf{p}$ represent epistemic uncertainty over which opinion to endorse. Unlike classical EDL models, which aggregate all evidence into a single unimodal Dirichlet, $\mathcal{F}$-EDL maintains the structure of uncertainty across hypotheses. This allows it to capture ambiguity between plausible alternatives—enabling a more expressive and faithful representation of model belief.

Table 7 summarizes this interpretation by comparing $\mathcal{F}$-EDL with representative EDL variants from the SL perspective. Unlike conventional EDL models that yield SL triplets per input, $\mathcal{F}$-EDL produces K distinct SL triplets—each representing class-specific subjective opinion. These triples share the belief and uncertainty masses $(\mathbf{b}, u)$, but differ in the base rate $\mathbf{a}$, reflecting the class-dependent bias. The overall opinion is expressed as a mixture of these K opinions, weighted by a categorical distribution $\mathbf{p}$ that is learned per input. This structured formulation enables principled UQ through evidence aggregation grounded in subjective logic.

In particular, it is noteworthy that, unlike EDL variants that yield a single SL triplet per input, $\mathcal{F}$-EDL produces K distinct SL triplet, sharing the belief and uncertainty massess $(\mathbf{b}, u)$ but differing in the base rate $\mathbf{a}$ due to class-dependent bias. The overall subjective opinion is distributed through these K subjective opinions, following the categorical distribution parameterized by $\mathbf{p}$, which is also learned per input. These allow principled uncertainty quantification through the structured evidence aggregation grounded by subjective logic.

Variance-based Uncertainty Measures in $\mathcal{F}$-EDL Classical SL metrics (e.g., inverse total evidence) are not directly applicable to this mixture. Instead, we quantify epistemic uncertainty using the total variance of the FD distribution:

$$\mathrm{EU}(\mathbf{x}^\star) = \sum_{k=1}^{K} \mathrm{Var}(\pi_k(\mathbf{x}^\star)).$$

This formulation captures both inter-hypothesis ambiguity, reflected in the hypothesis selection probabilities (i.e., allocation probabilities) $\mathbf{p}$, and intra-hypothesis variability, driven by the evidence parameters $\boldsymbol{\alpha}$ and dispersion τ. By explicitly modeling these two sources of uncertainty, $\mathcal{F}$-EDL offers more faithful and interpretable estimates of epistemic uncertainty, especially in ambiguous or complex prediction scenarios.

D Interpretation of FD Parameters and Justification for Its Use in UQ

This section provides additional explanation for adopting the FD distribution for UQ, demonstrating that its use represents a principled generalization rather than an arbitrary increase in flexibility. First, we clarify the semantics of the FD parameters $(\boldsymbol{\alpha}, \mathbf{p}, \tau)$. Second, we justify the suitability of the FD distribution for UQ, particularly in cases involving multiple plausible class hypotheses. Third, we describe how evidence is structurally extracted in $\mathcal{F}$-EDL.

D.1 Semantics of FD Distribution Parameters

In the FD distribution, the parameters $\boldsymbol{\alpha}$, $\mathbf{p}$, and τ jointly characterize different aspects of uncertainty. The vector $\boldsymbol{\alpha}$ represents the total amount of evidence supporting each class, analogous to the concentration parameters in a standard Dirichlet distribution. However, unlike the Dirichlet case, the FD introduces two additional parameters—$\mathbf{p}$ and τ—that disentangle how this evidence is distributed and how sharply it is expressed.

The parameter $\mathbf{p}$ specifies how belief is allocated among class-specific hypotheses. A sharp $\mathbf{p}$, where one component dominates, reflects a decisive belief in a particular class; conversely, a diffuse $\mathbf{p}$ indicates competing hypotheses and greater ambiguity in class assignment. Thus, $\mathbf{p}$ governs the *directional aspect* of uncertainty—how belief is distributed across class-specific hypotheses.

The scalar τ modulates the *intensity* or *concentration* of each hypothesis. A smaller τ yields more concentrated, confident distributions for each hypothesis, while a larger τ produces flatter, more dispersed components, effectively tempering overconfidence in uncertain or ambiguous regions. By dynamically adapting τ, the FD can represent varying degrees of epistemic caution in response to data complexity.

Together, $\mathbf{p}$ and τ provide structural flexibility that extends beyond the magnitude of evidence encoded by $\boldsymbol{\alpha}$. This decomposition allows the FD to express both *how much* evidence the model has and *how that evidence is organized* across each competing hypothesis—leading to more faithful and principled UQ under complex or unforeseen data conditions.

D.2 Justification for Using the FD Distribution for UQ

We employ the FD distribution to address a central limitation of traditional EDL—its inability to produce reliable uncertainty estimates in complex or ambiguous scenarios where multiple class hypotheses may be simultaneously plausible.

