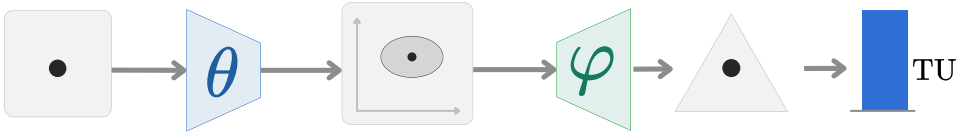
THE CREDAL INFORMATION BOTTLENECK

Dario Loi* **Gabriele Onorato**
Nicolò Bonacorsi **Francesca Meneghello** **Francesco Restuccia**
Department of Electrical and Computer Engineering
Northeastern University

ABSTRACT

The information bottleneck (IB) finds a compressed representation that preserves information relevant to a task. Traditional IB defines this compression–prediction tradeoff for a fixed joint input–output distribution. While robust IB variants account for perturbations in that distribution, they do not explicitly preserve the information needed to estimate the uncertainty at the output of the bottleneck. To address this fundamental limitation, we propose the *credal information bottleneck (CIB)*, which for the first time represents uncertainty in the data-generating distribution and the data processing, and propagates it over the bottleneck, enabling its estimation and categorization at the bottleneck output. Importantly, CIB disentangles aleatoric, data-induced and processing-induced uncertainty. This knowledge about the sources of uncertainty allows performing specific and selective interventions at deployment-time. To do so, CIB leverages the formalism of *credal sets*, i.e., convex sets of probability distributions, to characterize and propagate the uncertainty and the information about the uncertainty sources through the bottleneck. By replacing point-valued representations with credal sets, *CIB fundamentally changes the very definition of information object and information distortion*. We introduce variational CIB (VCIB) to make CIB differentiable, and prove that the VCIB objective upper-bounds the CIB up to an additive term. We further prove that every pooled-efficient IB representation incurs a strictly positive credal-distortion floor, which CIB provably overcomes. On three colored benchmarks with environment conflict, VCIB improves post-deployment worst-case accuracy by up to 6.4 percentage points over the strongest baseline.

(a) Variational Information Bottleneck



(b) Variational Credal Information Bottleneck

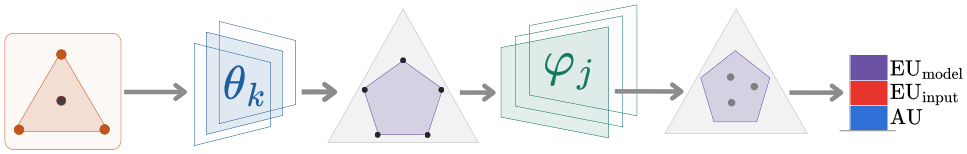


Figure 1: The IB compresses a point-valued input by optimizing a single pipeline of processing functions based on a compression–prediction tradeoff under one pooled law. The proposed CIB, instead, uses credal sets to represent the uncertainty in the input distribution and the processing functions. To implement the processing functions, **(a)** the variational IB (VIB) uses an encoder and a decoder respectively parametrized by θ and φ, while **(b)** the VCIB instantiates credal sets on the model parameters (identified by θ_k and φ_j) and the input distribution to distinguish among aleatoric uncertainty (AU), epistemic uncertainty (EU) on the input and EU on the encoder/decoder.

*Corresponding author: loi.d@northeastern.edu

1 INTRODUCTION

Learning representations that preserve task-relevant information is a fundamental problem of statistical machine learning (Bengio et al., 2013; Achille & Soatto, 2018). The information bottleneck (IB) (Tishby et al., 2000), whose variational approximation is presented in Figure 1**(a)**, formalizes this problem as a rate–relevance tradeoff, i.e., for an input X, target Y and representation Z, it seeks to preserve information about Y (maximal relevance) while limiting the information retained from X (minimal rate). Neural approximations extend this to high-dimensional data, using variational IB (VIB) (Alemi et al., 2017) and nonlinear IB (Kolchinsky et al., 2019). Robust IB methods penalize the sensitivity of representations to input perturbations (Pensia et al., 2020). Related minimax work examines worst-case inference risk under a description-rate constraint (Li et al., 2020).

Key Issues. While traditional IB selects a representation under a fixed joint distribution, in practical scenarios the data-generating distribution is unknown and the compressing–decompressing pipeline is itself uncertain (Mohajerin Esfahani & Kuhn, 2018; Gal & Ghahramani, 2016). Thus, a representation optimized for a nominal distribution may miss predictive information or encode spurious information whose correlation with the target fails under distribution shift (Federici et al., 2021; Zhou et al., 2021). Robust IB and minimax methods (Pensia et al., 2020; Li et al., 2020) do not explicitly distinguish irreducible stochasticity, i.e., the aleatoric uncertainty (AU), from uncertainty about the data-generating law and processing functions, i.e., the epistemic uncertainty (EU), when determining what information to retain. This distinction helps both compression and decision: downstream decisions must account for AU, while additional observations or model refinement may reduce EU. Thus, *the key challenge is to obtain representations and compression rates that retain the information necessary to disentangle AU and EU at the output*.

Our Approach. We introduce the *credal information bottleneck (CIB)*, a new theory for uncertainty-aware compression that for the first time preserves *both* input *and* processing uncertainty. As shown in Figure 1**(b)**, CIB departs from prior art since it allows the compressed representation's uncertainty to be separated into AU, data-induced EU (i.e., environments conflict) and processing-induced EU (i.e., model uncertainty). CIB relies on *credal sets* as mathematical backbone for modeling the EU (Abellán et al., 2006). *Thus, CIB learns a stochastic representation that is optimized to preserve a set-valued predictive target rather than a single pooled predictive law.* As such, CIB's objective is fundamentally different from prior robust IB work since it changes the information object into a set and the distortion into distortion among sets.

We provide a visual representation of the CIB approach in Figure 2. The results are obtained using the variational relaxation detailed in Section 4 on the colored MNIST dataset, where the ground truth is obtained as detailed in Appendix D. While VIB only measures total uncertainty, variational CIB (VCIB) separates its components, including EU associated with the source and the model (left panel). This distinction enables different responses: the system can abstain when model EU is high or adapt its decision with high input EU. The resulting accuracy improvement ($\uparrow$ **7.8** percentage points overall) is shown in the right panel.

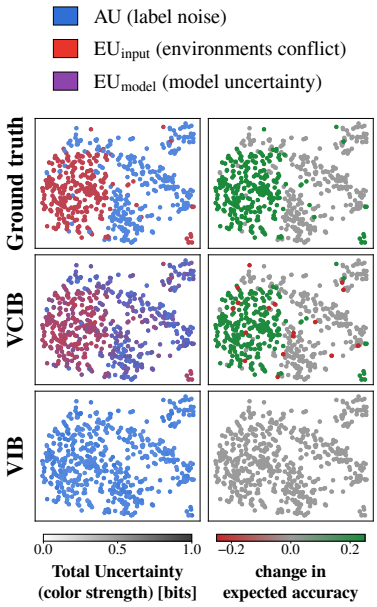


Figure 2: (**left**) Disentanglement of AU, input EU and processing EU at the task output using VIB (bottom) and VCIB (middle) against the ground-truth (top); (**right**) expected accuracy gain when accounting for the EU in making the decision.

Technical Challenges. *The first core technical challenge is preserving and disentangling uncertainty across diverse data-generating environments*. We tackle this by enforcing *credal sufficiency*, i.e., we ensure that the representation retains the full geometry of the credal set (Definition 1). *The second technical challenge is formalizing an IB objective over uncertain environments.* We address this by defining the *credal rate* and *credal distortion*. In Theorem 2, we prove that whenever an environment-specific labeling law is not almost surely determined by the pooled target, every pooled-efficient IB representation incurs a strictly positive credal-distortion floor. In contrast, CIB provably guarantees vanishing credal distortion as β increases, bounding the reconstruction error of the full credal set.

Summary of Novel Contributions

❶ We formulate the CIB, a new information bottleneck theory that preserves information about the uncertainty and its source through the bottleneck by transmitting a set-valued instead of a point-valued information, providing end-to-end uncertainty preservation;

❷ We formalize a new compression–prediction problem (Sections 4.1 and 4.2) modeling the bottleneck through credal sets of possible distributions over the input and the compressor/decompressor functions, enabling the decomposition of EU between the contributions of input and processing;

❸ We derive the VCIB, a variational relaxation of the CIB (Section 4.2). VCIB is fully differentiable and extends CIB to high-dimensional and continuous settings, which ultimately allows its implementation as neural networks (NNs);

❹ We evaluate VCIB on three colored benchmarks under shift (C-MNIST (Arjovsky et al., 2019), C-KMNIST, C-Fashion-MNIST) and show that it reduces the worst-environment credal distortion of competitive robust and Bayesian baselines by a factor of $1.8\times$ to $4.4\times$, improves worst-case accuracy after test-time environment reweighting by up to 6.4 percentage points over the strongest baseline, and detects environment conflict with an AUROC of up to 0.99, against at most 0.83 for the baselines.

Summary of Impact. By providing a new notion of information object and information distortion, the proposed CIB enables, for the first time, the estimation of the sources of uncertainty of the task output, i.e., the AU, which allows identifying spurious inputs, the EU on the input data, which isolates environment disagreement, and the EU on the compressor/decompressor functions, which quantifies the ignorance about the best processing realization. A representation trained against a pooled predictive law does not preserve environment-specific predictive distinctions. Such distinctions are not measurable from the pooled target, and Theorem 2 shows that every pooled-efficient representation incurs a strictly positive credal-distortion floor.

2 Related Work

Information Bottleneck and Robust Compression. The IB (Tishby et al., 2000) seeks a representation Z of X that balances compression, measured by $I(X;Z)$, against information relevant to a target Y, measured by $I(Y;Z)$, under a specified joint law $P(X,Y)$. The VIB and nonlinear IB make this tradeoff tractable for neural encoders using variational and nonparametric information bounds, respectively (Alemi et al., 2017; Kolchinsky et al., 2019). Conditional IB has also been applied to time-series imputation (Choi & Lee, 2024). These methods optimize with respect to a nominal training law, typically estimated from finite data. As such, discarded features may become relevant when task-relevant associations shift (Geirhos et al., 2020; Koh et al., 2021). Robust IB methods such as Fisher-information regularization target perturbations of the input (Pensia et al., 2020), while coded redundancy protects task-oriented representations against communication-channel degradation (Xie et al., 2023). Minimax distributed inference optimizes rate-constrained prediction under worst-case risk over a distributional ambiguity set (Li et al., 2020). IB-IRM (Ahuja et al., 2021) combines the IB with invariant risk minimization and seeks a representation on which a single classifier is optimal in every training environment, whereas CIB preserves environment disagreement. *These formulations do not jointly distinguish AU, input-level EU and model-level EU when selecting a representation.* Recently, Bonacorsi et al. (2026) introduced a credal rate–distortion theory. Yet, it does not account for the task relevance, which is the focus of CIB and requires a completely different formulation.

Credal Sets and Imprecise Deep Learning. Imprecise probability theory (Walley, 2000; Troffaes, 2007) models uncertainty through a *credal set* $\mathcal{C}$ of probability measures, thus separating AU from EU. While early credal models were confined to discrete variables due to the intractability of lower and upper expectations (Cozman, 2000; Antonucci & Cuzzolin, 2010; Utkin & Destercke, 2009), recent work has introduced credal priors in Bayesian NNs (BNNs) (Caprio et al., 2024a;b), credal deep ensembles (Wang et al., 2024; 2026), Random-Set NNs (Manchingal et al., 2025), and credal concept bottlenecks (Mukherjee et al., 2026). *All prior work has focused either on parameter-space priors or output sets*. Conversely, CIB formulates task-relevant latent compression with separately specified source and processing ambiguity.

Deep Uncertainty Quantification. Bayesian NNs (MacKay, 1992; Blundell et al., 2015; Rudner et al., 2022; Hobbhahn et al., 2022), deep ensembles (Lakshminarayanan et al., 2017), epistemic networks (Osband et al., 2023), evidential heads (Sensoy et al., 2018; Amini et al., 2020), or mutual information regularization (Alemi et al., 2018; Guo et al., 2024) have been proposed for uncertainty

quantification. In ensemble learning, diversity among members is enforced through information-theoretic objectives such as minimizing conditional redundancy in representation space (Rame & Cord, 2021). However, sampling and ensembling methods estimate model EU from the spread of their members without constraining what information each representation retains, while evidential models lack a rate–distortion regularizer that prunes non-robust features. Adams & Elhabian (2023) use Monte Carlo dropout and ensembles to approximate model EU from variation in predictions. However, this estimate depends on the chosen approximation to the weight posterior and can miss ambiguity across plausible source worlds. CIB models that source ambiguity separately from uncertain model parameters and accounts for both when choosing what information to retain. Its variational version, VCIB, derives uncertainty directly from the credal geometry of the output, providing robust compression alongside separate aleatoric and epistemic estimates for both the data distribution and the model.

3 Preliminary Notions

Information Bottleneck. Given an input $X \in \mathcal{X}$ and a target $Y \in \mathcal{Y}$ with nominal joint law $P_0(X, Y)$, the IB (Tishby et al., 2000) learns a stochastic encoder $P(Z \mid X)$ under the Markov chain $Y \to X \to Z$, by minimizing the rate–relevance Lagrangian $\mathcal{L}_{\text{IB}} = I(X;Z) - \beta I(Y;Z)$, where I denotes mutual information and $\beta \geq 0$ weights relevance ($I(Z;Y)$) against rate ($I(X;Z)$). Because these terms are intractable for high-dimensional data, the VIB (Alemi et al., 2017) bounds them using a parametric encoder $p_\theta(Z \mid X)$ and decoder $q_\phi(Y \mid Z)$, and a prior on the compressed representation $r(Z)$, which is typically set as Gaussian $\mathcal{N}(0, I)$. The resulting rate and relevance bounds are defined as

$$I(X;Z) \leq \mathbb{E}_{P_0(X)}\big[D_{\text{KL}}\big(p_\theta(Z \mid X) \| r(Z)\big)\big], \tag{1}$$

$$I(Y;Z) \geq H(Y) + \mathbb{E}_{P_0(X,Y)}\mathbb{E}_{p_\theta(Z|X)}\big[\log_2 q_\phi(Y \mid Z)\big], \tag{2}$$

where D_{KL} is the Kullback–Leibler (KL) divergence and $H(Y)$ is the entropy of Y.

