# Evidential Concept Bottleneck Models

Haifei Zhang*, Weixuan Xiao† Benjamin Quost‡ Marie-Helene Masson‡§
* UMR-CNRS 5516 Laboratoire Hubert Curien, Université Jean Monnet, Saint-Étienne, France
† College of Computer Science, Shandong Xiehe University, Jinan, China
‡ UMR-CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, Compiègne, France
§ IUT de l'Oise, Université de Picardie Jules Verne, Beauvais, France
haifei.zhang@univ-st-etienne.fr, weixuan.xiao@outlook.com, {benjamin.quost, mylene.masson}@hds.utc.fr

***Abstract*—Concept Bottleneck Models (CBMs) are attractive for interpretable image classification, but they often (i) ignore concept uncertainty at the label prediction stage and (ii) suffer when jointly training for concepts and labels, since label accuracy can improve at the expense of concept fidelity. We propose an Evidential Concept Bottleneck Model (EVCBM) that formalizes the concept-to-label phase within the Dempster-Shafer framework. Each concept is mapped to a learnable concept-for-class mass function, discounted according to its calibrated reliability, so that uncertain concepts contribute to ignorance rather than sharp class support. Discounted masses are pooled with a tunable linear combination of Dempster and Yager rules, and converted to class probabilities via the pignistic transform, along with an explicit uncertainty signal that supports abstention and risk control. Experiments show that under joint training, EVCBM achieves label accuracy comparable to jointly trained CBMs while consistently matching or surpassing sequentially trained CBMs in concept accuracy, improving the balance between predictive performance and concept fidelity.**

***Index Terms*—Explainable AI, Trustworthy AI, Concept Bottleneck Models, Dempster-Shafer Theory, Uncertainty**

## I. Introduction

Concept Bottleneck Models (CBMs) are a promising approach to interpretable image classification [1]. They predict human-understandable concepts and use the concepts as an intermediate representation for label prediction, producing explanations in terms of semantic attributes. This is appealing in high-stakes settings where users need both accuracy and traceable reasoning. CBMs can be trained in three ways: (i) independent, where a concept predictor is trained with concept supervision and the label predictor is then learned on ground-truth concept annotations; (ii) sequential, where the label predictor is instead trained on concepts predicted by the pretrained concept model, better matching test-time inputs; and (iii) joint, where concept prediction and label prediction are optimized simultaneously under a shared objective.

However, two practical issues limit standard CBMs. First, CBMs generally poorly manage uncertainty. The concept predictor typically outputs a vector of scores, used by the label predictor as if they were fully reliable assessments of the concepts' presence. As a result, model outputs are "confident", even when concept predictions are uncertain [2]. We argue that in high-stakes applications, it is not enough to output a label: rather, we may provide an assessment of the associated certainty, allowing to abstain when the evidence is too weak. Second, when the concept and prediction components of a CBM are jointly trained, end-to-end optimization can yield high label accuracy, but often at the expense of the concept layer drifting to a generic feature representation used by the label head [3]. As a consequence, concept prediction accuracy drops, and explanations become less faithful.

In this paper, we propose an Evidential Concept Bottleneck Model (EVCBM) that addresses both issues by replacing the concept-to-label predictor with an evidential one, rooted in the Dempster-Shafer (DS) theory [4]. The support of a concept to a given class is modeled by a parameterized mass function: this latter is discounted according to a reliability score pertaining to that concept-prediction pair, so that unreliable concepts support ignorance rather than determinate predictions. The discounted masses are pooled using a tunable combination rule [5] transformed into a probability [6]. The degree of support to ignorance provides a direct uncertainty quantification, allowing for flagging unreliable predictions and abstaining from making a decision [7]. Moreover, this evidential predictor improves joint training behavior: label prediction must pass through calibrated, discounted evidence and fusion, which reduces the tendency to bypass the bottleneck, keeping label accuracy comparable to jointly trained CBMs while preserving concept accuracy close to sequentially trained CBMs.

The main contributions of this paper are the following:

- **Uncertainty propagation from concept to label:** We propagate concept uncertainty to the label level via discounted evidence fusion, providing the mass on ignorance as an uncertainty signal that supports abstention.
- **Reliability calibration for evidence discounting:** We calibrate concept activations into reliability scores for more faithful evidence discounting operation.
- **Better joint-training balance:** The evidential interface mitigates bottleneck bypassing and preserves concept faithfulness during joint training.
- **Tunable conflict handling:** We use a single parameter to interpolate between Dempster's and Yager's combination rules, controlling how conflicting evidence is treated.

This paper is structured as follows. Section II reviews related work and background. Section III presents our EVCBM approach. Section IV reports and discusses experimental results. Section V concludes the paper. The code and the data used in this paper are publicly available on GitHub https://github.com/Haifei-ZHANG/EVCBM.