For instance, when an input resembles both a "7" and a "9", a standard EDL, which relies on a single Dirichlet distribution, produces a unimodal belief that tends to overcommit to the dominant class (see Figure 3, second and third panels). This representation fails to capture the inherent ambiguity of the input and leads to overconfident predictions. In contrast, $\mathcal{F}$-EDL employs a structured mixture of class-specific Dirichlet hypotheses, where each component encodes the belief *this image corresponds to class* k. All components share a common base evidence $\boldsymbol{\alpha}$, while the parameters $\mathbf{p}$ and τ respectively determine the mixture weights and concentration of each component. This formulation enables the model to represent multimodal beliefs in a principled way, naturally capturing conflicting hypotheses and reducing overconfidence in uncertainty regions of the input space.

D.3 Evidence Extraction in $\mathcal{F}$-EDL

For each input, $\mathcal{F}$-EDL extracts and structures evidence through three conceptually interpretable stages:

- **(1) Base Evidence Extraction:** The model first estimates the base evidence $\boldsymbol{\alpha}$, analogous to standard EDL. This represents the initial strength of belief across classes and may be overconfident in ambiguous cases.
- **(2) Construction of Class-Specific Hypotheses:** Using $\boldsymbol{\alpha}$, $\mathcal{F}$-EDL constructs a Dirichlet hypothesis for each class, treating each as a plausible explanation of the input. This enables the model to represent multiple competing class-level beliefs.
- **(3) Input-Adaptive Mixing via $\mathbf{p}$ and τ:** The parameters $\mathbf{p}$ and τ govern how these hypotheses are integrated. $\mathbf{p}$ assigns mixture weights to each class-specific Dirichlet—peaked for clean inputs, dispersed for ambiguous samples, and near-uniform for OOD data—reflecting how belief is distributed across hypotheses. τ modulates the sharpness of each component: it increases for ambiguous inputs to reduce spurious confidence, and remains low for clear inputs to preserve certainty.

This three-stage process aligns naturally with the mixture-based formulation of the FD distribution, emphasizing that $\mathcal{F}$-EDL's architecture arises from a principled probabilistic design rather than an ad-hoc extension of EDL.

E Additional Explanations about Experiments

In this section, we present a comprehensive explanation of our experiments. First, we describe the datasets used in our study. Second, we outline the implementation details.

E.1 Datasets

In line with recent works in EDL [23, 9, 10, 11], we used CIFAR-10 [43] as the primary ID dataset for our evaluations. To assess UQ in more complex scenarios, we also employed CIFAR-100 [43] as an additional ID dataset.

For CIFAR-10, we considered the Street View House Numbers (SVHN) [45] and CIFAR-100 as OOD datasets. When CIFAR-100 was used as the ID dataset, its corresponding OOD datasets included SVHN and Tiny-ImageNet-200 (TIN). To investigate the generalizability of $\mathcal{F}$-EDL in long-tailed ID scenarios, we employed CIFAR-10-LT [44], a version of CIFAR-10 with artificially imbalanced class distributions. Specifically, we evaluated two levels of imbalance: mild ($\rho = 0.1$) and heavy ($\rho = 0.01$), using the same OOD dataset as CIFAR-10. For noisy ID scenarios, we utilized Dirty-MNIST (DMNIST) [7], a noisy variant of MNIST [52], with Fashion-MNIST (FMNIST) [53] as its OOD dataset. Additionally, for assessing distribution shift detection capabilities, we adopted CIFAR-10-C [46], which introduces controlled perturbations to CIFAR-10. Detailed descriptions of these datasets are provided below.

CIFAR-10 [43] is one of the most widely used image classification datasets in the UQ literature. It consists of color images categorized into 10 classes, representing various animals and objects: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The dataset contains 50,000 training images and 10,000 testing images, with each image represented as a tensor of shape $3 \times 32 \times 32$ corresponding to RGB channels and spatial dimensions. The dataset is balanced, meaning each class contains an equal number of samples in both the training and testing splits. For our experiments, the training set is further divided into training and validation subsets with a ratio of $0.95 : 0.05$. CIFAR-10 served as the primary ID dataset and was also used to create CIFAR-10-LT, a long-tailed variant.

Street View House Numbers (SVHN) [45] is a dataset comprising cropped images of house numbers obtained from Google Street View. It contains 73,257 training images, 26,032 testing images, and 531,131 additional images, each represented as a tensor of the shape $3 \times 32 \times 32$. In our experiments, the test set of SVHN was employed as the OOD dataset when CIFAR-10, CIFAR-10-LT, or CIFAR-100 served as the ID dataset.

CIFAR-100 [43] is a more detailed and challenging version of CIFAR-10, comprising 100 classes, each representing a distinct animal and object category. It includes 50,000 training images and 10,000 testing images, with each image represented as a tensor of the shape $3 \times 32 \times 32$. In our experiments, the training set was further divided into training and validation subsets with a ratio of $0.95 : 0.05$. CIFAR-100 was used as the ID dataset in the classical setting. Additionally, its test set was employed as an OOD dataset when either CIFAR-10 or CIFAR-10-LT served as the ID dataset.

Tiny-ImageNet-200 (TIN) [54] is a smaller version of the ImageNet [55] dataset, comprising 200 classes, each representing a distinct category. It contains 100,000 training images, 10,000 validation images, and 10,000 testing images, with each image represented as a tensor of the shape $3 \times 64 \times 64$. In our experiments, the test set of TIN was used as the OOD dataset when CIFAR-100 served as the ID dataset. Due to the mismatch in the image sizes, we resized TIN images to $3 \times 32 \times 32$ for compatibility.