Imprecise Probability. Imprecise probability theory (Walley, 2000) represents EU by replacing a single law with a *credal set*, that is, a non-empty, closed and convex set of probability distributions. Given distributions $P_1, \ldots, P_m$ on a common measurable space, for instance the members of an ensemble, their convex hull is the finitely generated credal set (FGCS)

$$\mathcal{P} = \text{conv}(\{P_1, \ldots, P_m\}) \coloneqq \Big\{ P_{\boldsymbol{\lambda}} = \textstyle\sum_{i=1}^m \lambda_i P_i : \boldsymbol{\lambda} \in \Delta_m \Big\}, \tag{3}$$

$$\Delta_m \coloneqq \Big\{ \boldsymbol{\lambda} \in \mathbb{R}^m_{\geq 0} : \textstyle\sum_{i=1}^m \lambda_i = 1 \Big\}, \tag{4}$$

where conv denotes the convex hull and λ_i the i-th mixture weight; we write $[m] \coloneqq \{1, \ldots, m\}$ for the index set of plausible data-generating environments. Let $\Delta(\mathcal{Y})$ be the probability simplex on a finite label set $\mathcal{Y}$ and $H(p) = -\sum_{y \in \mathcal{Y}} p(y) \log_2 p(y)$ the Shannon entropy in bits. For a credal set $\mathcal{C} \subseteq \Delta(\mathcal{Y})$, the upper and lower entropies $\overline{H}(\mathcal{C}) = \sup_{p \in \mathcal{C}} H(p)$ and $\underline{H}(\mathcal{C}) = \inf_{p \in \mathcal{C}} H(p)$ yield the decomposition $\overline{H}(\mathcal{C}) = \underline{H}(\mathcal{C}) + \big[\overline{H}(\mathcal{C}) - \underline{H}(\mathcal{C})\big]$, as in Abellán et al. (2006); Hüllermeier & Waegeman (2021), where $\underline{H}(\mathcal{C})$ represents AU, i.e., uncertainty inherent to the data distribution (e.g., in the presence of label noise), and $\overline{H}(\mathcal{C})$–$\underline{H}(\mathcal{C})$ isolates EU, the reducible uncertainty stemming from lack of confidence about a given prediction, as reflected by the disagreement among the candidate distributions in $\mathcal{C}$. Thus it follows that when $\mathcal{C}$ is a singleton, EU cannot be disentangled.

4 The Credal Information Bottleneck Method

Section 4.1 defines the CIB objective and compares it with pooled IB; Section 4.2 derives a differentiable model through variational relaxation. Section 4.3 introduces the uncertainty decomposition over the input data and model parameters. Complete proofs for all theorems are available in Appendix C.

4.1 Credal Rate–Distortion

Let (X, Y) take values in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a standard Borel space (for example, $\mathbb{R}^d$) and $\mathcal{Y}$ is a finite label set with $n = |\mathcal{Y}|$. Environment $i \in [m]$ generates data according to $P_i(dx, y) = \mu_i(dx)\, p_i(y \mid x)$, where μ_i is its input distribution and $p_i(\cdot \mid x) \in \Delta(\mathcal{Y})$ is its labeling law, and $\mathcal{P} = \text{conv}(\{P_1, \ldots, P_m\})$ as in Equation (3). The *credal target* $T(x) = (p_1(\cdot \mid x), \ldots, p_m(\cdot \mid x))$ collects the labeling laws, and $\mathcal{C}^\star(x) = \text{conv}(\{p_i(\cdot \mid x)\}_{i \in [m]})$ is the true conditional credal set. An encoder is a Markov kernel $e(dz \mid x)$ into a code space $\mathcal{Z}$, an uncountable standard Borel space such as $\mathbb{R}^d$; since e observes only x, $Y \to X \to Z$ is a Markov chain under every environment. We

write I_i and H_i for mutual information and entropy under $\mu_i(dx)\, p_i(y \mid x)\, e(dz \mid x)$, and define the *environment distortion* δ_i, the *credal rate* $\mathcal{R}$ and the *credal distortion* $\mathcal{D}$ as

$$\delta_i(e) = I_i(X;Y) - I_i(Z;Y) = I_i(X;Y \mid Z) \in [0, \log_2 n], \tag{5}$$

$$\mathcal{R}(e) = \max_{i\in[m]} I_i(X;Z), \qquad \mathcal{D}(e) = \max_{i\in[m]} \delta_i(e). \tag{6}$$

Definition 1 (*Credal sufficiency*)**.** An encoder e is *credally sufficient* if $\mathcal{D}(e) = 0$, meaning that under every environment the representation preserves all task-relevant information: $p_i(\cdot \mid X) = P_i(Y \in \cdot \mid Z)$ almost surely under P_i for all $i \in [m]$.

Credal distortion controls how well $\mathcal{C}^\star(x)$ can be recovered from the compressed code. Let $q_i(z) = P_i(Y \in \cdot \mid Z = z)$ be the Bayes decoder of environment i, $\hat{p}_i(\cdot \mid x) = \int q_i(z)\, e(dz \mid x)$ its readout, and $\widehat{\mathcal{C}}_e(x) = \mathrm{conv}(\{\hat{p}_i(\cdot \mid x)\}_{i\in[m]})$ the reconstructed credal set. Each readout satisfies $\mathbb{E}_{\mu_i}\big[d_{\mathrm{TV}}\big(p_i(\cdot \mid X), \hat{p}_i(\cdot \mid X)\big)\big] \le \sqrt{(\ln 2/2)\, \mathcal{D}(e)}$, where $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\sum_{y\in\mathcal{Y}} |p(y) - q(y)|$ is the total variation distance; if $\mathcal{D}(e) = 0$ and $\mu_1, \dots, \mu_m$ are mutually absolutely continuous, then $\widehat{\mathcal{C}}_e(X) = \mathcal{C}^\star(X)$ almost surely under every environment (Proposition 7). The CIB objective trades rate against distortion in the worst of the m generating environments, i.e.,

$$\min_e F_\beta(e), \qquad F_\beta(e) = \max_{i\in[m]} \big[I_i(X;Z) + \beta\, \delta_i(e)\big] \ge \max\{\mathcal{R}(e), \beta\, \mathcal{D}(e)\}, \qquad \beta > 0. \tag{7}$$

Because the infimum does not need to be attained over general kernels, an encoder e is called an ε-*minimizer* of F_β if $F_\beta(e) \le \inf_{e'} F_\beta(e') + \varepsilon$ for a given $\varepsilon \ge 0$.

Pooling Information Bottleneck. A conventional approach to environmental uncertainty under the classical IB (Tishby et al., 2000) is to collapse the polytope $\mathcal{P}$ into one mixture $P_{\boldsymbol{\lambda}} = \sum_{i=1}^m \lambda_i P_i$ with weights $\lambda_i > 0$. Its input marginal is $\mu_{\boldsymbol{\lambda}} = \sum_{i=1}^m \lambda_i \mu_i$ and its labeling law is the *pooled target*

$$\bar{T}_{\boldsymbol{\lambda}}(x) \coloneqq p_{\boldsymbol{\lambda}}(\cdot \mid x) = \sum_{i=1}^m \rho_i^{\boldsymbol{\lambda}}(x)\, p_i(\cdot \mid x), \qquad \rho_i^{\boldsymbol{\lambda}}(x) \coloneqq \lambda_i \frac{d\mu_i}{d\mu_{\boldsymbol{\lambda}}}(x), \tag{8}$$

where the Radon–Nikodym derivative $d\mu_i/d\mu_{\boldsymbol{\lambda}}$ makes $\rho_i^{\boldsymbol{\lambda}}(x)$ the posterior probability of environment i given x (defined $\mu_{\boldsymbol{\lambda}}$-almost everywhere). Pooled IB minimizes $I_{\boldsymbol{\lambda}}(X;Z) - \beta I_{\boldsymbol{\lambda}}(Z;Y)$, where $I_{\boldsymbol{\lambda}}$ is mutual information under $P_{\boldsymbol{\lambda}}$. An encoder with $I_{\boldsymbol{\lambda}}(X;Z) < \infty$ is *pooled-efficient* if no other encoder has pooled rate $I_{\boldsymbol{\lambda}}(X;Z)$ smaller or equal and pooled relevance $I_{\boldsymbol{\lambda}}(Z;Y)$ larger or equal, with at least one inequality strict; every finite-rate minimizer of the pooled objective is pooled-efficient for any $\beta \ge 0$ (see Appendix C).

Theorem 2 (Fundamental Limits of Pooled IB)**.** *Assume $\lambda_i > 0$ for all $i \in [m]$, and define the pooling floor and the credal-target rate*

$$\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}} \coloneqq \max_{i\in[m]} I_i\big(X; Y \mid \bar{T}_{\boldsymbol{\lambda}}(X)\big), \qquad R_T \coloneqq \max_{i\in[m]} H_i\big(T(X)\big), \tag{9}$$

with $H_i(T(X)) \coloneqq \infty$ when $T(X)$ is not discrete under μ_i. The following hold.

- *(i)* ***Pooling floor.*** *Every pooled-efficient encoder e depends on x only through $\bar{T}_{\boldsymbol{\lambda}}(x)$, $\mu_{\boldsymbol{\lambda}}$-almost everywhere; therefore $\mathcal{D}(e) \ge \mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}}$ for every rate and every $\beta \ge 0$. Moreover, $\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}} > 0$ if and only if, for some $i \in [m]$, $p_i(\cdot \mid X)$ is not μ_i-almost surely a function of $\bar{T}_{\boldsymbol{\lambda}}(X)$.*
- *(ii)* ***Achievability.*** *If e_β is an ε_β-minimizer of F_β for every $\beta > 0$ and $\varepsilon_\beta/\beta \to 0$, then $\mathcal{D}(e_\beta) \to 0$ as $\beta \to \infty$. If $R_T < \infty$, every ε-minimizer e of F_β satisfies $\mathcal{D}(e) \le (R_T + \varepsilon)/\beta$.*
- *(iii)* ***Separation.*** *If $\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}} > 0$, $R_T < \infty$ and $\beta > (R_T + \varepsilon)/\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}}$, then every ε-minimizer $e^\star$ of F_β and every pooled-efficient encoder e satisfy $\mathcal{D}(e^\star) < \mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}} \le \mathcal{D}(e)$.*

The floor in (i) is a property of the pooled target and holds for every β, whereas (ii) shows that the CIB removes it as β grows. Corollary 13 further shows that any uncertainty estimate computed from pooled-efficient encoders has an irreducible expected error whenever the true uncertainty is not a function of $\bar{T}_{\boldsymbol{\lambda}}(X)$.

4.2 Variational Credal Information Bottleneck (VCIB)

Directly evaluating F_β is computationally intractable in continuous spaces, since both the representation marginals $p_i(z)$ and the Bayes decoders q_i are unknown. As such, we introduce VCIB, which parameterizes the credal sets over models and data environments through amortized variational bounds, thus optimizing a smooth minimax surrogate (Appendix A.2). Following Caprio et al. (2024a), VCIB represents model EU by a FGCS generated by K posteriors $Q_1, \dots, Q_K$

over encoder parameters θ. A shared deterministic backbone feeds K Bayesian encoder heads $e_k(z \mid x) = \int e_\theta(z \mid x)\, dQ_k(\theta)$, each producing a diagonal Gaussian code, where e_θ is the encoder with parameters θ. Each environment i has J Bayesian decoder heads $q_{ij}(y \mid z)$, $j \in [J]$, shared across encoder heads, and each pair (i, k) has a learnable code prior r_{ik}. Environment-specific priors are required because a single shared prior implicitly penalizes the information that Z carries about the environment index (Proposition 14). We write I_{ik} and δ_{ik} for I_i and δ_i evaluated with encoder e_k, and L_i^{ref} for the held-out cross-entropy of a reference network trained on environment i without a bottleneck, which estimates $H_i(Y \mid X)$ when the labeling laws are not available in closed form.

Theorem 3 (Variational credal bound)**.** *For $i \in [m]$ and $k \in [K]$, define the variational rate $\widehat{R}_{ik}$ and the average decoder cross-entropy $\bar{L}_{ik}$ as*

$$\widehat{R}_{ik} := \mathbb{E}_{\mu_i}\big[D_{\mathrm{KL}}\big(e_k(\cdot \mid X) \,\|\, r_{ik}\big)\big], \qquad \bar{L}_{ik} := \frac{1}{J}\sum_{j=1}^{J} \mathbb{E}_{P_i, e_k}\big[-\log_2 q_{ij}(Y \mid Z)\big], \tag{10}$$

and the resulting bracket as

$$\widehat{b}_{ik} := \frac{\widehat{R}_{ik}}{\beta} + \bar{L}_{ik} - L_i^{\mathrm{ref}}, \tag{11}$$

where $\mathbb{E}_{P_i, e_k}$ is the expectation over $(X, Y) \sim P_i$ and $Z \sim e_k(\cdot \mid X)$. If these quantities are finite, then

$$\beta\, \widehat{b}_{ik} \geq I_{ik}(X; Z) + \beta\, \delta_{ik}(e_k) - \beta\, \varepsilon_i^{\mathrm{ref}}, \qquad \varepsilon_i^{\mathrm{ref}} := L_i^{\mathrm{ref}} - H_i(Y \mid X), \tag{12}$$

with equality if and only if r_{ik} equals the code marginal $\int \mu_i(dx)\, e_k(\cdot \mid x)$ and every $q_{ij}(\cdot \mid z)$ equals the Bayes decoder of environment i under e_k for almost every z.

Taking the maximum over i in Equation (12) and summing over k gives $\sum_{k=1}^{K} F_\beta(e_k) \leq \beta \sum_{k=1}^{K} \max_{i \in [m]} \widehat{b}_{ik} + K\beta \max_{i \in [m]} \varepsilon_i^{\mathrm{ref}}$. The offset $\varepsilon_i^{\mathrm{ref}}$ is the estimation error of the reference network; it can have either sign and does not depend on the trained parameters. We therefore minimize, over the parameters of the encoders, decoders and priors, the objective $\mathcal{L}_{\mathrm{VCIB}}$ or its smooth surrogate $\mathcal{L}_{\mathrm{VCIB}}^{\tau}$ with temperature $\tau > 0$:

$$\mathcal{L}_{\mathrm{VCIB}} := \sum_{k=1}^{K} \max_{i \in [m]} \widehat{b}_{ik}, \qquad \mathcal{L}_{\mathrm{VCIB}}^{\tau} := \sum_{k=1}^{K} \tau \ln \sum_{i=1}^{m} \exp\big(\widehat{b}_{ik}/\tau\big). \tag{13}$$

Since $\mathcal{L}_{\mathrm{VCIB}} \leq \mathcal{L}_{\mathrm{VCIB}}^{\tau} \leq \mathcal{L}_{\mathrm{VCIB}} + K\tau \ln m$, the surrogate upper-bounds the minimax objective and converges to it as $\tau \to 0^+$; its gradient weights each environment by the softmax of its bracket. In training, each bracket additionally carries a KL penalty on the decoder parameter posteriors, scaled by the environment sample size (Appendix A.2).