## II. Preliminaries and Related Work

### A. Concept Bottleneck Models

We consider a supervised learning setting with concept annotations. Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}_{i=1}^N$ be a dataset where $\boldsymbol{x}_i \in \mathcal{X}$ denotes an input image, $\boldsymbol{c}_i \in \{0,1\}^J$ denotes a vector of $J$ human-interpretable concepts, and $y_i \in \Omega$ denotes a label in a finite label set $\Omega = \{\omega_1, \ldots, \omega_K\}$ with $K = |\Omega|$ classes. We write $c_{ij} \in \{0,1\}$ for the $j$-th concept of sample $i$. A Concept Bottleneck Model (CBM) decomposes the decision process into two stages: a concept predictor $g_\theta : \mathcal{X} \to [0,1]^J$ produces concept scores $\hat{\boldsymbol{c}} = g_\theta(\boldsymbol{x}) = (\hat{c}_1, \ldots, \hat{c}_J)$ with $\hat{c}_j \approx \mathbb{P}(c_j = 1 \,|\, x)$; and a label predictor $f_\phi : [0,1]^J \to \Delta^{K-1}$ then maps $\hat{\boldsymbol{c}}$ to class probabilities $\hat{\boldsymbol{p}} = f_\phi(\hat{\boldsymbol{c}}) = (\hat{p}_1, \ldots, \hat{p}_K)$, with $\Delta^{K-1}$ the probability simplex over $K$ classes [1].

CBMs are commonly trained in *independent*, *sequential*, or *joint* fashions [1]. Independent and sequential (two-stage) training first optimizes the concept predictor with concept supervision, and then train the label head either on ground-truth concepts annotation $\boldsymbol{c}$ or on predicted concepts $\hat{\boldsymbol{c}} = g_\theta(\boldsymbol{x})$, while keeping $g_\theta$ fixed. Joint training instead optimizes end-to-end for $\theta$ and $\phi$ with a combined objective

$$\min_{\theta,\phi} \; \mathbb{E}_{(\boldsymbol{x},\boldsymbol{c},y)\sim\mathcal{D}} \Big[ \mathcal{L}_y\big(f_\phi(g_\theta(\boldsymbol{x})), y\big) + \lambda_c \, \mathcal{L}_c\big(g_\theta(\boldsymbol{x}), \boldsymbol{c}\big) \Big], \quad (1)$$

where $\mathcal{L}_y$ is typically a cross-entropy loss over labels, $\mathcal{L}_c$ is a concept-wise loss (e.g., binary cross-entropy across $J$ concepts), and $\lambda_c > 0$ controls the trade-off between label accuracy and concept fidelity. While joint training can improve label accuracy, it may also encourage the concept layer to behave as an unconstrained intermediate feature representation, potentially degrading concept fidelity.

Several CBM variants modify the concept-to-label interface to improve accuracy, interventions, or uncertainty handling. CEM replaces scalar concept scores with learned concept embeddings [3]; Stochastic CBMs inject randomness at the bottleneck and aggregate predictions (e.g., via Monte Carlo) [8]; ProbCBM models each concept with an explicit parametric distribution [2]; and Post-hoc CBM builds a concept interface on top of a pretrained black-box model [9]. Collectively, these methods highlight how concepts are represented and utilized by the label predictor, which significantly impacts both interpretability and robustness to uncertainty.

### B. Dempster–Shafer Theory

Dempster-Shafer (DS) theory provides a principled framework for representing and combining uncertain evidence over a finite set of hypotheses [4]. Let $\Omega = \{\omega_1, \ldots, \omega_K\}$ denote the frame of discernment (the set of all labels), and let $2^\Omega$ denote its power set. A basic belief assignment (BBA), also called a mass function, is a mapping $m : 2^\Omega \to [0,1]$ satisfying

$$\sum_{A \subseteq \Omega} m(A) = 1, \quad (2a)$$

$$m(\emptyset) = 0. \quad (2b)$$

For any subset $A \subseteq \Omega$, the mass $m(A)$ represents the degree of support committed exactly to $A$ (i.e., not allocated to any strict subset): whenever $A = \{\omega_k\}$, it corresponds to the specific support to class $\omega_k$, while $m(\Omega)$ represents ignorance. The belief and plausibility functions associated with a mass function $m$ are defined as

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B), \quad (3a)$$

$$\mathrm{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - \mathrm{Bel}(\Omega \setminus A). \quad (3b)$$

DS theory makes it possible to take into account the reliability of a source via *discounting* [4]. Given a mass function $m$ and a reliability factor $r \in [0,1]$, the discounted mass $m^{(r)}$ is given by

$$m^{(r)}(A) = r\, m(A), \;\; A \subset \Omega, \;\; A \neq \Omega, \quad (4a)$$

$$m^{(r)}(\Omega) = (1 - r) + r\, m(\Omega). \quad (4b)$$

In a nutshell, discounting attenuates specific support and transfers the remaining belief to $\Omega$, reflecting that unreliable evidence should increase ignorance rather than enforce potentially misleading class commitments.