CIFAR-10-LT [56] is a long-tailed version of CIFAR-10, characterized by artifically imbalanced class distributions. The imbalance severity is controlled using an imbalance factor (ρ), defined as the ratio of the number of samples in the head class to the tail class. In our experiments, CIFAR-10-LT datasets with two imbalance factors, $\rho = 0.1$ and $\rho = 0.01$, were used to validate the generalizability of $\mathcal{F}$-EDL in long-tailed ID scenarios.

CIFAR-10-C [57] is a corrupted version of the CIFAR-10 dataset, designed to evaluate the robustness of image classification models against various types of corruption and noise. It introduces 19 types of real-world corruptions, including Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate,

Table 8: Comparison of the total parameters and the additional parameters introduced by the MLPs across different architectures. "# of Params." refers to the total number of parameters in the base model architectures (e.g., ConvNet, VGG-16, and ResNet-18). "# of Additional Params." denotes the number of parameters added by the MLPs. "Proportion" indicates the percentage increase in the total parameter count due to the addition of the MLPs.

ID Dataset	Architecture	# of Params.	# of additional params.	Proportion (%)
DMNIST	ConvNet	248,330	6,347	**2.56%**
CIFAR-10, CIFAR-10-LT	VGG-16	14,857,546	266,507	**1.79%**
CIFAR-100	ResNet-18	11,046,308	289,367	**2.62%**

Table 9: Average batch inference time (batch size = 64) on the CIFAR-10 dataset using the VGG-16 backbone. The mean and standard deviation are calculated over five independent runs.

Metric	EDL	DAEDL	$\mathcal{F}$-EDL
Inference Time (s)	1.262 ± 0.06	2.683 ± 0.05	1.279 ± 0.08

jpeg compression, speckle noise, Gaussian blur, splatter, and saturate. Each corruption is applied at five different severity levels $\mathcal{C} \in \{1, 2, 3, 4, 5\}$, resulting in a total of 950,000 corrupted images (10,000 images $\times$ 19 corruptions $\times$ 5 severities). In our experiments, CIFAR-10-C was used for distribution shift detection tasks. Specifically, models trained on CIFAR-10 performed OOD detection using CIFAR-10-C as the OOD dataset to assess their ability to capture distribution shifts through uncertainty measures.

Dirty-MNIST (DMNIST) [7] is a noisy variant of the MNIST [52] dataset, containing ambiguous data points. It is generated by integrating Ambiguous-MNIST (AMNIST) [7], which includes artificially synthesized MNIST samples with varying entropy levels, into the original MNIST dataset. DMNIST contains 120,000 training images (60,000 from MNIST and 60,000 from AMNIST) and 70,000 testing images (10,000 from MNIST and 60,000 from AMNIST). In our experiments, DMNIST was used to validate the generalizability of $\mathcal{F}$-EDL in noisy ID scenarios.

Fashion-MNIST (FMNIST) [53] is a modern replacement for the MNIST dataset, consisting of grayscale images from 10 classes that represent various fashion items, including clothing, footwear, and accessories. The dataset contains 60,000 training images and 10,000 testing images, each represented as a tensor of shape $1 \times 28 \times 28$, similar to MNIST. In our experiments, FMNIST served as the OOD dataset when DMNIST was used as the ID dataset.

E.2 Implementation Details

To ensure a fair comparison, we adopted VGG-16 [47] as the model architecture when CIFAR-10 served as the ID dataset, consistent with recent studies [23, 9, 10, 11]. The same architecture was utilized when CIFAR-10-LT was employed as the ID dataset. For CIFAR-100, ResNet-18 [48] was employed as the model architecture, while a simple convolutional neural network (ConvNet) consisting of three convolutional layers followed by three dense layers was implemented for DMNIST.

To compute the parameters $\mathbf{p}$ and τ, we added two shallow MLPs into each architecture. These MLPs consisted of a single layer when DMNIST was the ID dataset and two layers for the other dataset. The additional model complexity introduced by these MLPs was minimal, as demonstrated in Table 8. Moreover, the added overhead becomes negligible for larger architectures, indicating that $\mathcal{F}$-EDL scales efficiently. For instance, with WideResNet-28-10 (36.5M parameters) on TinyImageNet-200, the extra heads introduce only 346K parameters (0.95%). This relative overhead further diminishes as the model capacity increases.

In terms of inference efficiency, $\mathcal{F}$-EDL enables fast prediction and UQ without sampling or post-hoc steps (e.g., DAEDL's density estimation). As shown in Table 9, the average batch inference time, including FD parameter prediction and uncertainty computation, was 1.279s $\pm$ 0.08—only 1.35% slower than EDL (1.262s $\pm$ 0.06) and 52.34% faster than DAEDL (2.683s $\pm$0.05).

Table 10: Implementation details and hyperparameter settings for our experiments. Here, $T_{\max}$, B, and η denote the maximum number of epochs, batch size, and learning rate, respectively. Additionally, L and H represent the number of layers and hidden dimensions of the MLPs, respectively.