4.3 Uncertainty Quantification

Each input x yields $m \times K \times J$ predictive distributions $\widehat{p}_{i\ell}(\cdot \mid x) = \int q_{ij}(\cdot \mid z)\, e_k(dz \mid x)$, indexed by environment i and model hypothesis $\ell = (k, j) \in [K] \times [J]$. The predictive distributions define the per-environment and overall predictive credal sets

$$\widehat{\mathcal{C}}_i(x) = \mathrm{conv}\big(\{\widehat{p}_{i\ell}(\cdot \mid x)\}_{\ell \in [K] \times [J]}\big), \quad \widehat{\mathcal{C}}(x) = \mathrm{conv}\big(\textstyle\bigcup_{i=1}^{m} \widehat{\mathcal{C}}_i(x)\big). \tag{14}$$

Definition 4 (Nested credal uncertainty decomposition)**.** Let $W(\mathcal{C}) := \overline{H}(\mathcal{C}) - \underline{H}(\mathcal{C})$ be the entropy width of a credal set $\mathcal{C}$. The total uncertainty $\mathrm{TU} := \overline{H}(\widehat{\mathcal{C}}(x))$ decomposes as $\mathrm{TU} = \mathrm{AU} + \mathrm{EU}_{\mathrm{model}} + \mathrm{EU}_{\mathrm{input}}$, where

$$\mathrm{AU} := \underline{H}(\widehat{\mathcal{C}}(x)), \quad \mathrm{EU}_{\mathrm{model}} := \max_{i \in [m]} W(\widehat{\mathcal{C}}_i(x)), \quad \mathrm{EU}_{\mathrm{input}} := W(\widehat{\mathcal{C}}(x)) - \mathrm{EU}_{\mathrm{model}}. \tag{15}$$

Both EU terms are non-negative, since $\widehat{\mathcal{C}}_i(x) \subseteq \widehat{\mathcal{C}}(x)$ implies $W(\widehat{\mathcal{C}}_i(x)) \leq W(\widehat{\mathcal{C}}(x))$.

Uncertainty-guided intervention. At deployment-time, estimating the environment mixture that best models the current operating condition can reduce $\mathrm{EU}_{\mathrm{input}}$, thus improving task performance. Let $\mathcal{B}$ be a test-time batch of unlabeled inputs, and r_{ik}^{G} be the moment-matched Gaussian code prior of the environment–encoder pair (i, k) (Appendix A.6). We perform inference over the environment by computing the likelihood over the compressed representations at the outputs of the encoders. We obtain weighting factors $w_i(\mathcal{B})$ for the different worlds and use them to build a credal set $\widehat{\mathcal{C}}_{\mathcal{B}}(x)$ on the environment as follows:

$$\begin{aligned} w_i(\mathcal{B}) &\propto \exp\left(\sum_{x \in \mathcal{B}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{z \sim e_k(\cdot \mid x)}\big[\ln r_{ik}^{G}(z)\big]\right), \quad \sum_{i=1}^{m} w_i(\mathcal{B}) = 1, \\ \widehat{\mathcal{C}}_{\mathcal{B}}(x) &= \Big\{ \textstyle\sum_{i=1}^{m} w_i(\mathcal{B})\, \tilde{p}_i : \; \tilde{p}_i \in \widehat{\mathcal{C}}_i(x) \Big\}. \end{aligned} \tag{16}$$

The task prediction is then obtained as the E-admissible label (Levi, 1974) with the largest lower probability, $\hat{y}(x) = \arg\max_{y \in \mathcal{E}(\widehat{\mathcal{C}}_{\mathcal{B}}(x))} \sum_{i=1}^{m} w_i(\mathcal{B}) \min_{\ell \in [K] \times [J]} \widehat{p}_{i\ell}(y \mid x)$, where $\mathcal{E}(\mathcal{C})$ is the set of labels that are Bayes-optimal for some member of $\mathcal{C}$. The impact of the intervention procedure is shown in Figure 2 and Table 4. Implementation details shown in Appendices A.6 and A.7.

5 Experimental Results

Objectives. The evaluation focuses on three key research questions: ❶ Does VCIB overcome the pooling floor of Theorem 2? ❷ Does VCIB correctly disentangle AU and EU? ❸ Does CIB uncertainty disentanglement translate into gains in task performance at deployment?

Setup and metrics. We use Colored-MNIST (C-MNIST) (Arjovsky et al., 2019), Colored-KMNIST (C-KMNIST) and Colored Fashion-MNIST (C-Fashion-MNIST). Both C-KMNIST and C-Fashion-MNIST follow the exact same two-environment construction and binary shape-color correlation rule as C-MNIST (detailed in Appendix D), substituting the underlying digit images with Kuzushiji-MNIST (Clanuwat et al., 2018) characters and Fashion-MNIST (Xiao et al., 2017) items, respectively. All methods share the same backbone (Appendix A.1). We compare against VIB (Alemi et al., 2017), IB-IRM (Ahuja et al., 2021), credal Bayesian deep learning (CBDL) (Caprio et al., 2024a), a 5-member ensemble of VIB (Ens-VIB), and a variant of VCIB with credal sets on the model only, pooled over environments (VCIB_{mod}). We also report the rate-unconstrained limit VCIB ($\beta \to \infty$). Each method with a bottleneck is evaluated at the value of β in its grid whose training rate is closest to 2 bits; instead, CBDL and VCIB ($\beta \to \infty$) have no rate constraint. We report the worst-environment distortion $\mathcal{D}$, the worst-environment accuracy (WC-Acc) and the area under the risk–coverage curve (RC-AUC). Each uncertainty component is evaluated by the AUROC against its own signal: EU_{input} against inputs on which the environments disagree ($\text{AUROC}_{\text{conflict}}$), EU_{model} against under-trained inputs ($\text{AUROC}_{\text{model}}$), and AU against out-of-support inputs ($\text{AUROC}_{\text{OOD}}$). After deployment, we reweight the environments on unlabeled 100-element batches, and report the resulting worst-case accuracy ($\text{WC-Acc}_{\text{int}}$) and regret under intervention, defined as

$$\text{WC-Reg}_{\text{int}} = \max_{i \in [m]} \left(\text{Acc}_i^* - \text{Acc}_i \right) , \tag{17}$$

where Acc_i^* denotes the optimal-Bayes accuracy achievable on environment i, and Acc_i is the accuracy achieved by the model after batch-level environment reweighting. A lower regret indicates that the post-intervention decision closely matches the best possible predictor for that environment. VCIB uses Equation (16) to perform test-time intervention. The baselines use calibrated label-shift expectation maximization (EM) (Alexandari et al., 2020). The results are averaged over 10 seeds. Each method is compared with VCIB through a paired Wilcoxon test with Holm–Bonferroni correction (Wilcoxon, 1992; Holm, 1979). The details about the encoder/decoder NN implementation, the baselines and the evaluation methodology are reported in Appendix A.

Rate–distortion plane. Figure 3 depicts the worst-environment distortion against the worst-environment rate on C-MNIST. VCIB and VIB curves are reported over their β grids, together with the exact CIB and pooled IB. At rates smaller than 0.7 bits, VCIB, VIB and the exact CIB present the same distortion. Above this rate, the exact pooled IB curve saturates at the floor $\mathcal{D}_{\boldsymbol{\lambda}}^{\text{pool}} \approx 0.2$ bits defined in Theorem 2, thus the shaded region below it cannot be reached by any pooled code. VIB saturates slightly above the floor (0.24 bits distortion), and its rate does not exceed 1.5 bits despite increasing values of β. VCIB crosses the pooled IB floor at about 1 bit and reaches about 0.05 bit distortion at a rate of 2 bits, in agreement with Table 1. The exact CIB curve reaches zero distortion at about 2 bits, consistent with Lemma 5 (Appendix C).

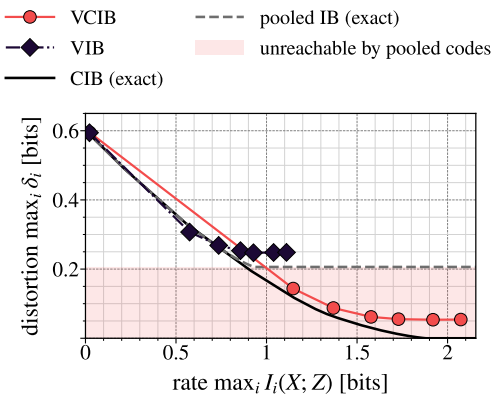


Figure 3: Rate–distortion plane on C-MNIST obtained by sweeping $\beta \in [2, 16]$. VCIB breaks the distortion lower bound of pooled bottlenecks.

Overcoming the pooling floor. Table 1 reports the results of the comparison at a 2-bit budget. The pooled methods (VIB, Ens-VIB, CBDL and VCIB_{mod}) reach a worst-environment distortion $\mathcal{D}$

Table 1: Performance at a 2-bit rate budget: for each method the results are obtained at the β of its grid whose training rate is closest to 2 bits. $\mathcal{R}$: rate of one code; $\bar{\mathcal{R}}_{\text{joint}}$: upper bound on the rate of all K codes (Kolchinsky & Tracey, 2017). **Bold**: best mean and every method not significantly different from it; ▲/▼: significantly better/worse than VCIB. Mean ± standard deviation over ten seeds.

Dataset	Method	β	$\mathcal{R}$	$\bar{\mathcal{R}}_{\text{joint}}$	$\mathcal{D}\downarrow$	WC-Acc↑	WC-Acc$_{\text{int}}$ ↑	WC-Reg$_{\text{int}}$ ↓	RC-AUC↓
C-MNIST	VIB	16	$1.115_{\pm.026}$	$1.341_{\pm.035}$	$0.248_{\pm.006}$▼	$\mathbf{0.810}_{\pm.006}$	$0.810_{\pm.006}$▼	$0.091_{\pm.004}$▼	$\mathbf{0.104}_{\pm.003}$
	Ens-VIB	16	$1.140_{\pm.015}$	$2.436_{\pm.071}$	$0.246_{\pm.006}$▼	$\mathbf{0.813}_{\pm.005}$	$0.812_{\pm.005}$▼	$0.088_{\pm.003}$▼	$\mathbf{0.104}_{\pm.004}$▲
	IB-IRM	12	$0.214_{\pm.091}$	$0.366_{\pm.160}$	$0.536_{\pm.022}$▼	$0.710_{\pm.020}$▼	$0.717_{\pm.018}$▼	$0.184_{\pm.019}$▼	$0.175_{\pm.028}$▼
	CBDL	–	–	–	$0.240_{\pm.007}$▼	$\mathbf{0.812}_{\pm.005}$	$0.812_{\pm.005}$▼	$0.089_{\pm.003}$▼	$\mathbf{0.104}_{\pm.003}$
	VCIB$_{\text{mod}}$	16	$1.149_{\pm.019}$	$2.023_{\pm.107}$	$0.243_{\pm.008}$▼	$\mathbf{0.811}_{\pm.005}$	$0.811_{\pm.005}$▼	$0.090_{\pm.003}$▼	$\mathbf{0.104}_{\pm.004}$
	VCIB($\beta\to\infty$)	–	–	–	$\mathbf{0.043}_{\pm.004}$▲	$\mathbf{0.809}_{\pm.006}$	$\mathbf{0.883}_{\pm.005}$▲	$\mathbf{0.017}_{\pm.002}$▲	$\mathbf{0.106}_{\pm.004}$
	VCIB (ours)	16	$2.065_{\pm.021}$	$3.430_{\pm.065}$	$0.055_{\pm.006}$	$\mathbf{0.807}_{\pm.010}$	$0.873_{\pm.005}$	$0.026_{\pm.006}$	$0.106_{\pm.004}$
C-KMNIST	VIB	16	$1.339_{\pm.028}$	$1.752_{\pm.045}$	$0.315_{\pm.012}$▼	$0.776_{\pm.011}$	$0.774_{\pm.011}$▼	$0.126_{\pm.008}$▼	$0.116_{\pm.004}$▲
	Ens-VIB	16	$1.404_{\pm.020}$	$4.081_{\pm.077}$	$0.317_{\pm.008}$▼	$\mathbf{0.781}_{\pm.009}$	$0.778_{\pm.009}$▼	$0.121_{\pm.005}$▼	$\mathbf{0.110}_{\pm.002}$▲
	IB-IRM	12	$0.084_{\pm.030}$	$0.125_{\pm.058}$	$0.586_{\pm.009}$▼	$0.596_{\pm.036}$▼	$0.618_{\pm.034}$▼	$0.283_{\pm.036}$▼	$0.291_{\pm.033}$▼
	CBDL	–	–	–	$0.312_{\pm.010}$▼	$0.777_{\pm.007}$▼	$0.777_{\pm.007}$▼	$0.123_{\pm.004}$▼	$0.114_{\pm.003}$▲
	VCIB$_{\text{mod}}$	16	$1.398_{\pm.023}$	$3.006_{\pm.100}$	$0.330_{\pm.010}$▼	$0.769_{\pm.009}$▼	$0.768_{\pm.009}$▼	$0.132_{\pm.006}$▼	$0.116_{\pm.004}$▲
	VCIB($\beta\to\infty$)	–	–	–	$\mathbf{0.157}_{\pm.012}$▲	$\mathbf{0.778}_{\pm.007}$	$\mathbf{0.830}_{\pm.006}$▲	$\mathbf{0.067}_{\pm.007}$▲	$0.122_{\pm.005}$
	VCIB (ours)	12	$2.041_{\pm.065}$	$3.563_{\pm.296}$	$0.177_{\pm.009}$	$\mathbf{0.780}_{\pm.005}$	$0.821_{\pm.006}$	$0.079_{\pm.006}$	$0.123_{\pm.003}$
C-Fashion-MNIST	VIB	16	$1.198_{\pm.041}$	$1.505_{\pm.053}$	$0.314_{\pm.006}$▼	$\mathbf{0.768}_{\pm.006}$	$0.766_{\pm.006}$▼	$0.136_{\pm.005}$▼	$\mathbf{0.115}_{\pm.004}$▲
	Ens-VIB	16	$1.244_{\pm.045}$	$2.766_{\pm.113}$	$0.316_{\pm.005}$▼	$\mathbf{0.772}_{\pm.006}$	$0.769_{\pm.006}$▼	$0.133_{\pm.004}$▼	$\mathbf{0.113}_{\pm.004}$▲
	IB-IRM	12	$0.419_{\pm.106}$	$0.671_{\pm.140}$	$0.483_{\pm.037}$▼	$0.708_{\pm.043}$▼	$0.735_{\pm.016}$▼	$0.166_{\pm.015}$▼	$0.161_{\pm.020}$▼
	CBDL	–	–	–	$0.323_{\pm.007}$▼	$\mathbf{0.767}_{\pm.005}$	$0.763_{\pm.006}$▼	$0.139_{\pm.004}$▼	$\mathbf{0.113}_{\pm.004}$▲
	VCIB$_{\text{mod}}$	16	$1.200_{\pm.026}$	$2.012_{\pm.066}$	$0.310_{\pm.007}$▼	$\mathbf{0.770}_{\pm.006}$	$0.770_{\pm.006}$▼	$0.132_{\pm.002}$▼	$0.116_{\pm.004}$▲
	VCIB($\beta\to\infty$)	–	–	–	$\mathbf{0.152}_{\pm.004}$▲	$\mathbf{0.772}_{\pm.008}$	$\mathbf{0.836}_{\pm.005}$	$\mathbf{0.066}_{\pm.002}$	$0.122_{\pm.004}$
	VCIB (ours)	16	$1.960_{\pm.033}$	$3.291_{\pm.069}$	$0.165_{\pm.007}$	$\mathbf{0.773}_{\pm.007}$	$\mathbf{0.834}_{\pm.005}$	$\mathbf{0.068}_{\pm.004}$	$0.123_{\pm.003}$

between 0.24 and 0.33 bit on every dataset, while IB-IRM reaches between 0.48 and 0.59 bit. VCIB lowers $\mathcal{D}$ to values between 0.055 and 0.177, which is 1.8 to 4.4 times lower than the best baseline. At 2 bits, VCIB is close to its rate-unconstrained limit VCIB($\beta\to\infty$), with a gap in $\mathcal{D}$ of at most 0.02. At the grid point closest to 2 bits, the pooled baselines use between 1.1 and 1.4 bits, and VCIB uses between 1.9 and 2.1 bits.