For two mass functions $m_1$ and $m_2$ defined on the same frame $\Omega$, the conjunctive (unnormalized) combination rule is

$$(m_1 \cap m_2)(A) = \sum_{\substack{B,C \subseteq \Omega \\ B \cap C = A}} m_1(B)\, m_2(C), \;\; \forall A \subseteq \Omega. \quad (5)$$

This rule can assign non-zero mass to the empty set, interpreted as the degree of conflict:

$$\mathcal{K} \triangleq (m_1 \cap m_2)(\emptyset). \quad (6)$$

Dempster's rule [4] resolves conflict by normalization:

$$(m_1 \oplus m_2)(A) = \frac{(m_1 \cap m_2)(A)}{1 - \mathcal{K}}, \;\; A \subseteq \Omega, A \neq \emptyset, \quad (7a)$$

$$(m_1 \oplus m_2)(\emptyset) = 0; \quad (7b)$$

whereas Yager's rule [5] transfers $m(\emptyset)$ to $\Omega$:

$$(m_1 \otimes m_2)(A) = (m_1 \cap m_2)(A), \;\; A \subsetneq \Omega, A \neq \emptyset, \quad (8a)$$

$$(m_1 \otimes m_2)(\Omega) = (m_1 \cap m_2)(\Omega) + \mathcal{K}, \quad (8b)$$

$$(m_1 \otimes m_2)(\emptyset) = 0. \quad (8c)$$

To make a decision, a given mass function $m$ can be transformed into a *pignistic probability distribution* over singletons:

$$\mathrm{BetP}(\omega_k) = \sum_{\substack{A \subseteq \Omega \\ \omega \in A}} \frac{m(A)}{|A|}, \;\; \forall \omega_k \in \Omega, \quad (9a)$$

$$\hat{y} = \arg\max_{\omega_k \in \Omega} \mathrm{BetP}(\omega_k), \quad (9b)$$

where $|A|$ denotes the cardinality of $A$.

The DS theory framework presented above provides the mathematical foundation for our evidential concept-to-label predictor, which will be explained in the next section.

## III. Evidential Concept Bottleneck Models

Building on the CBM setting and DS theory preliminaries in Sec. II, we introduce EVCBM in this section.

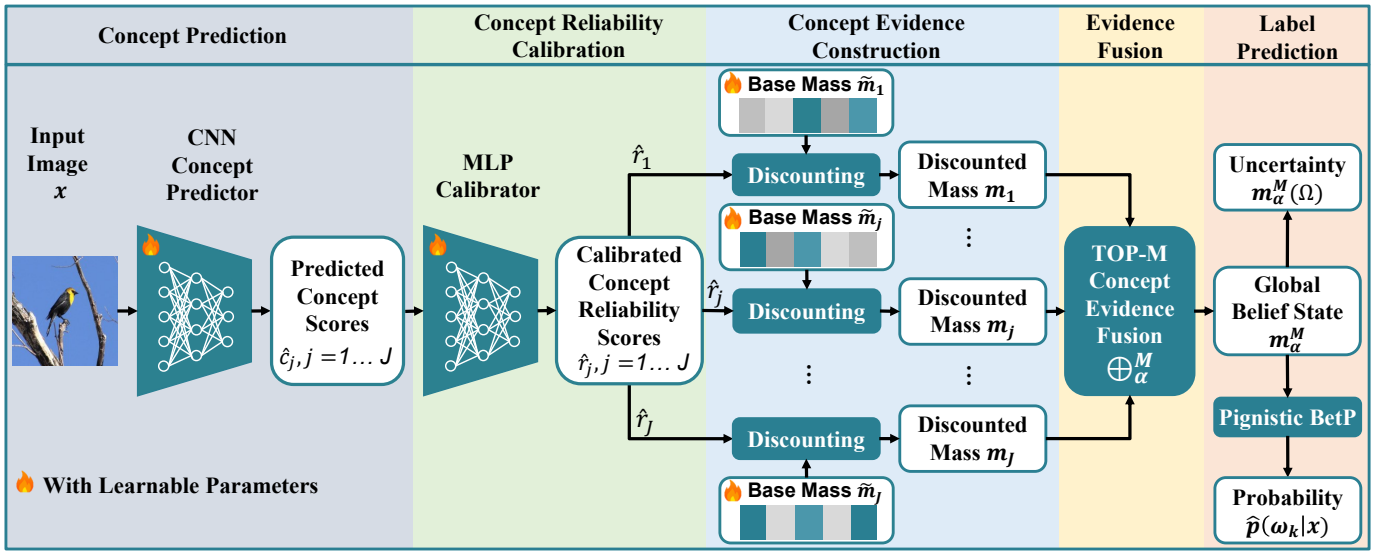



Fig. 1. Architecture of Evidential CBM.