ID Dataset	**Architecture**	**Optimizer**	**Scheduler**	$\mathbf{T}_{\max}$	**B**	$\boldsymbol{\eta}$	**Step size**	**L**	**H**
DMNIST	ConvNet	Adam	StepLR	50	64	5×10^{-4}	20	1	64
CIFAR-10	VGG-16	Adam	StepLR	100	64	5×10^{-4}	30	2	256
CIFAR-10-LT	VGG-16	Adam	StepLR	100	64	5×10^{-4}	30	2	256
CIFAR-100	ResNet-18	Adam	StepLR	100	64	10^{-4}	30	2	256

Based on the validation loss, early stopping was applied to mitigate overfitting across all experiments. A fixed batch size of $B = 64$ was used in all settings. Training was conducted for up to 50 epochs on DMNIST and 100 epochs on other datasets. The Adam optimizer [58] and the StepLR scheduler were consistently employed. Hyperparameters for both our proposed method and the baselines were selected through grid search. Importantly, $\mathcal{F}$-EDL eliminates the need for hyperparameter tuning in its objective function, significantly simplifying the training process. The detailed experimental setups and hyperparameter configurations are provided in Table 10. All experiments were implemented in PyTorch. Depending on availability, we used either an RTX 4060 GPU with 8GB memory or a TITAN V GPU with 12GB memory.

Table 11: Comprehensive results for UQ-related downstream tasks in the classical setting using CIFAR-10 as the ID dataset, with additional results from PostNet [23] and DUQ [6].

Method	Test.Acc.	Conf.	SVHN / CIFAR-100
Dropout	82.84 ±0.1	97.15 ±0.0	51.39 ±0.1 / 45.57 ±1.0
PostNet	84.85 ±0.0	97.76 ±0.0	77.71 ±0.3 / 81.96 ±0.8
DUQ	89.33 ±0.2	97.89 ±0.3	80.23 ±3.4 / 84.75 ±1.1
EDL	83.55 ±0.6	97.86 ±0.2	79.12 ±3.7 / 84.18 ±0.7
$\mathcal{I}$-EDL	89.20 ±0.3	98.72 ±0.1	82.96 ±2.2 / 84.84 ±0.6
R-EDL	90.09 ±0.3	98.98 ±0.1	85.00 ±1.2 / 87.73 ±0.3
DAEDL	91.11 ±0.2	99.08 ±0.0	85.54 ±1.4 / 88.19 ±0.1
$\mathcal{F}$-EDL	**91.19** ±0.2	**99.10** ±0.0	**91.20** ±1.3 / **88.37** ±0.3

Table 12: AUROC scores for OOD detection using epistemic uncertainty estimates in the classical setting. Results are reported using CIFAR-10 as the ID dataset with SVHN and CIFAR-100 (C-100) as OOD datasets, and CIFAR-100 as the ID dataset with SVHN and TinyImageNet (TIN) as OOD datasets.

Method	ID: CIFAR-10 OOD: SVHN / C-100	ID: CIFAR-100 OOD: SVHN / TIN
EDL	81.06 ±4.5 / 80.63 ±1.0	63.95 ±3.4 / 65.32 ±2.3
$\mathcal{I}$-EDL	86.79 ±1.3 / 82.15 ±0.5	77.85 ±1.5 / 73.34 ±0.3
R-EDL	87.47 ±1.2 / 85.26 ±0.4	77.06 ±2.2 / 71.10 ±1.0
DAEDL	89.24 ±1.0 / 86.04 ±0.1	81.07 ±3.0 / 75.04 ±1.3
$\mathcal{F}$-EDL	**93.74** ±1.5 / **86.37**±0.3	**81.59** ±1.8 / **79.24** ±0.2

F Additional Results for Quantitative Study on UQ-Related Downstream Tasks

Classical Setting. Table 11 presents comprehensive results for UQ-related downstream tasks in the classical setting using CIFAR-10 as the ID dataset, comparing $\mathcal{F}$-EDL with additional baselines:

Table 13: AUROC scores for OOD detection using epistemic uncertainty estimates in the long-tailed setting. Results are reported using CIFAR-10-LT as the ID dataset under mild ($\rho = 0.1$) and heavy ($\rho = 0.01$) imbalance, with SVHN and CIFAR-100 (C-100) as OOD datasets.

Method	ID: CIFAR-10-LT ($\rho = 0.1$) OOD: SVHN / C-100	ID: CIFAR-10-LT ($\rho = 0.01$) OOD: SVHN / C-100
EDL	80.36 $\pm$1.5 / 77.03 $\pm$0.6	58.25 $\pm$4.4 / 61.05 $\pm$0.7
$\mathcal{I}$-EDL	84.65 $\pm$3.8 / 80.83$\pm$0.5	61.47 $\pm$7.4 / 65.12 $\pm$1.4
R-EDL	77.39 $\pm$5.3 / 72.12 $\pm$0.9	62.24 $\pm$2.9 / 64.03 $\pm$0.6
DAEDL	80.34 $\pm$2.9 / 73.92 $\pm$1.3	64.21 $\pm$4.3 / 63.49 $\pm$0.6
$\mathcal{F}$-EDL	**89.22** $\pm$1.7 / **81.54** $\pm$0.6	**71.62** $\pm$1.9 / **67.74** $\pm$1.8

Table 14: UQ-related downstream task results are reported using the DMNIST dataset as the ID dataset. "Brier." denotes the Brier score for calibration. "FMNIST.AUPR." and "FMNIST.AUROC." indicate AUPR and AUROC scores for OOD detection using epistemic uncertainty, with FMNIST as the OOD dataset.