Deployment-time adaptation. Before intervention, the worst-case accuracy of VCIB approaches the best baseline on all three datasets. After intervention based on the disentangled uncertainty sources, VCIB improves WC-Acc$_{\text{int}}$ over the best pooled baseline by 6.1, 4.3 and 6.4 percentage points on C-MNIST, C-KMNIST and C-Fashion-MNIST, respectively, and reduces WC-Reg$_{\text{int}}$ by a factor of 1.5× to 3.4×. In contrast, the accuracy of VIB, Ens-VIB, CBDL and VCIB$_{\text{mod}}$ changes by at most 0.4 percentage points after intervention. For selective prediction, VCIB is slightly worse than the best baseline, with an RC-AUC higher by at most 0.013 on every dataset.

Table 2: Uncertainty disentanglement (AUROC). AUROC$_{\text{conflict}}$: EU$_{\text{input}}$ against inputs on which the environments disagree; AUROC$_{\text{model}}$: EU$_{\text{model}}$ against under-trained inputs (the training images for digits 2 and 7 are reduced to 2% maintaining the same β); AUROC$_{\text{OOD}}$: AU against out-of-distribution (OOD) inputs. Methods without the split use the best available uncertainty cue (VIB: TU; Ens-VIB: mutual information; CBDL and VCIB$_{\text{mod}}$: EU$_{\text{model}}$). **Bold**: best mean and methods performing similarly; ▲/▼: much better/worse than VCIB. Mean ± standard deviation over 10 seeds.

Dataset	Method	AUROC$_{\text{conflict}}$ ↑	AUROC$_{\text{model}}$ ↑	AUROC$_{\text{OOD}}$ ↑
C-MNIST	VIB	$0.827_{\pm.013}$▼	$0.600_{\pm.035}$	$\mathbf{0.661}_{\pm.065}$
	Ens-VIB	$0.224_{\pm.011}$▼	$0.642_{\pm.019}$	$\mathbf{0.681}_{\pm.042}$
	IB-IRM	$0.616_{\pm.116}$▼	$0.607_{\pm.024}$	$0.540_{\pm.020}$▼
	CBDL	$0.299_{\pm.037}$▼	$0.640_{\pm.040}$	$0.620_{\pm.087}$▼
	VCIB$_{\text{mod}}$	$0.203_{\pm.026}$▼	$0.558_{\pm.012}$▼	$\mathbf{0.687}_{\pm.060}$
	VCIB($\beta\to\infty$)	$\mathbf{0.994}_{\pm.001}$▲	$\mathbf{0.759}_{\pm.037}$▲	$\mathbf{0.689}_{\pm.062}$
	VCIB (ours)	$0.992_{\pm.001}$	$0.608_{\pm.043}$	$\mathbf{0.729}_{\pm.049}$
C-KMNIST	VIB	$0.777_{\pm.020}$▼	$0.599_{\pm.041}$	$0.480_{\pm.070}$▼
	Ens-VIB	$0.294_{\pm.016}$▼	$0.637_{\pm.009}$▲	$0.447_{\pm.034}$▼
	IB-IRM	$0.546_{\pm.107}$▼	$0.531_{\pm.017}$▼	$0.432_{\pm.063}$▼
	CBDL	$0.318_{\pm.021}$▼	$0.599_{\pm.021}$	$0.444_{\pm.035}$▼
	VCIB$_{\text{mod}}$	$0.256_{\pm.026}$▼	$0.571_{\pm.012}$	$0.435_{\pm.073}$▼
	VCIB($\beta\to\infty$)	$\mathbf{0.949}_{\pm.007}$▲	$\mathbf{0.723}_{\pm.018}$▲	$\mathbf{0.635}_{\pm.058}$
	VCIB (ours)	$0.935_{\pm.015}$	$0.590_{\pm.028}$	$\mathbf{0.661}_{\pm.054}$
C-Fashion-MNIST	VIB	$0.794_{\pm.014}$▼	$0.456_{\pm.015}$▼	$\mathbf{0.624}_{\pm.067}$
	Ens-VIB	$0.347_{\pm.019}$▼	$0.530_{\pm.012}$	$0.598_{\pm.034}$▼
	IB-IRM	$0.722_{\pm.046}$▼	$0.451_{\pm.017}$▼	$\mathbf{0.642}_{\pm.077}$
	CBDL	$0.340_{\pm.047}$▼	$0.483_{\pm.042}$▼	$0.488_{\pm.076}$▼
	VCIB$_{\text{mod}}$	$0.230_{\pm.014}$▼	$\mathbf{0.559}_{\pm.011}$▲	$0.534_{\pm.069}$▼
	VCIB($\beta\to\infty$)	$\mathbf{0.957}_{\pm.003}$▲	$0.480_{\pm.039}$▼	$\mathbf{0.671}_{\pm.044}$
	VCIB (ours)	$0.940_{\pm.004}$	$0.540_{\pm.020}$	$\mathbf{0.672}_{\pm.029}$

Uncertainty disentanglement. Table 2 reports the detection AUROC of each component. $\mathrm{EU}_{\mathrm{input}}$ detects environment conflict with an AUROC between 0.93 and 0.99. The best baseline reaches at most 0.83 (VIB total uncertainty). The epistemic scores of Ens-VIB, CBDL and $\mathrm{VCIB}_{\mathrm{mod}}$ reach an AUROC between 0.20 and 0.35, below chance, so they assign lower uncertainty to conflicting inputs than to the others. AU detects out-of-support inputs with the best mean AUROC on all three datasets. $\mathrm{EU}_{\mathrm{model}}$ is the weakest component at 2 bits. VCIB is not significantly different from VIB, Ens-VIB, IB-IRM and CBDL on C-MNIST, and is worse than Ens-VIB on C-KMNIST. On C-Fashion-MNIST, VCIB is better than VIB, IB-IRM and CBDL but worse than $\mathrm{VCIB}_{\mathrm{mod}}$. On C-MNIST and C-KMNIST, VCIB($\beta \to \infty$) reaches 0.759 and 0.723, whereas on C-Fashion-MNIST the ordering is reversed. On the first two datasets this is consistent with a rate–distortion tradeoff: under a finite rate budget, the code retains task-discriminative and environment-disagreement information before the residual information that expresses parameter uncertainty, so the capacity available for $\mathrm{EU}_{\mathrm{model}}$ decreases as β decreases.

6 Conclusions

We introduced the CIB, an information-theoretic framework that preserves set-valued uncertainty together with task-relevant information through rate-constrained bottlenecks. By replacing nominal pooled targets with credal sets, CIB provably overcomes the strictly positive credal distortion floor inherent to pooled representations. Its variational realization, VCIB, enables end-to-end differentiable training and decomposes output uncertainty into aleatoric, environment-conflict, and model-related components, facilitating effective unsupervised test-time adaptation.

Summary of technical results. We proved that every pooled-efficient IB encoder factors through the pooled target $\bar{T}_{\boldsymbol{\lambda}}(X)$ and therefore incurs a credal distortion of at least $\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}}$ at every rate and every β; this floor is strictly positive whenever some environment's labeling law is not almost surely a function of the pooled target (Theorem 2). When $R_T < \infty$, CIB near-minimizers instead satisfy $\mathcal{D} \leq (R_T + \varepsilon)/\beta$, which gives strict separation from pooled codes for $\beta > (R_T + \varepsilon)/\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}}$. Consequently, an uncertainty estimate computed from pooled-efficient encoders has an irreducible expected error whenever the true uncertainty is not a function of $\bar{T}_{\boldsymbol{\lambda}}(X)$ (Corollary 13). The VCIB objective upper-bounds the CIB objective up to an additive term set by the reference-network error (Theorem 3). At a 2-bit budget on C-MNIST, C-KMNIST and C-Fashion-MNIST, VCIB lowers the worst-environment distortion by $1.8\times$ to $4.4\times$ relative to the best baseline and detects environment conflict with an AUROC between 0.93 and 0.99, against at most 0.83 for the baselines. After label-free intervention, it raises worst-case accuracy over the best pooled baseline by 6.1, 4.3 and 6.4 percentage points, reducing worst-case regret by $1.5\times$ to $3.4\times$. $\mathrm{EU}_{\mathrm{model}}$ remains the weakest component at finite rate.

Limitations and Future Directions. While VCIB establishes a rigorous information-theoretic formulation and an effective variational framework for set-valued uncertainty compression, several promising extensions remain for future research. Currently, our formulation focuses on finitely generated credal sets over m distinct observed environments. Extending the variational objective to accommodate continuous distributions of environments or non-parametric uncertainty sets represents a natural – yet not trivial – theoretical progression. In addition, our deployment-time adaptation uses an unsupervised likelihood weighting scheme across code priors to resolve environment uncertainty at test time. Investigating alternative decision rules, such as active sampling or partial human-in-the-loop verification when both AU and EUs are elevated, could broaden applicability in safety-critical deployments. Finally, while our focus is on foundational classification under distribution shift where exact information limits can be established, applying the CIB principle to structured output regimes, such as dense segmentation or sequential decision-making, deserves a separate investigation.

Reproducibility Statement. To enable the reproducibility of our theoretical and empirical findings, complete measure-theoretic proofs for all propositions, lemmas, and theorems (including Theorems 2 and 3) are provided in Appendix C. The exact data generation processes, environment configurations, and closed-form labeling laws for the semi-synthetic benchmarks are detailed in Appendix D. Comprehensive implementation specifications including model architectures, optimization hyperparameters, Bayesian prior and posterior parameterizations, training protocols, evaluation metrics, and hardware details are documented in Appendix A. Furthermore, full source code along with configuration files, scripts, and seeded execution routines to replicate all tables and figures will be made publicly available upon acceptance.

Disclosure of AI Use. We used generative AI tools to assist with the experimental evaluation, including code implementation and experiment execution, running the experiments under our direction, producing figures and tables. We also used them to improve the readability of the text and detect typos; no text appearing in the manuscript was entirely AI-generated. The research questions, the theory and the choice of experiments are entirely the authors'. All AI-assisted material were manually reviewed and verified by the authors, who take full responsibility for the content of this paper.

ACKNOWLEDGEMENTS

This work was supported in part by NSF under Grants CNS-2312875 and OAC-2530896, DARPA under Cooperative Agreement D25AC00374-00, ONR under Award N00014-23-1-2221, and AFOSR under Award FA9550-23-1-0261.

REFERENCES

J. Abellán, G.J. Klir, and S. Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1):29–44, 2006. doi: 10.1080/03081070500473490. URL https://doi.org/10.1080/03081070500473490.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.*, 19:50:1–50:34, 2018. URL https://jmlr.org/papers/v19/17-646.html.

Jadie Adams and Shireen Y Elhabian. Fully Bayesian VIB-DeepSSM. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 346–356. Springer, 2023. doi: 10.1007/978-3-031-43898-1_34.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=jlchsFOLfeF.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. Uncertainty in the variational information bottleneck. In *Uncertainty in Deep Learning Workshop at the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018. URL https://arxiv.org/abs/1807.00906.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232. PMLR, 2020. URL http://proceedings.mlr.press/v119/alexandari20a.html.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

Alessandro Antonucci and Fabio Cuzzolin. Credal sets approximation by lower probabilities: Application to credal networks. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann (eds.), *Computational Intelligence for Knowledge-Based Systems Design*, pp. 716–725, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-14049-5. URL https://doi.org/10.1007/978-3-642-14049-5_73.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Koenraad M R Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127, jun 2007. doi: 10.1088/1751-8113/40/28/S18. URL https://doi.org/10.1088/1751-8113/40/28/S18.

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

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015. doi: 10.5555/3045118.3045290. URL https://dl.acm.org/doi/10.5555/3045118.3045290.

Nicolò Bonacorsi, Michele Caprio, Francesca Meneghello, and Francesco Restuccia. Credal Rate-Distortion Theory. *Proceedings of the IEEE Information Theory Workshop (IEEE ITW)*, 2026.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal Bayesian deep learning. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=4NHF9AC5ui.

Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=AH5KwUSsln.

MinGyu Choi and Changhee Lee. Conditional information bottleneck approach for time series imputation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=K1mcPiDdOJ.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(00)00029-1. URL https://www.sciencedirect.com/science/article/pii/S0004370200000291.

Marco Federici, Ryota Tomioka, and Patrick Forré. An information-theoretic approach to distribution shifts. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=GrZmKDYCp6H.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. doi: 10.1038/S42256-020-00257-Z. URL https://doi.org/10.1038/s42256-020-00257-z.

Robert M. Gray. *Entropy and Information Theory*. Springer Publishing Company, Incorporated, 2nd edition, 2011. ISBN 9781441979698.

Ling Guo, Hao Wu, Yan Wang, Wenwen Zhou, and Tao Zhou. IB-UQ: Information bottleneck based uncertainty quantification for neural function regression and neural operator learning. *Journal of Computational Physics*, 510:113089, 2024. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2024.113089. URL https://www.sciencedirect.com/science/article/pii/S0021999124003383.

Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with Bayesian deep networks. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 822–832. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/hobbhahn22a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021. doi: 10.1007/S10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

Artemy Kolchinsky and Brendan D. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017. doi: 10.3390/E19070361. URL https://doi.org/10.3390/e19070361.

Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12), 2019. ISSN 1099-4300. doi: 10.3390/e21121181. URL https://www.mdpi.com/1099-4300/21/12/1181.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Isaac Levi. On indeterminate probabilities. *J. Philos.*, 71(13):391, July 1974.

Cheuk Ting Li, Xiugang Wu, Ayfer Özgür, and Abbas El Gamal. Minimax learning for distributed inference. *IEEE Transactions on Information Theory*, 66(12):7929–7938, 2020. doi: 10.1109/TIT.2020.3029182.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, 1992. doi: 10.1162/NECO.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pdjkikvCch.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018. doi: 10.1007/s10107-017-1172-1.

Tanmoy Mukherjee, Thomas Bailleux, Pierre Marquis, and Zied Bouraoui. Credal concept bottleneck models for epistemic–aleatoric uncertainty decomposition. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42961–42980, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1991. URL https://aclanthology.org/2026.acl-long.1991/.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 2795–2823. Curran Associates, Inc., 2023. doi: 10.52202/

075280-0124. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/07fbde96bee50f4e09303fd4f877c2f3-Paper-Conference.pdf.

Ankit Pensia, Varun S. Jog, and Po-Ling Loh. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 1(1):131–144, 2020. doi: 10.1109/JSAIT.2020.2991005.