### A. Overall Architecture

EVCBM consists of five stages, as it illustrated in Fig. 1: concept prediction, concept reliability calibration, concept evidence construction via discounting, evidence fusion, and label prediction.

Given an input image $\boldsymbol{x} \in \mathcal{X}$, the concept predictor $g_\theta : \mathcal{X} \to [0,1]^J$ outputs concept probabilities

$$\hat{\boldsymbol{c}} = g_\theta(\boldsymbol{x}) = (\hat{c}_1, \ldots, \hat{c}_J), \tag{10}$$

where $\hat{c}_j$ estimates the activation level of concept $c_j$. In practice, these probabilities can be miscalibrated, especially under a training-test distribution shift. We therefore introduce a lightweight MLP calibrator $h_\psi : [0,1]^J \to [0,1]^J$ that takes $\hat{\boldsymbol{c}}$ as input and produce concept reliability scores as output:

$$\hat{\boldsymbol{r}} = h_\psi(\hat{\boldsymbol{c}}) = (\hat{r}_1, \ldots, \hat{r}_J), \tag{11}$$

where $\hat{r}_j \in [0,1]$ quantifies how much we trust the evidence contributed by concept $c_j$ for the current input.

Each concept $c_j$ is associated with a learnable base mass function $\tilde{m}_j$ defined on $\Omega$, encoding to which extent $c_j$ supports that class *a priori*, i.e. without taking the reliability degree $\hat{r}_j$ into account. Then, for a given input, $\hat{r}_j$ is used as a discounting factor to obtain an input-dependent mass $m_j(\cdot \,|\, x)$ (Sec. III-B). The discounted concept-wise masses are then fused into a global mass $m_\alpha(\cdot \,|\, x)$ (Sec. III-C), which is mapped to class probabilities via the pignistic transform for training and inference (Sec. III-D).

For efficiency, we perform sparse fusion by selecting the top-$M$ most reliable concepts for each input:

$$\mathcal{S}(\boldsymbol{x}) \triangleq \mathrm{TopM}(\hat{\boldsymbol{r}}, M), \tag{12}$$

i.e., $\mathcal{S}(\boldsymbol{x})$ contains the indices of the $M$ largest entries in $\hat{\boldsymbol{r}}$. Only masses $\{m_j(\cdot \,|\, \boldsymbol{x})\}_{j \in \mathcal{S}(\boldsymbol{x})}$ participate in fusion, resulting in the final mass $m_\alpha^M(\cdot \,|\, \boldsymbol{x})$, which reduces computational load and limits the influence of the least reliable concepts.

### B. Concept Reliability Calibration and Evidence Construction

EVCBM converts each concept into an evidence source over the label frame $\Omega = \{\omega_1, \ldots, \omega_K\}$. This conversion relies on two ingredients: a concept-specific base mass function that encodes how the concept supports classes when it is fully reliable, and an input-dependent reliability score that discounts this support when the concept prediction is uncertain.

**Learnable base masses.** For each concept $c_j$, we learn a base mass function $\tilde{m}_j$ on $\Omega$. In this work, we restrict focal elements to singleton labels, which keeps the interpretability and makes fusion efficient. Concretely,

$$\tilde{m}_j(\{\omega_k\}) = \pi_{j,k}, \;\; k = 1, \ldots, K, \tag{13a}$$

$$\sum_{k=1}^{K} \pi_{j,k} = 1, \;\; \pi_{j,k} \geq 0. \tag{13b}$$

We parameterize $\pi_{j,\cdot}$ with a $K$-way softmax over learnable logits (one vector per concept). Intuitively, $\tilde{m}_j(\{\omega_k\})$ captures the typical support that concept $c_j$ provides for class $\omega_k$ when the concept is present and perfectly reliable.

**Reliability calibration.** Given concept scores $\hat{\boldsymbol{c}} = g_\theta(\boldsymbol{x})$, the calibrator outputs $\hat{\boldsymbol{r}} = h_\psi(\hat{\boldsymbol{c}})$. The value $\hat{r}_j \in [0,1]$ represents the degree to which concept $c_j$ should influence label inference for the current input. While $\hat{c}_j$ reflects the predicted presence of a concept, $\hat{r}_j$ reflects the reliability of using the corresponding concept evidence; this distinction is important when concept predictions are overconfident or unstable under distribution shift.