Method	Test.Acc.	Conf.	Brier.	FMNIST.AUPR.	FMNIST.AUROC.
MSP	83.90 $\pm$0.1	96.01 $\pm$0.0	2.38 $\pm$0.0	98.10 $\pm$0.3	89.30 $\pm$1.5
Dropout	84.13 $\pm$0.1	96.09 $\pm$0.0	2.44 $\pm$0.0	94.68 $\pm$0.6	68.97 $\pm$2.8
DDU	84.05 $\pm$0.1	82.73 $\pm$0.1	2.35 $\pm$0.0	98.49 $\pm$0.4	90.16 $\pm$2.1
EDL	77.37 $\pm$5.8	95.19 $\pm$0.3	3.25 $\pm$0.5	92.23 $\pm$1.7	61.51 $\pm$6.6
$\mathcal{I}$-EDL	83.46 $\pm$0.2	95.58 $\pm$0.0	2.74 $\pm$0.0	94.11 $\pm$0.9	68.70 $\pm$4.0
R-EDL	83.41 $\pm$0.1	95.58 $\pm$0.1	2.80 $\pm$0.0	90.91 $\pm$5.6	59.45 $\pm$13.9
DAEDL	84.12 $\pm$0.1	95.93 $\pm$0.0	2.78 $\pm$0.1	99.44 $\pm$0.2	96.34 $\pm$1.4
$\mathcal{F}$-EDL	**84.28** $\pm$0.1	**96.17** $\pm$0.1	**2.32** $\pm$0.0	**99.76** $\pm$0.1	**98.46** $\pm$0.7

Table 15: Ablation study results on $\mathcal{F}$-EDL using the CIFAR-100 dataset as the ID dataset. "Fix-$\mathbf{p}$ (U), τ" fixes $\mathbf{p}$ to a uniform vector ($\mathbf{1}/K$) and τ to 1, while "Fix-$\mathbf{p}$ (N), τ" fixes $\mathbf{p}$ to the normalized concentration parameters ($\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_1$) and τ to 1. "Fix-$\mathbf{p}$ (U)" and "Fix-$\mathbf{p}$ (N)" fix $\mathbf{p}$ but allow τ to be learned. "Fix-τ" fixes $\tau = 1$ while learning $\mathbf{p}$. The full "$\mathcal{F}$-EDL" model learns both $\mathbf{p}$ and τ.

Variant	Test.Acc.	Conf.	SVHN / TIN
Fix-$\mathbf{p}$ (U), τ	63.54 $\pm$0.8	90.95 $\pm$0.4	71.49 $\pm$4.2 / 77.57 $\pm$0.4
Fix-$\mathbf{p}$ (N), τ	64.03 $\pm$0.5	91.27 $\pm$0.4	71.73 $\pm$1.7 / 77.53 $\pm$0.4
Fix-$\mathbf{p}$ (U)	63.74 $\pm$0.3	91.04 $\pm$0.2	69.06 $\pm$3.4 / 77.40 $\pm$0.6
Fix-$\mathbf{p}$ (N)	63.77 $\pm$1.3	91.22 $\pm$0.6	69.59 $\pm$2.0 / 77.60 $\pm$0.5
Fix-τ	65.68 $\pm$0.8	92.60 $\pm$0.5	73.38 $\pm$4.6 / 78.76 $\pm$0.4
$\mathcal{F}$-EDL	**69.40** $\pm$0.2	**94.00** $\pm$0.1	**75.40** $\pm$2.3 / **80.60** $\pm$0.2

Table 16: Ablation study on the effect of different regularization terms in $\mathcal{F}$-EDL using CIFAR-10 as the ID dataset. "No Reg." denotes training without regularization, "KL-Div." indicates KL-based regularization applied to the Dirichlet parameters, "CE." refers to cross-entropy regularization applied on $\mathbf{p}$, and "Brier" represents the Brier-based regularization adopted in $\mathcal{F}$-EDL.

Regularization	Test.Acc.	Conf.	SVHN / C-100
No Reg.	91.13 $\pm$0.2	99.02 $\pm$0.4	89.32 $\pm$0.1 / 87.65 $\pm$1.0
KL-Div.	34.61 $\pm$27.8	50.69 $\pm$34.7	45.83 $\pm$21.3 / 58.93 $\pm$13.9
CE	91.11 $\pm$0.1	98.87 $\pm$0.1	89.83 $\pm$0.7 / 87.77 $\pm$0.2
Brier (Ours)	**91.19** $\pm$0.2	**99.10** $\pm$0.0	**91.20** $\pm$1.3 / **88.37** $\pm$0.3

DUQ [6] and PostNet [23]. DUQ represents deterministic uncertainty methods, while PostNet exemplifies a density-based EDL approach. $\mathcal{F}$-EDL outperforms both by a significant margin, underscoring its robustness. Results on CIFAR-100 are omitted due to the high computational cost and training instability these baselines face when applied to larger label spaces. Table 12 reports AUROC scores for OOD detection in the classical setting, confirming that $\mathcal{F}$-EDL maintains strong performance across different datasets and metrics.