Alexandre Rame and Matthieu Cord. DICE: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=R2ZlTVPx0Gk.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. URL https://dl.acm.org/doi/10.5555/3600270.3601918.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3183–3193, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Matthias C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2006.06.001. URL https://www.sciencedirect.com/science/article/pii/S0888613X06000387.

Lev Utkin and Sebastien Destercke. Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning*, 50(5):778–798, 2009. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2009.02.004. URL https://www.sciencedirect.com/science/article/pii/S0888613X09000401.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2):125–148, 2000. ISSN 0888-613X. doi: https://doi.org/10.1016/S0888-613X(00)00031-1. URL https://www.sciencedirect.com/science/article/pii/S0888613X00000311.

Kaizheng Wang, Fabio Cuzzolin, Shireen Kudukkil Manchingal, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal deep ensembles for uncertainty quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=PCgnTiGC9K.

Kaizheng Wang, Fabio Cuzzolin, David Moens, and Hans Hallez. Credal ensemble distillation for uncertainty quantification. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, AAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i31.39837. URL https://doi.org/10.1609/aaai.v40i31.39837.

Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pp. 196–202. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_16. URL https://doi.org/10.1007/978-1-4612-4380-9_16.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Songjie Xie, Shuai Ma, Ming Ding, Yuanming Shi, Mingjian Tang, and Youlong Wu. Robust information bottleneck for task-oriented communication with digital modulation. *IEEE Journal on Selected Areas in Communications*, 41(8):2577–2591, 2023. doi: 10.1109/JSAC.2023.3288252.

Zhengmin Zhang. Estimating mutual information via Kolmogorov distance. *IEEE Transactions on Information Theory*, 53(9):3280–3282, 2007. doi: 10.1109/TIT.2007.903122.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12857–12867. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhou21g.html.

CREDAL INFORMATION BOTTLENECK

SUPPLEMENTARY MATERIAL

CONTENTS

A Experimental Setup **16**

A.1 Architectures 16

A.2 Objective and optimization 16

A.3 Baselines 17

A.4 Choice of β and rate estimates 17

A.5 Rate estimates and bounds 18

A.6 Moment-matched code priors 18

A.7 Evaluation 19

A.8 Statistics and compute 19

B Additional Results **20**

C Proofs **20**

C.1 Measure-theoretic conventions and technical background 20

C.2 Properties of the credal target code 24

C.3 Credal-set recovery 25

C.4 The pooling floor and separation 26

C.5 Consequence for uncertainty quantification 28

C.6 Variational Relaxation 29

D The Colored-MNIST dataset and uncertainty map **31**

A Experimental Setup

This section lists the data, the architectures, the objective and its optimization, the baselines, how β is chosen, and how each reported quantity is computed. The construction of Colored-MNIST and its closed-form labeling laws are given in Appendix D. The code, the configuration files and the scripts that produce every table and figure will be released with the paper.

Out-of-support inputs. For C-MNIST and C-KMNIST the out-of-support inputs are the Fashion-MNIST test images, and for C-Fashion-MNIST the MNIST test images. They are colored by the rule of an environment drawn uniformly at random, so they differ from the training data in shape but not in color.

Under-trained inputs. $\mathrm{AUROC}_{\mathrm{model}}$ is measured on a second copy of each colored dataset in which classes 2 and 7 keep a random 2% of their training images in both environments; the test split is unchanged, and its images of these two classes are the under-trained inputs. Every method is trained on this variant with the same configuration and the same β as on the full dataset, and only $\mathrm{AUROC}_{\mathrm{model}}$ is taken from it.

Augmentation of the colored datasets. During training only, every colored image is translated at random by up to 2 pixels (zero padding, both channels together), for every method. The labeling laws depend only on the class and the color, so the translation leaves the environments unchanged. Without it, the 25% label noise of small environments is memorized within a few hundred steps. Validation, test and all evaluations use the untranslated images.

A.1 Architectures

Encoder. Every method uses the same encoder NN. On the colored datasets, it consists of three 3×3 convolutions with 32, 64 and 64 channels and strides 1, 2 and 2, each followed by GroupNorm with 8 groups and a leaky ReLU, and an average pooling of the last feature map to 2×2, which gives 256 features. GroupNorm normalizes each input separately, so no prediction depends on the composition of the batch. The backbone is deterministic and shared by the K encoder hypotheses. Each hypothesis adds a Bayesian head with one hidden layer of 256 units that outputs the mean and log-variance of a diagonal Gaussian code of dimension $d_z = 16$; log-variances are clipped to $[-12, 4]$.

Decoders and hypotheses. Each environment has J Bayesian decoders q_{ij} with one hidden layer of 64 units, shared by all encoder hypotheses. Credal methods use $(K, J) = (5, 3)$. Bayesian layers have mean-field Gaussian posteriors over their weights. Following Caprio et al. (2024a), the vertices of the prior credal set are Gaussian priors whose scales are spread geometrically over $[0.8, 1.25]$ across the K encoder and the J decoder hypotheses, in units of each layer's Kaiming standard deviation $g/\sqrt{\text{fan-in}}$ ($g = \sqrt{2}$ before a nonlinearity, 1 at the output). Posteriors are initialized with standard deviations of 0.1 times the prior's. Deterministic layers use Kaiming-normal initialization.

Code priors. Each pair (i, k) has its own code prior r_{ik}, a Gaussian mixture with 10 components learned jointly with the model. After one epoch (one pass over the largest training environment) the components are placed by k-means on 4,096 sampled codes of each pair; from then on they are fitted by their own maximum-likelihood term, which is not weighted by the fidelity weights. VIB, Ens-VIB and IB-IRM use the fixed prior $\mathcal{N}(0, I)$.

A.2 Objective and optimization

Brackets. Training minimizes the smoothed objective of Equation (13). For every encoder hypothesis k, the brackets $\widehat{b}_{ik}$ of Equation (11) are combined across environments with fidelity weights $w_i^{(k)} = \mathrm{softmax}_i\big(\bar{b}_{ik}/\tau_t\big)$, where $\bar{b}_{ik}$ is an exponential moving average (factor 0.9) of past brackets, including the weight-KL term described below. The weights are held fixed within a step. The inverse temperature $1/\tau_t$ rises linearly from 0, where the environments are averaged, to β over the whole step budget, so that the final temperature is $\tau = 1/\beta$. Because the weights use a moving average and are held fixed within a step, the update approximates, rather than equals, the gradient of $\mathcal{L}^{\tau}_{\mathrm{VCIB}}$. The backbone and every decoder receive the sum over k of the weighted gradients; nothing is detached or

routed by hypothesis, except the code priors as above. The floor L_i^{ref} is the closed-form $H_i(Y \mid X)$ on the colored datasets.

Weight posteriors. The weight-KL terms are scaled by $\alpha = 0.1$, a tempered posterior: with $\alpha = 1$, the KL of the several hundred thousand head weights outweighs the likelihood and the heads collapse onto their priors. The coefficient rises linearly from 0 over the first 2,000 steps. Each environment's term is divided by the size of its training split.

Optimizer. All methods and datasets use Adam with learning rate 10^{-3}, without a schedule, in single precision. A batch holds 256 images per environment; pooled methods draw m times as many images from the union of the environments. Training runs for at most 8,000 steps. Every 200 steps the objective is evaluated on the validation split with 4 weight samples, and training stops after 10 evaluations without improvement; the parameters with the best validation objective are kept. The validation split is used for nothing else, apart from calibrating the label-shift baseline.

Warm starts. For each method, dataset and seed, the β grid is traversed in decreasing order: the largest β trains from scratch and each smaller β starts from the checkpoint of the previous one, without re-placing the code priors. This traces the rate–distortion curve by continuation and keeps small β away from the trivial encoder.

A.3 Baselines

All baselines use the backbone, heads, optimizer, step budget and early stopping above. Table 3 summarizes how they differ.

VIB (Alemi et al., 2017) trains one stochastic encoder and one decoder on the union of the environments with the prior $\mathcal{N}(0, I)$. **Ens-VIB** (Lakshminarayanan et al., 2017) trains five such models independently, each with its own backbone, head and decoder. **IB-IRM** (Ahuja et al., 2021) adds to the VIB objective, averaged uniformly over the environments, the IRMv1 penalty with weight 1 during the first epoch and 100 afterwards. **CBDL** (Caprio et al., 2024a) is the credal Bayesian network with $(K, J) = (5, 3)$, deterministic codes and one decoder row trained on the pooled data. **VCIB**$_{\mathrm{mod}}$ keeps VCIB's model credal set, stochastic codes, code priors and rate term, but pools the environments into one decoder row. **VCIB**($\beta \to \infty$) keeps VCIB's environment decoders and worst-case weighting but has deterministic codes and no rate term; its temperature uses $\beta = 16$.

Table 3: Methods. R: decoder rows (one per environment, or one pooled); K, J: encoder and decoder hypotheses.

Method	R	K	J	Bayesian	Rate term	Environments combined by
VIB	1	1	1	no	yes, $\mathcal{N}(0, I)$	pooling
Ens-VIB	1	5	1	no	yes, $\mathcal{N}(0, I)$	pooling
IB-IRM	1	1	1	no	yes, $\mathcal{N}(0, I)$	uniform average + IRMv1
CBDL	1	5	3	yes	no	pooling
VCIB$_{\mathrm{mod}}$	1	5	3	yes	yes, r_k	pooling
VCIB($\beta \to \infty$)	m	5	3	yes	no	worst case
VCIB	m	5	3	yes	yes, r_{ik}	worst case

A.4 Choice of β and rate estimates

Rate budget. Every method with a rate term is trained over $\beta \in \{8, 12, 16\}$ for all ten seeds, and over the full grid $\{2, 3, 4, 6, 8, 12, 16\}$ for seed 0, which traces the curves of Figure 3. For each method and dataset we report the β of the ten-seed grid whose seed-averaged rate on training inputs is closest to 2 bits, a budget fixed before the experiments, just above the 1.88 bits that the exact CIB needs for zero distortion on the colored alphabet. No validation or test quantity enters this choice; selecting by validation distortion would always pick the largest β. This gives $\beta = 16$ for every method except VCIB on C-KMNIST and IB-IRM on every dataset ($\beta = 12$). The baselines never reach 2 bits: over the whole grid their rate saturates between 1.1 and 1.4 bits (IB-IRM below 0.5), so they are reported at $\beta = 16$ (IB-IRM at 12).

A.5 RATE ESTIMATES AND BOUNDS

Per-code rate. For any density $\hat{r}$ on $\mathcal{Z}$, $I_i(X;Z_k) = \mathbb{E}_{\mu_i}\big[D_{\mathrm{KL}}(e_k(\cdot \mid X) \,\|\, p_{ik})\big] \le \mathbb{E}_{\mu_i}\big[D_{\mathrm{KL}}(e_k(\cdot \mid X) \,\|\, \hat{r})\big]$, where p_{ik} is the code marginal of hypothesis k under environment i; the gap is $D_{\mathrm{KL}}(p_{ik} \,\|\, \hat{r})$. We use the same $\hat{r}$ for every method, whatever prior it was trained with. After training, with the encoder frozen, we fit by EM a Gaussian mixture $\hat{r}_{ik}$ with 10 components to 4,096 codes sampled from the training inputs of environment i, pooled over 4 weight draws. The bound is then estimated by Monte Carlo,

$$\widehat{I}_i(X;Z_k) = \frac{1}{4N}\sum_{n=1}^{N}\sum_{t=1}^{4}\Big[\log e_k^{(t)}\big(z_{nt} \mid x_n\big) - \log \hat{r}_{ik}\big(z_{nt}\big)\Big], \quad z_{nt} \sim e_k^{(t)}(\cdot \mid x_n), \tag{18}$$

where $e_k^{(t)}$ is the Gaussian code of hypothesis k under weight draw t. The reported rate $\mathcal{R} = \max_{i,k} \widehat{I}_i(X;Z_k)$ takes $x_1, \dots, x_N$ from the test split; the training rate, which selects β (Appendix A.4), takes 4,096 fresh training inputs. Both are in bits.

Brackets from pairwise distances. To check $\mathcal{R}$ and to bound the rate of all K codes together, we also bracket the mutual information with the pairwise-distance estimators of Kolchinsky & Tracey (2017). Fix a weight draw θ and N inputs $x_1, \dots, x_N$ of environment i, taken with equal weights. Each input has a Gaussian code $e_a = \mathcal{N}(m_a, \operatorname{diag} s_a^2)$, and the code marginal is the mixture $\frac{1}{N}\sum_b e_b$. Since $I(X;Z \mid \theta) = H\big(\frac{1}{N}\sum_b e_b\big) - \frac{1}{N}\sum_a H(e_a)$, the bounds of Kolchinsky & Tracey (2017) on the entropy of a mixture give, in nats,

$$-\frac{1}{N}\sum_{a=1}^{N}\log\frac{1}{N}\sum_{b=1}^{N} e^{-D_{\mathrm{B}}(e_a,e_b)} \;\le\; I(X;Z \mid \theta) \;\le\; -\frac{1}{N}\sum_{a=1}^{N}\log\frac{1}{N}\sum_{b=1}^{N} e^{-D_{\mathrm{KL}}(e_a \,\|\, e_b)}, \tag{19}$$

with D_{B} the Bhattacharyya distance. For diagonal Gaussians both distances are sums over coordinates,

$$D_{\mathrm{KL}}(e_a \,\|\, e_b) = \tfrac{1}{2}\textstyle\sum_l \Big[\log\frac{s_{bl}^2}{s_{al}^2} + \frac{s_{al}^2 + (m_{al}-m_{bl})^2}{s_{bl}^2} - 1\Big],$$
$$D_{\mathrm{B}}(e_a,e_b) = \textstyle\sum_l \Big[\frac{(m_{al}-m_{bl})^2}{8\,v_l} + \frac{1}{2}\log v_l - \frac{1}{4}\log\big(s_{al}^2 s_{bl}^2\big)\Big], \qquad v_l = \tfrac{1}{2}\big(s_{al}^2 + s_{bl}^2\big).$$

Given the weights, the K codes of an input are independent, so the code $(Z_1, \dots, Z_K)$ is again a diagonal Gaussian and its distances are the sums over k of the per-code distances. The same formula then brackets the joint rate $I(X; Z_1, \dots, Z_K \mid \theta)$. We use $N = 2{,}048$ test inputs per environment, one weight draw shared by all N inputs, and average the bounds over 4 draws. The per-code bracket takes the maximum over k, and both brackets the maximum over environments. $\widehat{\mathcal{R}}_{\mathrm{joint}}$ in Table 1 is the upper end of the joint bracket. Two properties matter for reading these numbers. The bounds refer to the empirical distribution of the N inputs, so they cannot exceed $\log N$ (11 bits here); no reported value comes close. They condition on the weights, as the rate term of the objective does, so the spread of the weight posterior is not counted as information about X. For VCIB, Ens-VIB, $\mathrm{VCIB}_{\mathrm{mod}}$ and VIB at the selected β, $\mathcal{R}$ lies inside the per-code bracket in every seed and dataset.