**Discounting.** We use $\hat{r}_j$ as a DS discounting factor to transform the base mass $\tilde{m}_j$ into an input-dependent mass $m_j(\cdot \,|\, \boldsymbol{x})$. Under the singleton-only base masses in Eq. (13), discounting reduces to

$$m_j(\{\omega_k\} \,|\, \boldsymbol{x}) = \hat{r}_j \, \tilde{m}_j(\{\omega_k\}), \;\; k = 1, \ldots, K, \tag{14a}$$

$$m_j(\Omega \,|\, \boldsymbol{x}) = 1 - \hat{r}_j. \tag{14b}$$

Thus, fully reliable concepts preserve their class-specific support, whereas unreliable concepts shift mass to $\Omega$, thereby supporting ignorance: when $\hat{r}_j = 0$, the mass function $m$ is vacuous, i.e., $m_j(\Omega \mid \boldsymbol{x}) = 1$; and $c_j$ does not contribute to the final decision.

### C. Evidential Fusion of Concepts

A set of top-$M$ input-dependent concept-wise masses $\{m_j(\cdot \mid \boldsymbol{x})\}_{j \in \mathcal{S}(\boldsymbol{x})}$ on the common frame $\Omega$ is fused into a global mass $m_\alpha^M(\cdot \mid \boldsymbol{x})$ using a conflict-aware combination rule that interpolates between Dempster's normalization and Yager's conflict-to-ignorance transfer.

Given two mass functions $m_a$ and $m_b$ from $\{m_j(\cdot \mid \boldsymbol{x})\}_{j \in \mathcal{S}(\boldsymbol{x})}$, let $m^D = m_a \oplus m_b$ and $m^Y = m_a \otimes m_b$ denote their combination via Dempster's rule and Yager's rule, respectively (as defined in Sec. II-B). We then define a tunable fusion rule as a linear mixture of $m^D$ and $m^Y$:

$$m_\alpha(\{\omega_k\}) = (1-\alpha)\, m^D(\{\omega_k\}) + \alpha\, m^Y(\{\omega_k\}),\ \omega_k \in \Omega, \tag{15a}$$

$$m_\alpha(\Omega) = (1-\alpha)\, m^D(\Omega) + \alpha\, m^Y(\Omega). \tag{15b}$$

where $\alpha \in [0,1]$ controls how conflict is handled: we retrieve Dempster's rule when $\alpha = 0$, and Yager's rule when $\alpha = 1$.

For multiple concepts, we apply Eq. (15) sequentially over $j \in \mathcal{S}(\boldsymbol{x})$. While Dempster's and Yager's rules are associative, their convex mixture in Eq. (15) is not guaranteed to be; we therefore pool masses in a fixed order, sorting concepts by decreasing reliability $\hat{r_j}$. Pooling $M$ concepts in $\mathcal{S}(\boldsymbol{x})$ results a global mass function $m_\alpha^M(\cdot \mid \boldsymbol{x})$ defined on the set $\Omega$ of classes.

### D. Output Layer and Loss Function

Eventually, $m_\alpha^M(\cdot \mid \boldsymbol{x})$ is converted into class probabilities via the pignistic transform (9). Under our focal-set restriction (singletons and $\Omega$), the pignistic probability simplifies to

$$\mathrm{BetP}(\omega_k \mid \boldsymbol{x}) = m_\alpha^M(\{\omega_k\} \mid \boldsymbol{x}) + \frac{m_\alpha^M(\Omega \mid \boldsymbol{x})}{|\Omega|},\ \omega_k \in \Omega, \tag{16}$$

and the predicted label is $\hat{y} = \arg\max_{\omega_k \in \Omega} \mathrm{BetP}(\omega_k \mid \boldsymbol{x})$.

We train EVCBM end-to-end with label supervision and concept supervision. Given a training set $\{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}_{i=1}^N$ with $\boldsymbol{c}_i \in \{0,1\}^J$ and $y \in \Omega$, we minimize

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \Big[ \mathcal{L}_y(\boldsymbol{x}_i, y_i) + \lambda_c\, \mathcal{L}_c(\hat{\boldsymbol{c}}_{\boldsymbol{i}}, \boldsymbol{c}_i) \Big], \tag{17a}$$

where

$$\mathcal{L}_y(\boldsymbol{x}_i, y_i) = -\log \mathrm{BetP}(y_i \mid \boldsymbol{x}_i), \tag{17b}$$

$$\mathcal{L}_c(\hat{\boldsymbol{c}}_{\boldsymbol{i}}, \boldsymbol{c}_i) = -\sum_{j=1}^{J} \Big[ c_{ij} \log \hat{c}_{ij} + (1-c_{ij}) \log(1-\hat{c}_{ij}) \Big], \tag{17c}$$

where $\hat{\boldsymbol{c}}_i = g_\theta(\boldsymbol{x}_i)$ are predicted concept probabilities and $\lambda_c > 0$ controls the trade-off between label accuracy and concept fidelity. The objective is minimized jointly over the concept predictor parameters $\theta$, calibrator parameters $\psi$, and the base-mass parameters $\{\pi_{j,k}\}, j = 1, \ldots J, k = 1, \ldots, K$.