Long-Tailed Setting. Table 13 presents additional OOD detection results in the long-tailed setting using the CIFAR-10-LT dataset, evaluated with AUROC. The results show that $\mathcal{F}$-EDL consistently outperforms its competitors regardless of the evaluation metric.

Noisy Setting. Table 14 reports the full results for UQ-related downstream tasks in the noisy setting using the DMNIST dataset. Metrics include the Brier score [59], a standard measure of calibration, and AUROC scores for OOD detection based on epistemic uncertainty estimates. For the Brier score, lower values indicate better calibration. The results show that $\mathcal{F}$-EDL consistently outperforms all competing methods across all metrics.

Ablation Study. Table 15 presents the additional ablation study results using CIFAR-100 as the ID dataset. The results demonstrate that the trend observed in the ablation study using the DMNIST dataset holds consistently: fixing either $\mathbf{p}$ or τ leads to degraded performance, while the full model that jointly learns both components yields the strongest results. This confirms that the observed benefits are not dataset-specific but arise from the principled generalization enabled by the FD distribution.

In addition, Table 16 presents the ablation study on the regularization term using CIFAR-10 as the ID dataset. The results indicate that the Brier-based regularization yields the best empirical performance, owing to its improved training stability and better-calibrated allocation of class probabilities.

G Additional Explanations and Results for Qualitative Study

In this section, we provide additional details on the toy experiment from Figure 1 and the posterior multimodality analysis from Section 6.3, including extended motivation, minor experimental settings, and supplementary figures.

G.1 Toy Experiment in the Figure 1

We provide additional explanations about the toy experiments described in Section 1. First, we outline the experimental setup, including the procedures followed and the uncertainty measures utilized. Second, we present the results for the aleatoric uncertainty distributions, along with corresponding figures and explanations, to complement the epistemic uncertainty distribution results shown in Section 1

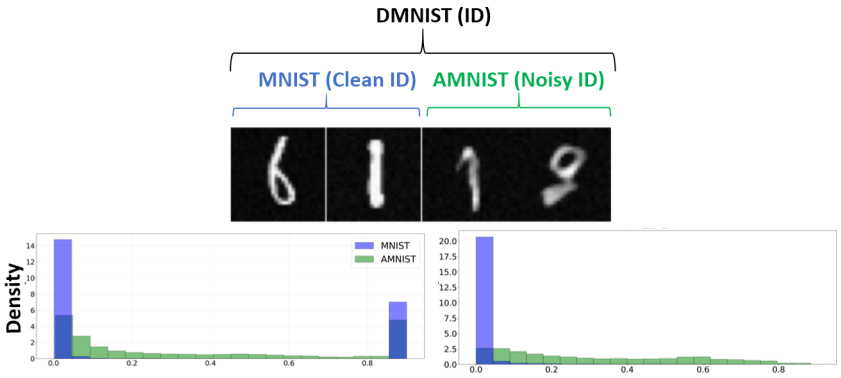
Details on Experimental Setup. To investigate the challenges EDL models face in generalizing to difficult scenarios and assess whether $\mathcal{F}$-EDL addresses these limitations, we conducted toy experiments to compare their uncertainty distributions. These experiments use DMNIST, a noisy variant of MNIST, as the ID dataset. The objective is to evaluate the ability of EDL and $\mathcal{F}$-EDL to generalize to noisy ID scenarios and reliably distinguish among three types of data: (i) clean ID data (MNIST), (ii) noisy ID data (AMNIST), and (iii) OOD data (FMNIST).

We trained both EDL and $\mathcal{F}$-EDL using the training set of DMNIST. After training, we computed the aleatoric and epistemic uncertainties for data points in the testing sets of DMNIST and FMNIST. For EDL, the aleatoric uncertainty ($\mathrm{AU}(\mathbf{x}^\star)$) and epistemic uncertainty ($\mathrm{EU}(\mathbf{x}^\star)$) for a testing example $\mathbf{x}^\star$ are defined as follows:

$$\mathrm{AU}(\mathbf{x}^\star) = 1 - \max_{k\in[K]} \mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\pi_k], \;\; \mathrm{EU}(\mathbf{x}^\star) = \frac{K}{\alpha_0},$$

where $\alpha_0 = \sum_{k=1}^K \alpha_k$ and $\mathbb{E}_{\boldsymbol{\pi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\pi_k] = \alpha_k/\alpha_0$.