A.6 MOMENT-MATCHED CODE PRIORS

Deployment and the environment weights of the default decision do not use the trained mixture priors r_{ik}. They use one Gaussian per pair (i,k) fitted after training by matching the first two moments of environment i's training codes:

$$r_{ik}^{G} = \mathcal{N}\big(\bar{m}_{ik}, \operatorname{diag}\bar{s}_{ik}^2\big), \tag{20}$$

$$\bar{m}_{ik} = \frac{1}{TN}\sum_{t,n} m_k^{(t)}(x_n), \tag{21}$$

$$\bar{s}_{ik}^2 = \frac{1}{TN}\sum_{t,n}\Big[s_k^{(t)}(x_n)^2 + \big(m_k^{(t)}(x_n) - \bar{m}_{ik}\big)^2\Big], \tag{22}$$

where $m_k^{(t)}(x)$ and $s_k^{(t)}(x)^2$ are the mean and variance of the code of hypothesis k under weight draw t, over $N = 4{,}096$ training inputs of environment i and $T = 4$ draws (squares taken coordinate-wise).

This is the Gaussian with the mean and covariance diagonal of the code marginal of Appendix A.5. The fitted mixtures would also do, but their narrow components sit on the training codes, so a test code falls between them and a single batch would identify the environment with a confidence that the data do not support. The two moments ignore where exactly the training codes lie. They enter in two ways. For a deployment batch $\mathcal{B}$, the score of environment i is $\sum_{x\in\mathcal{B}} \frac{1}{K}\sum_k \mathbb{E}_{z\sim e_k(\cdot|x)}[\log r^G_{ik}(z)]$, in closed form for Gaussians, $\mathbb{E}_{z\sim\mathcal{N}(m,\operatorname{diag} s^2)}[\log r^G_{ik}(z)] = -\frac{1}{2}\sum_l \big[\log(2\pi\bar{s}^2_{ikl}) + \big(s_l^2 + (m_l - \bar{m}_{ikl})^2\big)/\bar{s}^2_{ikl}\big]$, averaged over 8 weight draws, and the batch weights of Equation (16) are the softmax of these scores (a uniform prior over environments). Summed over a batch drawn from one environment, the difference of two scores is a log-likelihood ratio in code space, which is why a batch of 100 inputs identifies the environment reliably. For the default decision on a single input, the weight of environment i is instead $\frac{1}{K}\sum_k \operatorname{softmax}_i \log \mathbb{E}_{z\sim e_k(\cdot|x)}[r^G_{ik}(z)]$, the ratio of the smoothed code densities of the environments at x; the expectation is again a Gaussian integral. The average runs over the hypotheses outside the softmax because densities in different hypotheses' code spaces cannot be compared.

A.7 Evaluation

Predictions. Each vertex of the predictive credal set is estimated with 32 joint draws of weights and codes. Methods with environment decoders decide by the credal minimax rule: the Bayes rule against least-favorable weights over the mKJ environment–hypothesis vertices, fitted by mirror ascent on 4,096 unlabeled training inputs and conditioned on each input's environment evidence in code space. VIB and IB-IRM take the argmax of their predictive distribution, Ens-VIB of its members' average, and CBDL and VCIB_{mod} decide by the same minimax rule over their pooled credal set.

Distortion and accuracy. $\mathcal{D}$ is the worst environment's expected excess log-loss on the test split, in bits. On the colored datasets it is computed under the exact labeling laws. WC-Acc is the lowest accuracy of the default decision over the environments' test splits, in expectation under the exact laws on the colored datasets. RC-AUC is the area under the risk–coverage curve when the model abstains in order of decreasing total uncertainty.

Detection. $\text{AUROC}_{\text{conflict}}$ scores EU_{input} on test inputs of classes 0–4 against inputs of classes 5–9; $\text{AUROC}_{\text{model}}$ scores EU_{model} on the under-trained test inputs against the rest; $\text{AUROC}_{\text{OOD}}$ scores AU on the out-of-support inputs against the test inputs. Methods without the decomposition use one score for both epistemic columns: VIB and IB-IRM the predictive entropy, Ens-VIB the mutual information between label and member, CBDL and VCIB_{mod} the entropy width of their credal set.

Deployment. Each environment's test inputs are partitioned at random into disjoint unlabeled batches of 100. Methods with environment decoders turn the code likelihoods of a batch into weights over the environments (Equation (16)) and predict, for every input of the batch, the E-admissible label with the largest lower probability over the collapsed credal set. Methods without environment decoders receive bias-corrected temperature scaling fitted on the validation split, followed by label-shift EM (Alexandari et al., 2020) on the environment's entire test split, which gives them more unlabeled data than a batch. $\text{WC-Acc}_{\text{int}}$ is the lowest accuracy over environments after this step, and $\text{WC-Reg}_{\text{int}} = \max_i[\text{Acc}^\star_i - \text{Acc}_i]$ its largest gap to the environment's Bayes accuracy $\text{Acc}^\star_i$, both in expectation under the exact laws.

A.8 Statistics and compute

Seeds and tests. The colored results use 10 seeds; a seed fixes the data split, the labels and colors, the initialization and the evaluation. Each method is compared with VCIB by an exact two-sided Wilcoxon signed-rank test on seed-paired values (zero differences handled by Pratt's method), with Holm–Bonferroni correction within each dataset and metric at level 0.05. A value is bold when it is the best mean or is not significantly different from it under the same test.

Compute. All runs used one NVIDIA GeForce RTX 5060 Ti with 16 GB of memory, with up to eight colored runs in parallel. A colored run takes 1 to 10 minutes. The colored results rest on 1,140 runs: for each of the three datasets and its under-trained variant, 34 runs for each of the five methods with a rate term (ten seeds at three values of β and four further values on seed 0) and ten for each rate-free method.

B Additional Results

Table 4: Label-free deployment on C-MNIST, the numbers behind Figure 2 (seed 0, $\beta = 16$): accuracy before and after the model, deployed in one training environment, performs a test-time intervention. Results reported over 2000 random batches of size $N = 100$ per environment; percentages are the share of test images in each group. Ground truth: pooled Bayes rule, then the environment's Bayes rule. VCIB: decision on the full credal set (before), then on the set collapsed by the batch (after), as in Equation (16); it identifies the environment in 97% of the batches. VIB: argmax, then calibrated label-shift EM (Alexandari et al., 2020). On the digits 0–4, where the environments conflict, VCIB achieves a relative gain of 21% (0.726 to 0.879), as much as switching to the true environment's Bayes rule. Label-shift EM leaves VIB unchanged: the environments share their label marginal and differ only in how color maps to the label, and VIB's pooled decoder ignores color on these digits.

	Digits 0–4 conflict (51%)		Digits 5–9 color agrees (34%)		Digits 5–9 color disagrees (15%)		All test images (100%)		Worst environment	
	before	after	before	after	before	after	before	after	before	after
Ground truth	0.742	0.898	0.964	0.964	0.750	0.750	0.818	0.898	0.816	0.897
VCIB	0.726	0.879	0.948	0.956	0.749	0.729	0.804	0.882	0.801	0.867
VIB	0.734	0.734	0.955	0.954	0.750	0.750	0.810	0.810	0.803	0.803

Label-free deployment. Table 4 details the test-time intervention behind Figure 2. The model is deployed in one of the training environments without being told which, and receives its test inputs in unlabeled batches of 100. The ground-truth row shows where an intervention can help at all: knowing the environment changes nothing on the digits 5–9, on which the environments agree, and raises the accuracy on the conflicting digits 0–4 from 0.742 to 0.898. VCIB recovers almost all of this gain without labels or retraining. From the code likelihoods of a batch it identifies the environment in 97% of the cases, collapses its credal set accordingly (Equation (16)), and raises its accuracy on the conflicting digits from 0.726 to 0.879, a relative gain of 21%, matching the gain of the true environment's Bayes rule. Over all test images, its accuracy rises from 0.804 to 0.882, and its worst-environment accuracy from 0.801 to 0.867, close to the known-environment ceiling of 0.897; the only cost is a small drop on the digits whose color disagrees with their label, which make up 15% of the inputs.

Representative inputs. Figures 4 to 6 show, for each colored dataset, test images whose largest uncertainty component is AU, $\mathrm{EU}_{\mathrm{input}}$ or $\mathrm{EU}_{\mathrm{model}}$, drawn at random among the clearly uncertain ones ($\mathrm{TU} \geq 0.5$ bits) of the rare-class variant, in which classes 2 and 7 keep 2% of their training images. $\mathrm{EU}_{\mathrm{input}}$ consistently isolates the classes 0–4 on which the two environments prescribe different labeling laws (93%, 80% and 97% of the images on C-MNIST, C-KMNIST and C-Fashion-MNIST). On C-MNIST, every image dominated by $\mathrm{EU}_{\mathrm{model}}$ belongs to the thinned classes, i.e., the model's own uncertainty concentrates where training data were scarce. High-noise inputs, due to both label noise and visual quality degradation, concentrate in the AU blocks and never appear among the images dominated by $\mathrm{EU}_{\mathrm{model}}$. This confirms that the disentanglement of AU and EU proposed in Section 4.3 correctly identifies the source of the uncertainty in the data.

C Proofs

In the following section, we provide formal proofs supporting the theorems introduced in the main manuscript (Section 4), as well as additional theoretical results.

C.1 Measure-theoretic conventions and technical background

Conventions and representations. Throughout, random variables take values in standard Borel spaces. The mutual information $I(U;V)$ is the Kullback–Leibler divergence between the joint distribution of (U,V) and the product of their marginals, taking values in $[0,\infty]$. Standard properties, including the chain rule $I(U;V,W) = I(U;V) + I(U;W \mid V)$ and the data-processing inequality, hold in this general setting (Gray, 2011). Conditional distributions are taken as regular conditional probability measures.

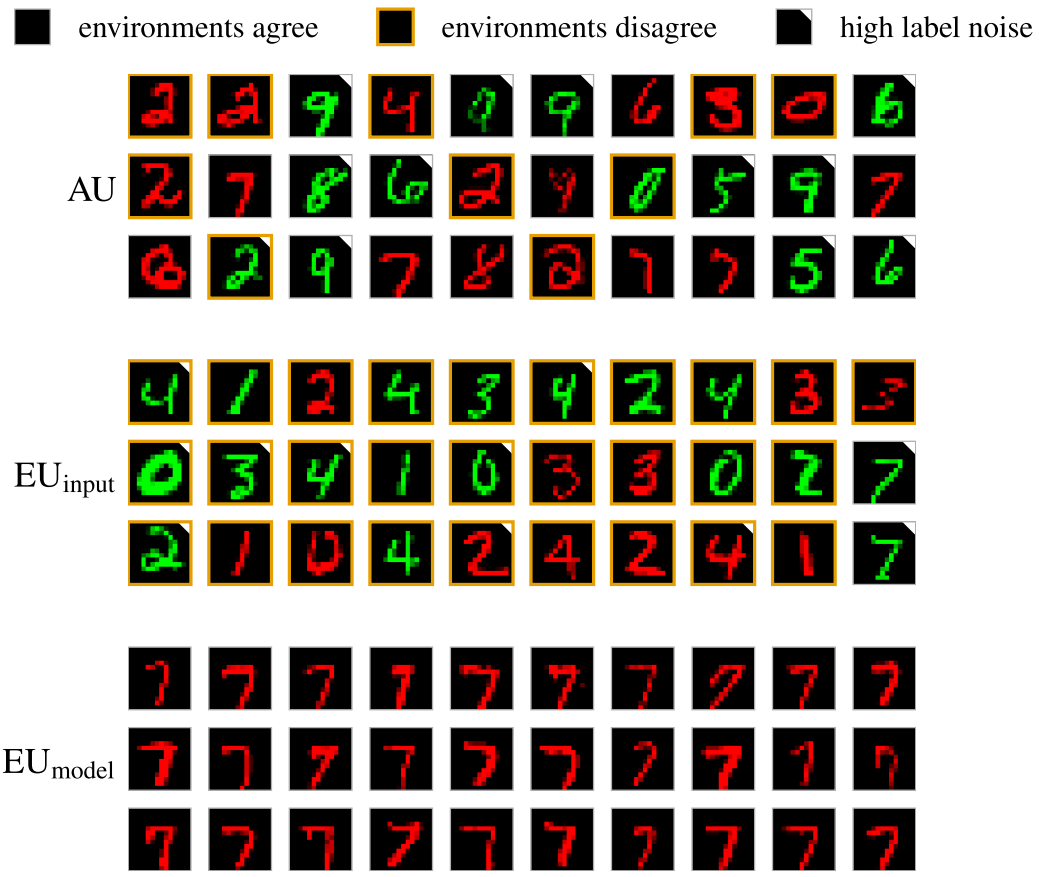


Figure 4: C-MNIST test images whose largest uncertainty component is AU, $\mathrm{EU}_{\mathrm{input}}$ or $\mathrm{EU}_{\mathrm{model}}$ (VCIB, $\beta = 16$; classes 2 and 7 thinned to 2%).

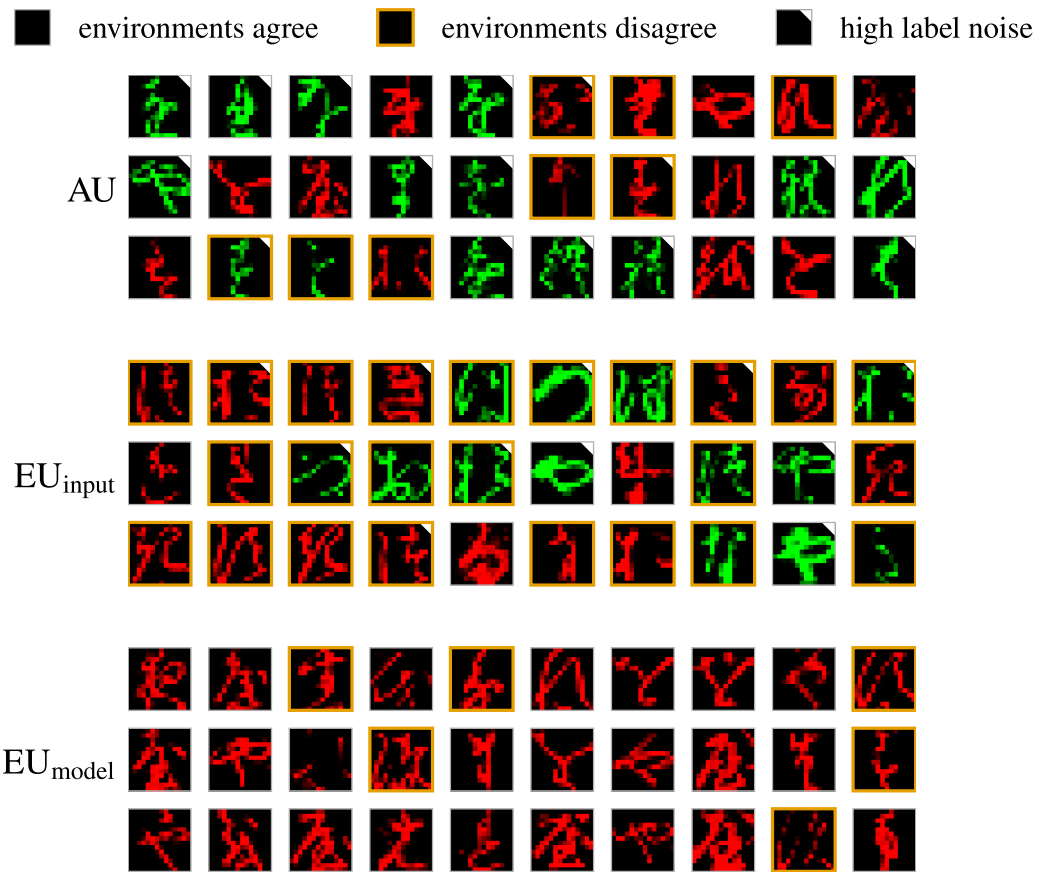


Figure 5: As Figure 4, on C-KMNIST ($\beta = 12$).