### E. Discussion

Under our focal-set restriction (mass is assigned to singletons and $\Omega$ only), a key byproduct of EVCBM is the final mass function $m_\alpha^M(\Omega \mid \boldsymbol{x})$, which directly measures the residual uncertainty after aggregating concept evidence. This mass increases when few concepts are reliable, or when the pooled mass functions substantially disagree. It can be used for mitigating the misclassification risk, by abstaining to make a decision when the uncertainty is high, i.e. $m_\alpha^M(\Omega \mid \boldsymbol{x}) \geq \tau_\Omega$, rather than picking a potentially wrong label. More generally, outputs can be turned into set-valued predictions using standard belief and plausibility criteria in the DS theory [10].

Our fusion procedure combines concept masses by treating concepts as independent (distinct) sources of evidence [4]. In practice, this assumption is not guaranteed to hold, and concepts may be strongly correlated in CBMs [11], which can induce double-counting and overly sharp beliefs after fusion. A more principled treatment of non-distinct concept evidence would be an interesting research direction [12].

Finally, we stress that interpretability is preserved by construction. The learned base masses $\tilde{m}_j$ have a clear semantics (how reliable concepts $c_j$ support classes), while calibrated reliability degrees $\hat{r}_j$ explain how much concepts are to be trusted for a given input. In addition, concept contributions to label predictions can be quantified with existing attribution tools (e.g., SHAP) applied to the pignistic output, as explored in prior work [13].

## IV. Experiments

This section evaluates EVCBM on three concept-annotated benchmarks and compares it with standard CBMs and competitive variants. We first describe the experimental setup and then report results on concept and label prediction accuracy, followed by additional analyses and a case study.

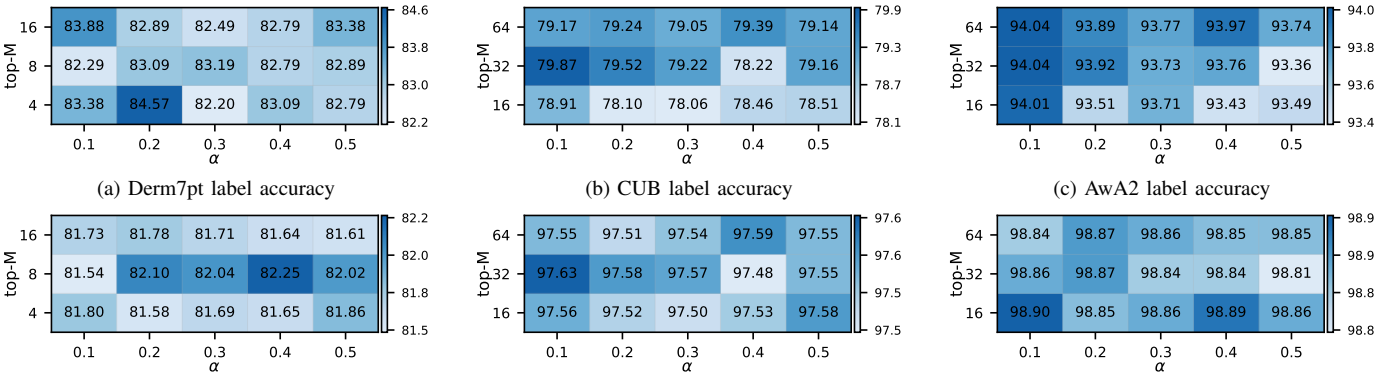
### A. Experimental Setup

We conducted evaluations on Derm7pt (7-point checklist for skin lesion malignancy) [14], CUB (Caltech-UCSD Birds-200-2011) [15], and AwA2 (Animals with Attributes 2) [16] using 5-fold cross-validation. In each fold, we split the training portion into inner training and inner validation sets (20% of the fold training data), used the inner validation set for model selection, and reported results on the held-out test split. Unless otherwise stated, we fixed the random seed to 42 and reported the mean and standard deviation across folds.

We compared against black-box ResNet50 (label-only), CBM-Sequential, CBM-Joint, CEM, and ProbCBM. All methods use a ResNet50 backbone and a batch size of 64. EVCBM was trained with AdamW (lr=$10^{-3}$), while the baselines were trained with SGD (lr=$10^{-2}$), which performed better than AdamW for these models in our setting. We trained for 50 epochs on Derm7pt, 40 on CUB, and 20 on AwA2, selecting the best checkpoint per fold based on inner validation performance. For all jointly-trained models (CBM-Joint, CEM, and EVCBM), we set $\lambda_c = 0.1$, which is good for concept-label balance according to the literature.