For $\mathcal{F}$-EDL, we utilized the aleatoric and epistemic uncertainty measures outlined in Section 3.3. To enhance the clarity of the figures, we normalized the uncertainty estimates. For epistemic uncertainty, a logarithmic transformation was applied. Both uncertainty estimates were then scaled to the



(a) EDL Aleatoric Uncertainty **(b) $\mathcal{F}$-EDL Aleatoric Uncertainty**

Figure 5: Aleatoric uncertainty distributions with DMNIST as the ID dataset. The top row presents sample images from MNIST and AMNIST. Panels (a) and (b) display histograms depicting the aleatoric uncertainty distributions obtained by the EDL and $\mathcal{F}$-EDL (proposed) models, respectively, across these datasets.

range $[0, 1]$ as follows:

$$\mathrm{AU}(\mathbf{x}^\star) \leftarrow \frac{\mathrm{AU}(\mathbf{x}^\star) - \mathrm{AU}_{\min}}{\mathrm{AU}_{\max} - \mathrm{AU}_{\min}}, \quad \mathrm{EU}(\mathbf{x}^\star) \leftarrow \frac{\log(\mathrm{EU}(\mathbf{x}^\star)) - \log(\mathrm{EU}_{\min})}{\log(\mathrm{EU}_{\max}) - \log(\mathrm{EU}_{\min})}.$$

Here, $\mathrm{AU}_{\max}$ and $\mathrm{AU}_{\min}$ denote the maximum and minimum values of aleatoric uncertainty, while $\mathrm{EU}_{\max}$ and $\mathrm{EU}_{\min}$ represent the maximum and minimum values of epistemic uncertainty. These values were computed using the combined test sets of DMNIST and FMNIST to ensure consistent scaling across both datasets.

Results for Aleatoric Uncertainty Distributions. The middle panel of Figure 5 illustrates the aleatoric uncertainty distributions generated by the EDL model. Ideally, a UQ model should assign higher aleatoric uncertainty to noisy samples, effectively distinguishing them from clean samples.

The results highlight the limitations of EDL in handling aleatoric UQ in noisy ID scenarios. First, a substantial portion of the MNIST test set exhibits high aleatoric uncertainty, nearing the maximum value shown in the figure. This suggests that EDL struggled to effectively train on the noisy ID dataset consisting of ambiguous data points. Second, there is a significant overlap between the uncertainty distributions of MNIST and AMNIST, suggesting that the model incorrectly assigned low aleatoric uncertainty to AMNIST test points, despite their artificial construction to exhibit higher aleatoric uncertainty than MNIST.

In contrast, the bottom panel of Figure 5, which presents the aleatoric uncertainty distributions produced by $\mathcal{F}$-EDL, demonstrates notable improvements over EDL in generalizing to noisy ID scenarios and delivering robust UQ for ID data. Specifically, AMNIST demonstrates significantly higher aleatoric uncertainty compared to MNIST, with minimal overlap between their uncertainty distributions. The small overlap observed between distributions is expected, as certain AMNIST data points with relatively low aleatoric uncertainty may naturally align with the uncertainty levels of MNIST.

G.2 Multimodal Uncertainty Representations for Ambiguous Inputs.

Motivation. We investigate whether the theoretical flexibility of $\mathcal{F}$-EDL to represent multimodal class probability distribution translates to semantically meaningful and interpretable uncertainty in practice. While Theorem 4.4 shows that $\mathcal{F}$-EDL can express predictive distributions as a mixture of Dirichlet components, it is crucial to assess whether this capacity emerges empirically—especially for perceptually ambiguous inputs. In particular, we aim to verify whether the model reflects structured hesitation between multiple plausible class hypotheses instead of defaulting to a single, potentially overconfident mode (Proposition 4.6).