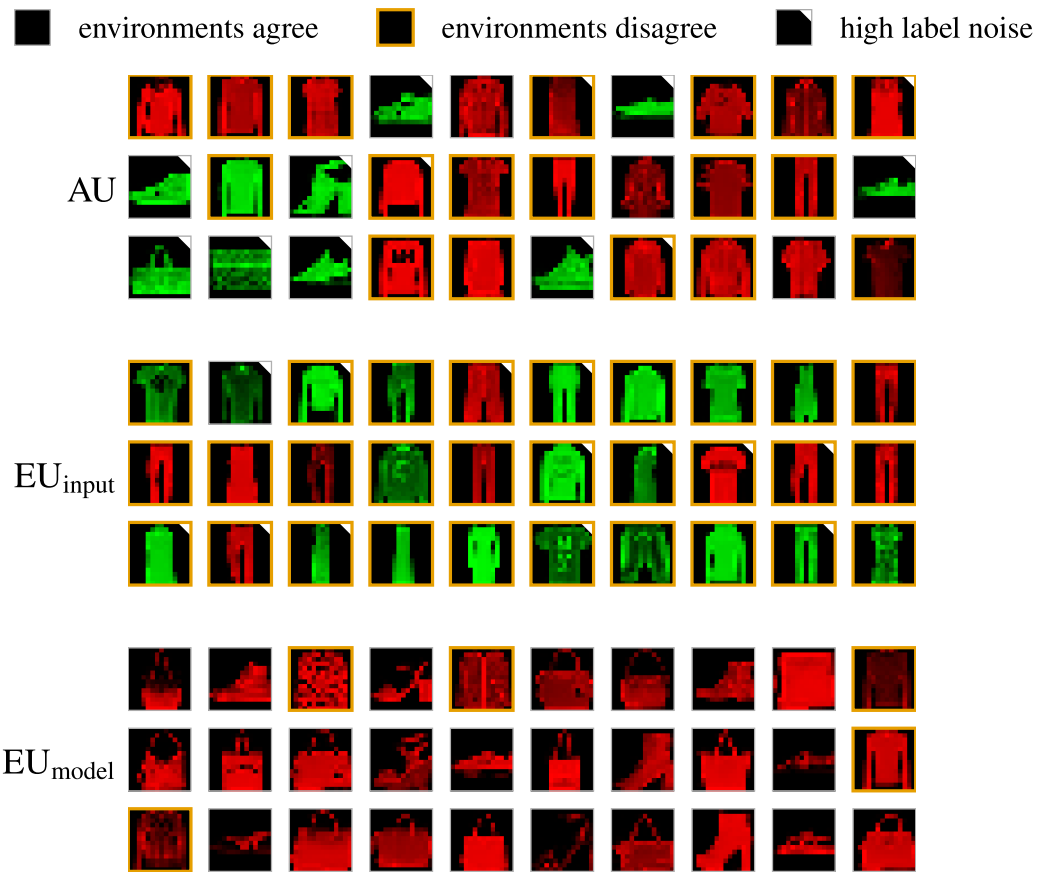


Figure 6: As Figure 4, on C-Fashion-MNIST ($\beta = 16$).

For each environment $i \in [m]$, we fix a measurable version of the Markov kernel $p_i(y \mid x)$ defined μ_i-almost everywhere. The credal target $T : \mathcal{X} \to \Delta(\mathcal{Y})^m$, defined by $T(x) = (p_1(\cdot \mid x), \ldots, p_m(\cdot \mid x))$, is therefore measurable, where $\Delta(\mathcal{Y}) \subset \mathbb{R}^n$ denotes the probability simplex on the label set $\mathcal{Y}$ ($n = |\mathcal{Y}|$). Because the mixing weights satisfy $\lambda_i > 0$ for all i, every environment marginal is absolutely continuous with respect to the mixture ($\mu_i \ll \mu_{\boldsymbol{\lambda}}$). Consequently, the Radon–Nikodym derivatives $\rho_i^{\lambda}(x) = \lambda_i \, (d\mu_i / d\mu_{\boldsymbol{\lambda}})(x)$ and the pooled target $\bar{T}_{\boldsymbol{\lambda}}(x) = \sum_i \rho_i^{\lambda}(x) p_i(\cdot \mid x)$ are well-defined $\mu_{\boldsymbol{\lambda}}$-almost everywhere and measurable.

We write $I_{\boldsymbol{\lambda}}$, $H_{\boldsymbol{\lambda}}$, and $\mathbb{E}_{\boldsymbol{\lambda}}$ for information-theoretic quantities evaluated under the mixture law $\mu_{\boldsymbol{\lambda}}(dx)\, p_{\boldsymbol{\lambda}}(y \mid x)\, e(dz \mid x)$. A random variable is discrete if its support is at most countable; H denotes the Shannon entropy in bits. When $T(X)$ is not discrete under μ_i, we adopt the convention $H_i(T(X)) = \infty$. Because an infimum of the objective F_β over general Markov kernels on continuous spaces may not be attained, an encoder e is termed an *ε-minimizer* (or near-minimizer) if $F_\beta(e) \leq \inf_{e'} F_\beta(e') + \varepsilon$ for a given $\varepsilon \geq 0$. Similarly, $\mathbb{E}_i$ denotes expectation under $\mu_i(dx)\, p_i(y \mid x)\, e(dz \mid x)$, and $\mathcal{R}_{\mathcal{P}}(D) = \inf_e \{\mathcal{R}(e) : \mathcal{D}(e) \leq D\}$ denotes the credal rate–distortion function, that is, the smallest credal rate achieving credal distortion at most D.

Key information-theoretic identities. Our analysis relies on two fundamental identities:

(a) *Conditional mutual information as expected divergence:* Let $G = g(X)$ for a measurable mapping g, and let $p(\cdot \mid X)$ be a regular conditional distribution of Y given X. Defining the intermediate predictor $\pi(\cdot \mid G) = \mathbb{E}[p(\cdot \mid X) \mid G]$, we have

$$I(X; Y \mid G) = \mathbb{E}\big[H(\pi(\cdot \mid G))\big] - \mathbb{E}\big[H(p(\cdot \mid X))\big] = \mathbb{E}\left[D_{\mathrm{KL}}\big(p(\cdot \mid X) \,\|\, \pi(\cdot \mid G)\big)\right]. \quad (23)$$

By the positivity of the KL divergence, $I(X; Y \mid G) = 0$ if and only if $p(\cdot \mid X) = \pi(\cdot \mid G)$ almost surely, that is, if and only if $p(\cdot \mid X)$ is $\sigma(G)$-measurable. Furthermore, by the chain rule, $I(X; Y) = I(G; Y) + I(X; Y \mid G)$.

Applying this identity to an encoder Z with Markov chain $Y \to X \to Z$, the environment distortion simplifies to $\delta_i(e) = I_i(X; Y \mid Z)$. Hence, $\delta_i(e) = 0$ if and only if $p_i(\cdot \mid X) = P_i(Y \in \cdot \mid Z)$ almost surely under P_i, recovering the condition in Definition 1.

(b) *Continuity of entropy:* For any two distributions $p, q \in \Delta(\mathcal{Y})$ with total variation distance $d_{\mathrm{TV}}(p, q) \leq \xi \leq 1 - \frac{1}{n}$, the entropy difference is bounded by (Audenaert, 2007; Zhang, 2007)

$$|H(p) - H(q)| \leq \omega(\xi) \coloneqq \xi \log_2(n - 1) + h(\xi), \quad (24)$$

where $h(\xi) = -\xi \log_2 \xi - (1-\xi) \log_2 (1-\xi)$ is the binary entropy function, with $\lim_{\xi \to 0} \omega(\xi) = 0$.

C.2 Properties of the credal target code

Lemma 5 (Sufficiency of the credal target)**.** *The deterministic encoder $Z = T(X)$ is credally sufficient. Moreover, whenever $T(X)$ is discrete under μ_i, its rate is given by $I_i(X; T(X)) = H_i(T(X))$.*

Proof. Under each environment law P_i, the conditional distribution $p_i(\cdot \mid X) = T_i(X)$ is a coordinate projection of $T(X)$, and hence is trivially $\sigma(T(X))$-measurable. Equation (23) with $G = T(X)$ implies that $I_i(X; Y \mid T(X)) = 0$, which yields $\delta_i = 0$ and $I_i(T(X); Y) = I_i(X; Y)$ for all $i \in [m]$. If $T(X)$ is discrete, determinism implies $H_i(T(X) \mid X) = 0$, giving $I_i(X; T(X)) = H_i(T(X)) - H_i(T(X) \mid X) = H_i(T(X))$. □

Lemma 6 (Uniform quantization of the credal target)**.** *For any $\xi \in (0, 1 - 1/n]$, there exists a measurable partition $Q_\xi : \Delta(\mathcal{Y})^m \to \{1, \ldots, N(\xi)\}$ with at most $N(\xi) \leq \lceil (n-1)/\xi \rceil^{m(n-1)}$ cells such that the quantized representation $Z = Q_\xi(T(X))$ satisfies*

$$I_i(X; Z) \leq \log_2 N(\xi) \quad \textit{and} \quad \delta_i(Z) \leq \omega(\xi) \quad \forall i \in [m]. \quad (25)$$

Consequently, $\mathcal{R}_{\mathcal{P}}(\omega(\xi)) \leq \log_2 N(\xi)$, and $\mathcal{R}_{\mathcal{P}}(D) < \infty$ for every target distortion $D > 0$.

Proof. Fix $\xi \in (0, 1 - 1/n]$, set grid width $\eta = \xi/(n-1)$ and resolution $V = \lceil 1/\eta \rceil$. On each factor simplex $\Delta(\mathcal{Y})$, partition points q using the coordinate-wise index vector $(\min\{\lfloor q_y/\eta \rfloor, V - 1\})_{y<n}$. Each resulting cell is formed by the intersection of the simplex with a hypercube of side length at most η along the first $n-1$ coordinates; this set is convex, and any two distributions within its closure

differ by at most η per coordinate. For the remaining coordinate $y = n$, the discrepancy is bounded by $(n-1)\eta$, yielding a total variation diameter of at most $(n-1)\eta = \xi$.

Taking the Cartesian product of these partitions across all m environments yields a composite code Q_ξ with cardinality $N(\xi) \le V^{m(n-1)}$. Under law P_i, the true target $p_i(\cdot \mid X) = T_i(X)$ belongs to the i-th marginal cell of Z. Because the closure of each cell is convex, the conditional mean $\pi_i(\cdot \mid Z) = \mathbb{E}_i[p_i(\cdot \mid X) \mid Z]$ lies in the same cell, ensuring that $d_{\mathrm{TV}}(p_i(\cdot \mid X), \pi_i(\cdot \mid Z)) \le \xi$ almost surely. Applying Equation (23) alongside the continuity bound (Equation (24)) yields

$$\delta_i(Z) = I_i(X; Y \mid Z) \le \mathbb{E}_i\big[\big|H(\pi_i(\cdot \mid Z)) - H(p_i(\cdot \mid X))\big|\big] \le \omega(\xi). \tag{26}$$

Because Z takes at most $N(\xi)$ discrete values, $I_i(X; Z) \le H_i(Z) \le \log_2 N(\xi)$ holds across all environments. Since $\omega(\xi) \to 0$ as $\xi \to 0$, any $D > 0$ strictly exceeds $\omega(\xi)$ for sufficiently small ξ, guaranteeing $\mathcal{R}_{\mathcal{P}}(D) < \infty$. □

C.3 Credal-set recovery

Proposition 7 (Credal-set recovery)**.** *For any encoder e and environment $i \in [m]$, let $q_i(z) = P_i(Y \in \cdot \mid Z = z)$ denote the environment-specific Bayes decoder and $\hat{p}_i(\cdot \mid x) = \int q_i(z)\, e(dz \mid x)$ the expected readout. Then:*

$$\mathbb{E}_i\left[d_{\mathrm{TV}}\big(p_i(\cdot \mid X), \hat{p}_i(\cdot \mid X)\big)\right] \le \mathbb{E}_i\left[d_{\mathrm{TV}}\big(p_i(\cdot \mid X), q_i(Z)\big)\right] \le \sqrt{\tfrac{\ln 2}{2}\,\delta_i(e)} \le \sqrt{\tfrac{\ln 2}{2}\,\mathcal{D}(e)}. \tag{27}$$

If, in addition, the likelihood ratios across environments $i, l \in [m]$ are bounded by $\chi = \max_{i,l} \operatorname{ess\,sup}(d\mu_l / d\mu_i) < \infty$, then under the Hausdorff distance d_H induced by d_{TV},

$$\mathbb{E}_{\mu_l}\left[d_H\big(\widehat{\mathcal{C}}_e(X), \mathcal{C}^\star(X)\big)\right] \le \chi\, m \sqrt{\tfrac{\ln 2}{2}\,\mathcal{D}(e)}\,, \tag{28}$$

for every $l \in [m]$. For exact recovery, mutual absolute continuity of $\mu_1, \dots, \mu_m$ is sufficient, without requiring bounded likelihood ratios. If $\mathcal{D}(e) = 0$, then $\widehat{\mathcal{C}}_e(X) = \mathcal{C}^\star(X)$ almost surely in every environment.

Proof. Under the Markov chain $Y \to X \to Z$, we have $P_i(Y \in \cdot \mid X, Z) = p_i(\cdot \mid X)$. The chain rule for mutual information expands the environment distortion as $\delta_i(e) = I_i(X; Y \mid Z) = \mathbb{E}_i\left[D_{\mathrm{KL}}(p_i(\cdot \mid X) \,\|\, q_i(Z))\right]$. Pinsker's inequality $d_{\mathrm{TV}}(p, q) \le \sqrt{\frac{\ln 2}{2} D_{\mathrm{KL}}(p \| q)}$ combined with Jensen's inequality yields the second inequality in Equation (27). Since the conditional distribution of Z given $X = x$ is $e(\cdot \mid x)$, the readout satisfies $\hat{p}_i(\cdot \mid X) = \mathbb{E}_i[q_i(Z) \mid X]$. The first inequality then follows from the convexity of total variation in its second argument.