TABLE I
LABEL AND CONCEPT PREDICTION ACCURACY (%) OF DIFFERENT CBMS.

| | Derm7pt | | CUB | | AwA2 | |
|---|---|---|---|---|---|---|
| | Label Acc, | Concept Acc. | Label Acc. | Concept Acc. | Label Acc. | Concept Acc. |
| Resnet50 | 80.71 ± 2.92 | / | **83.19 ± 0.54** | / | 95.48 ± 0.26 | / |
| CBM-Sequential | 81.51 ± 1.96 | **82.03 ± 0.77** | 68.24 ± 0.54 | 94.75 ± 0.10 | 95.35 ± 0.29 | 99.07 ± 0.05 |
| CBM-Joint | 80.71 ± 2.08 | 63.40 ± 4.99 | 81.15 ± 0.34 | 60.25 ± 0.61 | **95.59 ± 0.25** | 63.10 ± 0.07 |
| CEM | 81.10 ± 2.20 | 69.40 ± 3.87 | 77.93 ± 0.63 | 85.96 ± 0.65 | 94.74 ± 0.34 | 96.96 ± 0.27 |
| ProbCBM | 81.80 ± 1.11 | 81.15 ± 0.47 | 70.17 ± 0.88 | 95.22 ± 0.09 | 95.48 ± 0.26 | **99.10 ± 0.05** |
| EVCBM | **84.57 ± 2.58** | 81.58 ± 0.88 | 79.87 ± 0.99 | **97.63 ± 0.13** | 94.04 ± 0.29 | 98.86 ± 0.08 |




(a) Derm7pt label accuracy (b) CUB label accuracy (c) AwA2 label accuracy

(d) Derm7pt concept accuracy (e) CUB concept accuracy (f) AwA2 concept accuracy

Fig. 2. Heatmaps of Label accuracy (top row) and Concept accuracy (bottom row) across $\alpha$ and top-$M$ for three datasets.

For EVCBM, the concept reliability calibrator $h_\psi$ was a 2-layer MLP with 128 hidden units per layer. We selected the two EVCBM-specific hyperparameters, $\alpha$ and $M$, via grid search over $(\alpha, M)$ and chose the best-performing combination based on average label accuracy across folds; see Sec. IV-C for comprehensive details.

### B. Label and Concept Prediction Accuracy

Table I summarizes label and concept accuracies on three datasets. The main pattern is a compromise between label predictive performance and concept bottleneck faithfulness. Under joint training, CBM-Joint tends to prioritize the label prediction objective and allows the concept layer change into whatever helps the classifier, which yields strong label accuracy but severely degrades concept accuracy. In contrast, CBM-Sequential protects concept fidelity by construction, yet this often comes at a noticeable cost in label accuracy, especially on fine-grained recognition, e.g., on CUB.

EVCBM reduces this trade-off. Its evidential interface makes label prediction rely on reliability-discounted concept evidence rather than unrestricted concept features. Empirically, EVCBM maintains strong concept accuracy across datasets while keeping label accuracy close to the best baselines. In particular, it delivers the best label accuracy on Derm7pt with strong concept accuracy, and on CUB, it markedly improves concept accuracy while remaining competitive in label accuracy. Overall, these results suggest that EVCBM yields more stable joint training behavior, preserving task performance without sacrificing the quality of concept explanations.

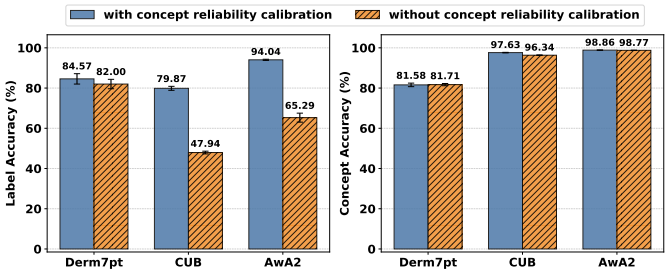



Fig. 3. Ablation study evaluating the impact of concept reliability calibration on label accuracy (left) and concept accuracy (right).

### C. Ablation Study

We conducted two ablations to study hyperparameter sensitivity and isolate key components of EVCBM. First, we ran a grid search over the conflict-handling parameter $\alpha$ and the top-$M$ concept selection. From Fig. 2, we find that fusing a moderate number of concept evidences (4 for Derm7pt, 32 for CUB and AwA2) and using a lower value of $\alpha$ (0.2 or 0.1) generally yields better results, especially on label accuracy.

We also compared EVCBM with and without the concept reliability calibrator using the best hyperparameters. In the no-calibration variant, we directly set $\hat{r}_j = \hat{c}_j$ for discounting. As shown in Fig. 3, removing the calibrator significantly reduces the label accuracy (especially on CUB and AwA2, about 30%), while concept accuracy remains nearly unchanged, suggesting that calibration mainly improves downstream evidence quality.