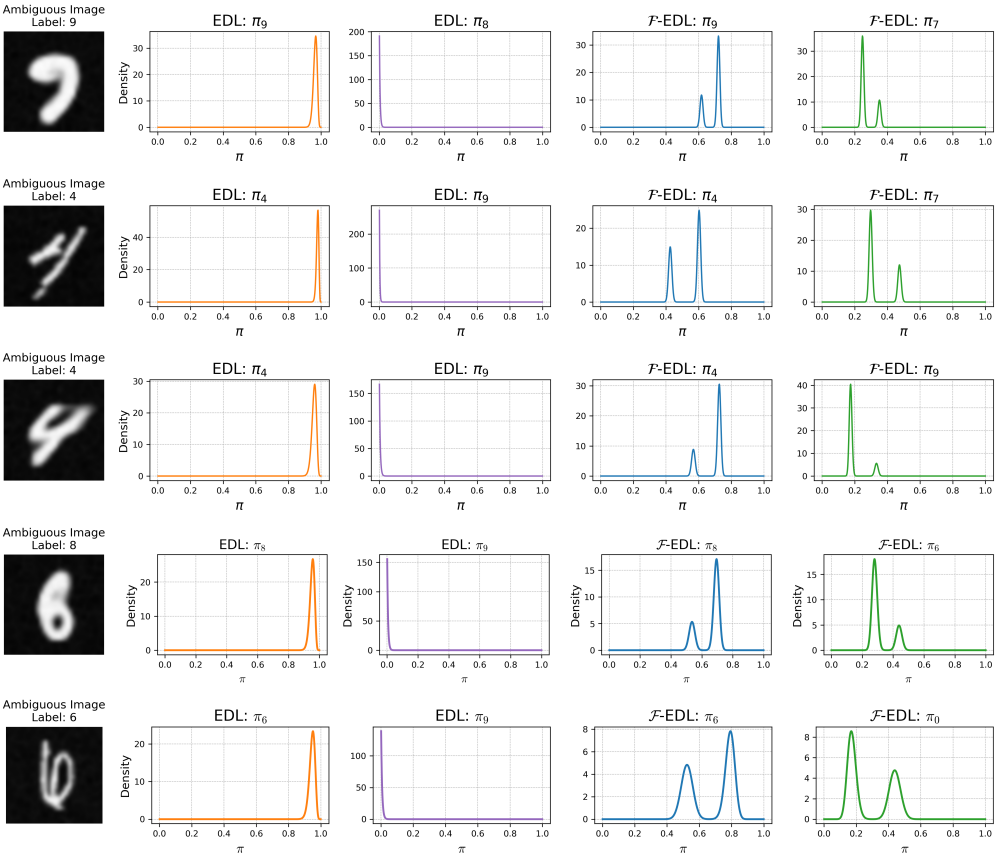


Figure 6: Additional examples supplementing Figure 3, showing posterior class probability distributions for ambiguous DMNIST inputs. Each subfigure includes: the input image, followed by the marginal distributions over the two most probable classes as predicted by EDL (second and third panels) and by $\mathcal{F}$-EDL (fourth and fifth panels).

Experimental Setup. We analyze test examples from the DMNIST dataset that exhibit visual ambiguity. For each input $\mathbf{x}^\star$, we identify the top two predicted classes based on the expected class probabilities:

$$y_1^\star = \underset{k\in[K]}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\pi}}[\pi_k(\mathbf{x}^\star)], \quad y_2^\star = \underset{k\in[K],k\neq y_1^\star}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\pi}}[\pi_k(\mathbf{x}^\star)].$$

We visualize the marginal distributions $p(\pi_{y_1^\star}|\mathbf{x}^\star)$ and $p(\pi_{y_2^\star}|\mathbf{x}^\star)$. Under the FD distribution, each marginal is a mixture of Beta distributions, parameterized by shared base evidence $\boldsymbol{\alpha}$, allocation probabilities $\mathbf{p}$, and a dispersion parameter τ:

$$p(\pi_k|\mathbf{x}^\star) = p_k \operatorname{Beta}(\alpha_k+\tau, \alpha_0-\alpha_k) + (1-p_k)\operatorname{Beta}(\alpha_k, \alpha_0-\alpha_k+\tau), k \in \{y_1^\star, y_2^\star\}.$$

These marginals exhibit multimodal behavior when the allocation probabilities $\mathbf{p}$ are dispersed, i.e., when p_k is not close to 1 and—leading to a nontrivial mixture of hypotheses. Moreover, the degree of mode separation is captured by:

$$\Delta_{\text{mode}} = \left|\frac{\tau}{\alpha_0+\tau-2}\right|,$$

which increases with larger τ, as higher dispersion sharpens and shifts each Beta component toward opposite ends, amplifying their separation.

Additional Results. In Figure 6, we present five additional examples of posterior multimodality on visually ambiguous inputs, extending the results from Figure 3. Each examples display the class

probability distributions predicted by EDL and $\mathcal{F}$-EDL for images with overlapping or unclear digit structures. For both models, we extract the marginal distributions of the two most probable classes, denoted as $\pi_{y_1^\star}$ and $\pi_{y_2^\star}$. EDL produces unimodal Beta distributions that average over competing hypotheses, often concentrating mass on a single class. This unimodality results in overconfident predictions—especially problematic in ambiguous cases. In contrast, $\mathcal{F}$-EDL produces multimodal marginals with distinct peaks, each reflecting a plausible interpretation—such as “7" and “9" for the hybrid digit. By retaining *structured* multimodality, $\mathcal{F}$-EDL preserves input ambiguity rather than collapsing it into a single smoothed belief.

This ability to model structured multimodal beliefs offers key advantages. First, it improves interpretability: multimodal outputs reveal not only that the model is uncertain but also *why*, by highlighting plausible alternatives. This property is particularly valuable in safety-critical domains like medical diagnosis, where knowing *which* alternatives the model considers plausible can inform downstream decisions. Second, this flexibility contributes to stronger UQ performance. Structured multimodality helps avoid overconfident beliefs on ambiguous inputs, thereby reducing misclassification and enhancing robustness in OOD detection. Intuitively, EDL is biased toward collapsing uncertainty into a single mode. While this may be sufficient for downstream tasks when the dominant mode aligns with the true class, it risks failure when it favors an incorrect one—leading to confident misclassification or false OOD rejection of ambiguous ID inputs. In contrast, $\mathcal{F}$-EDL preserves a mixture of multiple plausible hypotheses, preventing premature convergence and allowing the model to gradually learn and better represent the underlying uncertainty. As demonstrated in our quantitative evaluations (Section 6.2), $\mathcal{F}$-EDL consistently outperforms existing methods, particularly under ambiguity (e.g., AMNIST) or data imbalance (e.g., CIFAR-10-LT tail classes).