Blur severity: 0.00
Concept accuracy: 100%
Uncertainty $m(\Omega)$: 0.342
Label prediction $\hat{y}$: Zebra ✔

Blur severity: 0.30
Concept accuracy: 97.6%
Uncertainty $m(\Omega)$: 0.442
Label prediction $\hat{y}$: Zebra ✔

Blur severity: 0.60
Concept accuracy: 65.9%
Uncertainty $m(\Omega)$: 0.797
Label prediction $\hat{y}$: Horse ✘

Blur severity: 0.90
Concept accuracy: 62.4%
Uncertainty $m(\Omega)$: 0.976
Label prediction $\hat{y}$: Hamster ✘

Fig. 4. AwA2 case study of prediction uncertainty under increasing fog severity.

### D. Case Study

In Fig. 4, we illustrate how EVCBM reacts to increasing visual degradation on an AwA2 example (a zebra image). We gradually blur the same image, which makes the visual information less distinct and therefore makes concept prediction harder (e.g., stripes, furry, long leg, grazer, etc.). As blur severity increases, concept prediction accuracy drops significantly, and EVCBM responds by assigning more mass to $\Omega$, i.e., increasing $m(\Omega)$ as an explicit uncertainty signal.

This behavior is aligned with the evidential reasoning framework: when the input becomes ambiguous, and concepts are less reliable, the model becomes less confident about a single class. In our example, the prediction is correct under mild degradation, but at higher blur levels, the model starts to make wrong label predictions. Importantly, these failures coincide with a large increase in $m(\Omega)$, indicating that EVCBM is aware that its decision is fragile. In practice, one can set a suitable abstention threshold on $m(\Omega)$ and reject predictions when uncertainty is too high, so as to avoid incorrect high-stakes decisions.

## V. Conclusion

In this paper, we introduced EVCBM, an evidential concept bottleneck model that replaces point-estimate concept-to-label reasoning with reliability-aware evidence aggregation in the Dempster-Shafer framework. By discounting concept evidence based on calibrated reliability and fusing concept-wise masses with a tunable compromise between Dempster's and Yager's combination rules, EVCBM propagates concept uncertainty to the label level and yields an explicit uncertainty signal. Empirically, EVCBM improves the balance between label accuracy and concept fidelity under joint training, helping preserve faithful concept-based explanations while keeping label prediction performance competitive.

Several directions follow naturally from this preliminary exploratory work. First, the uncertainty signal can be further exploited for principled abstention and set-valued prediction [17]. Second, relaxing the implicit source-independence assumption by introducing dependency-aware fusion [12] or structured concept evidence models may better handle correlated concepts.

## References


[1] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[2] E. Kim, D. Jung, S. Park, S. Kim, and S. Yoon, "Probabilistic concept bottleneck models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 16 521–16 540.

[3] M. E. Zarlenga, P. Barbiero, F. Precioso, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, S. Melacci, A. Weller, P. Lio *et al.*, "Concept embedding models: Beyond the accuracy-explainability trade-off," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 21 400–21 413.

[4] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 2020.

[5] R. R. Yager, "On the dempster-shafer framework and new combination rules," *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[6] P. Smets, "Constructing the pignistic probability function in a context of uncertainty." in *UAI*, vol. 89, 1989, pp. 29–40.

[7] T. Denoeux, "Decision-making with belief functions: A review," *International Journal of Approximate Reasoning*, vol. 109, pp. 87–110, 2019.

[8] M. Vandenhirtz, S. Laguna, R. Marcinkevičs, and J. Vogt, "Stochastic concept bottleneck models," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 51 787–51 810.

[9] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[10] H. Zhang, B. Quost, and M.-H. Masson, "Cautious classifier ensembles for set-valued decision-making," *International Journal of Approximate Reasoning*, vol. 177, p. 109328, 2025.

[11] L. Heidemann, M. Monnet, and K. Roscher, "Concept correlation and its effects on concept-based models," in *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2023, pp. 4780–4788.

[12] T. Denœux, "Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence," *Artificial Intelligence*, vol. 172, no. 2-3, pp. 234–264, 2008.

[13] H. Zhang, "Shaded: Shapley value-based deceptive evidence detection in belief functions," in *International Conference on Belief Functions*. Springer, 2024, pp. 171–179.

[14] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[15] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[16] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[17] Z. Guo, Z. Wan, Q. Zhang, X. Zhao, Q. Zhang, L. M. Kaplan, A. Jøsang, D. H. Jeong, F. Chen, and J.-H. Cho, "A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning," *Information Fusion*, vol. 101, p. 101987, 2024.