To establish Equation (28), define, for each $i \in [m]$,

$$\xi_i(x) \coloneqq d_{\mathrm{TV}}\big(p_i(\cdot \mid x), \hat{p}_i(\cdot \mid x)\big). \tag{29}$$

For any weights $\alpha \in \Delta_m$, convexity gives

$$d_{\mathrm{TV}}\left(\sum_i \alpha_i \hat{p}_i(\cdot \mid x), \sum_i \alpha_i p_i(\cdot \mid x)\right) \le \sum_i \alpha_i \xi_i(x) \le \max_i \xi_i(x). \tag{30}$$

Pairing mixtures with the same weights in both directions therefore yields

$$d_H\big(\widehat{\mathcal{C}}_e(x), \mathcal{C}^\star(x)\big) \le \max_{i \in [m]} \xi_i(x) \le \sum_{i=1}^m \xi_i(x). \tag{31}$$

Under the likelihood-ratio assumption, for every $i, l \in [m]$,

$$\mathbb{E}_{\mu_l} \xi_i(X) = \mathbb{E}_{\mu_i}\left[\frac{d\mu_l}{d\mu_i}(X)\, \xi_i(X)\right] \le \chi\, \mathbb{E}_{\mu_i} \xi_i(X). \tag{32}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mu_l} d_H\big(\widehat{\mathcal{C}}_e(X), \mathcal{C}^\star(X)\big) &\le \sum_{i=1}^{m} \mathbb{E}_{\mu_l} \xi_i(X) \\ &\le \chi \sum_{i=1}^{m} \mathbb{E}_{\mu_i} \xi_i(X) \\ &\le \chi m \sqrt{\frac{\ln 2}{2}\, \mathcal{D}(e)}.\end{aligned} \tag{33}$$

This proves Equation (28).

Finally, if $\mathcal{D}(e) = 0$, then Equation (27) gives $\xi_i(X) = 0$ μ_i-almost surely for every i. Under mutual absolute continuity, all these equalities hold simultaneously μ_j-almost surely for every j. The pointwise Hausdorff bound then proves exact recovery. □

C.4 The pooling floor and separation

Definition 8 (Pooled efficiency)**.** An encoder e producing code Z with finite pooled rate $I_{\boldsymbol{\lambda}}(X;Z) < \infty$ is *pooled-efficient* if no alternative encoder e' producing code Z' satisfies both $I_{\boldsymbol{\lambda}}(X;Z') \le I_{\boldsymbol{\lambda}}(X;Z)$ and $I_{\boldsymbol{\lambda}}(Z';Y) \ge I_{\boldsymbol{\lambda}}(Z;Y)$ with at least one inequality strict.

Every minimizer of finite rate of the standard pooled objective $I_{\boldsymbol{\lambda}}(X;Z) - \beta I_{\boldsymbol{\lambda}}(Z;Y)$ is pooled-efficient for any $\beta \ge 0$: for $\beta > 0$, any strictly dominating code would achieve a lower Lagrangian value; for $\beta = 0$, any minimizer has $I_{\boldsymbol{\lambda}}(X;Z) = 0$, forcing $I_{\boldsymbol{\lambda}}(Z;Y) = 0$ by the data-processing inequality.

We denote the operational quantities of interest by:

$$\mathcal{D}_{\boldsymbol{\lambda}}^{\mathrm{pool}} = \max_{i\in[m]} I_i\big(X;Y \mid \bar{T}_{\boldsymbol{\lambda}}(X)\big), \quad R_T = \max_{i\in[m]} H_i\big(T(X)\big), \quad D_{\max} = \max_{i\in[m]} I_i(X;Y), \tag{34}$$

representing the pooling distortion floor, the entropy rate of the credal target, and the distortion incurred by a non-informative constant representation, respectively.

Lemma 9 (Invariance under pooled-target conditioning)**.** *Given an encoder e, let $\kappa(\cdot \mid t) = P_{\boldsymbol{\lambda}}(Z \in \cdot \mid \bar{T}_{\boldsymbol{\lambda}}(X) = t)$ and define the conditioned encoder $\bar{e}(\cdot \mid x) = \kappa(\cdot \mid \bar{T}_{\boldsymbol{\lambda}}(x))$ with associated code $\bar{Z}$. Then*

$$I_{\boldsymbol{\lambda}}(\bar{Z};Y) = I_{\boldsymbol{\lambda}}(Z;Y), \qquad I_{\boldsymbol{\lambda}}(X;\bar{Z}) = I_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X);Z), \tag{35}$$

$$\textit{and} \quad I_{\boldsymbol{\lambda}}(X;Z) = I_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X);Z) + I_{\boldsymbol{\lambda}}(X;Z \mid \bar{T}_{\boldsymbol{\lambda}}(X)). \tag{36}$$

Proof. Let $G = \bar{T}_{\boldsymbol{\lambda}}(X)$. For any measurable set B in code space and label $y \in \mathcal{Y}$, noting that $p_{\boldsymbol{\lambda}}(y \mid X) = G(y)$ and $\mathbb{E}_{\boldsymbol{\lambda}}[e(B \mid X) \mid G] = \kappa(B \mid G)$, we have

$$P_{\boldsymbol{\lambda}}(Z \in B, Y = y) = \mathbb{E}_{\boldsymbol{\lambda}}\big[G(y)\, e(B \mid X)\big] = \mathbb{E}_{\boldsymbol{\lambda}}\big[G(y)\, \kappa(B \mid G)\big] = P_{\boldsymbol{\lambda}}(\bar{Z} \in B, Y = y), \tag{37}$$

which proves that (Z,Y) and $(\bar{Z},Y)$ are identically distributed, establishing the first equality. An identical calculation with an indicator $1\{G \in C\}$ shows that (G,Z) and $(G,\bar{Z})$ share the same law. Because $\bar{Z}$ is conditionally independent of X given G by construction, the chain rule gives $I_{\boldsymbol{\lambda}}(X;\bar{Z}) = I_{\boldsymbol{\lambda}}(X,G;\bar{Z}) = I_{\boldsymbol{\lambda}}(G;\bar{Z}) = I_{\boldsymbol{\lambda}}(G;Z)$. The final equation is the chain rule expansion of $I_{\boldsymbol{\lambda}}(X;Z)$ through G. □

Lemma 10 (Pooled efficiency of the pooled target)**.** *If $\bar{T}_{\boldsymbol{\lambda}}(X)$ is discrete with $H_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X)) < \infty$ and can be embedded into the code space $\mathcal{Z}$, then the deterministic encoder $Z = \bar{T}_{\boldsymbol{\lambda}}(X)$ is pooled-efficient.*

Proof. Applying Equation (23) with $G = \bar{T}_{\boldsymbol{\lambda}}(X)$ shows that $Z = \bar{T}_{\boldsymbol{\lambda}}(X)$ achieves the maximum possible mutual information $I_{\boldsymbol{\lambda}}(Z;Y) = I_{\boldsymbol{\lambda}}(X;Y)$ at a finite rate $H_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X))$. Now suppose an alternative encoder Z' achieves the same relevance: $I_{\boldsymbol{\lambda}}(Z';Y) = I_{\boldsymbol{\lambda}}(X;Y)$. Expanding $I_{\boldsymbol{\lambda}}(X,Z';Y)$ via the chain rule shows that $I_{\boldsymbol{\lambda}}(X;Y \mid Z') = 0$, meaning $p_{\boldsymbol{\lambda}}(\cdot \mid X) = P_{\boldsymbol{\lambda}}(Y \in \cdot \mid Z')$ almost surely. Thus $\bar{T}_{\boldsymbol{\lambda}}(X)$ is a deterministic function of Z' almost surely, implying $I_{\boldsymbol{\lambda}}(X;Z') \ge I_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X);Z') = H_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X))$. No representation can attain full pooled relevance at a strictly smaller rate. □

Theorem 11 (Full statement of Theorem 2)**.** *Let $\lambda_i > 0$ for all $i \in [m]$.*

(a) Rate-independent distortion floor. *Every pooled-efficient encoder e factors through the pooled target: there exists a kernel κ such that $e(\cdot \mid x) = \kappa(\cdot \mid \bar{T}_{\boldsymbol{\lambda}}(x))$ holds $\mu_{\boldsymbol{\lambda}}$-almost everywhere. Consequently, for every environment $i \in [m]$ and with $\pi_i(\cdot \mid t) = \mathbb{E}_i[p_i(\cdot \mid X) \mid \bar{T}_{\boldsymbol{\lambda}}(X) = t]$,*

$$\delta_i(e) \geq I_i\big(X; Y \mid \bar{T}_{\boldsymbol{\lambda}}(X)\big) = \mathbb{E}_i\, D_{\mathrm{KL}}\big(p_i(\cdot \mid X) \,\|\, \pi_i(\cdot \mid \bar{T}_{\boldsymbol{\lambda}}(X))\big). \tag{38}$$

In particular, $\mathcal{D}(e) \geq \mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}}$ holds uniformly across all code rates and trade-off parameters $\beta \geq 0$. Furthermore, when $\bar{T}_{\boldsymbol{\lambda}}(X)$ is discrete with $H_{\boldsymbol{\lambda}}(\bar{T}_{\boldsymbol{\lambda}}(X)) < \infty$, the encoder $Z = \bar{T}_{\boldsymbol{\lambda}}(X)$ attains this floor with equality in every environment.

(b) Strict positivity. *The pooling floor satisfies $\mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}} > 0$ if and only if there exists at least one environment i in which the true conditional distribution $p_i(\cdot \mid X)$ is not μ_i-almost surely a function of the pooled distribution $\bar{T}_{\boldsymbol{\lambda}}(X)$.*

(c) Achievability and vanishing distortion under CIB. *For any target distortion $D > 0$, the credal rate is finite ($\mathcal{R}_{\mathcal{P}}(D) < \infty$). Moreover, for any $\beta > 0$, $\varepsilon \geq 0$, and resolution $\xi \in (0, 1 - 1/n]$, every ε-minimizer e of F_β satisfies*

$$\mathcal{D}(e) \leq \omega(\xi) + \frac{\log_2 N(\xi) + \varepsilon}{\beta}. \tag{39}$$

Consequently for any family $\{e_\beta\}_{\beta>0}$ such that e_β is an ε_β-minimizer of F_β with $\varepsilon_\beta = o(\beta)$,

$$\lim_{\beta\to\infty} \mathcal{D}(e_\beta) = 0. \tag{40}$$

In addition, if $R_T < \infty$ and $D_{\max} > 0$, then $\mathcal{R}_{\mathcal{P}}(D) \leq R_T(1 - D/D_{\max})$ for all $D \in [0, D_{\max}]$, with near-minimizers satisfying $\mathcal{R}(e) \leq R_T + \varepsilon$ and $\mathcal{D}(e) \leq (R_T + \varepsilon)/\beta$.

(d) Strict separation between CIB and pooled representations. *Suppose $\mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}} > 0$. Then no pooled-efficient encoder can achieve a distortion below $\mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}}$, whereas $\mathcal{R}_{\mathcal{P}}(D) < \infty$ for all $D \in (0, \mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}})$ (and at $D = 0$ whenever $R_T < \infty$). In particular, for any parameter*

$$\beta > \frac{R_T + \varepsilon}{\mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}}}, \tag{41}$$

every ε-minimizer $e^\star$ of F_β achieves strictly smaller credal distortion than every pooled-efficient representation: $\mathcal{D}(e^\star) < \mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}} \leq \mathcal{D}(e)$.

Proof. Part (a): Let e be a pooled-efficient encoder, and construct $\bar{e}$ as in Lemma 9. Because $\bar{e}$ preserves the pooled relevance ($I_{\boldsymbol{\lambda}}(\bar{Z}; Y) = I_{\boldsymbol{\lambda}}(Z; Y)$) while reducing the rate by $I_{\boldsymbol{\lambda}}(X; Z \mid \bar{T}_{\boldsymbol{\lambda}}(X))$, pooled efficiency dictates that $I_{\boldsymbol{\lambda}}(X; Z \mid \bar{T}_{\boldsymbol{\lambda}}(X)) = 0$. Thus, $e(\cdot \mid x)$ depends on x only through $\bar{T}_{\boldsymbol{\lambda}}(x)$ for $\mu_{\boldsymbol{\lambda}}$-almost every x. Because $\mu_i \ll \mu_{\boldsymbol{\lambda}}$, this conditional factorization holds under every environment i. Under law P_i, the code forms the Markov chain $Y \to X \to \bar{T}_{\boldsymbol{\lambda}}(X) \to Z$. By the data-processing inequality, $I_i(Z; Y) \leq I_i(\bar{T}_{\boldsymbol{\lambda}}(X); Y)$, which expands through Equation (23) to give

$$\delta_i(e) = I_i(X; Y) - I_i(Z; Y) \geq I_i(X; Y) - I_i(\bar{T}_{\boldsymbol{\lambda}}(X); Y) = I_i(X; Y \mid \bar{T}_{\boldsymbol{\lambda}}(X)). \tag{42}$$

Taking the maximum over $i \in [m]$ establishes $\mathcal{D}(e) \geq \mathcal{D}^{\mathrm{pool}}_{\boldsymbol{\lambda}}$. When $\bar{T}_{\boldsymbol{\lambda}}(X)$ is discrete, Lemma 10 confirms that $Z = \bar{T}_{\boldsymbol{\lambda}}(X)$ achieves equality.

Part (b): This follows directly from the characterization in Equation (23), where $I_i(X; Y \mid \bar{T}_{\boldsymbol{\lambda}}(X)) = 0$ if and only if $p_i(\cdot \mid X) = \mathbb{E}_i[p_i(\cdot \mid X) \mid \bar{T}_{\boldsymbol{\lambda}}(X)]$ almost surely under μ_i.

Part (c): Finite rate $\mathcal{R}_{\mathcal{P}}(D) < \infty$ follows from Lemma 6. Because both rate and distortion are non-negative, the CIB objective provides the lower bound $\beta\mathcal{D}(e) \leq F_\beta(e)$. Evaluating F_β on the quantized code $Q_\xi(T(X))$ from Lemma 6 gives

$$\beta\mathcal{D}(e) \leq F_\beta(Q_\xi) + \varepsilon \leq \log_2 N(\xi) + \beta\omega(\xi) + \varepsilon, \tag{43}$$

which proves Equation (39). For every fixed $\xi > 0$,

$$\limsup_{\beta\to\infty} \mathcal{D}(e_\beta) \leq \omega(\xi), \tag{44}$$

since $\varepsilon_\beta/\beta \to 0$. Letting $\xi \downarrow 0$ and using $\omega(\xi) \to 0$ yields

$$\lim_{\beta\to\infty} \mathcal{D}(e_\beta) = 0. \tag{45}$$

When $R_T < \infty$, let $B \sim \text{Bernoulli}(q)$ be independent of (X, Y) under every environment, and define

$$Z_q = \begin{cases} T(X), & B = 1, \\ \dagger, & B = 0, \end{cases} \tag{46}$$

where $\dagger$ is a distinguished erasure symbol outside the range of T. Then

$$\mathcal{R}(Z_q) = qR_T, \qquad \mathcal{D}(Z_q) = (1-q)D_{\max}. \tag{47}$$

Choosing $q = 1 - D/D_{\max}$ proves the linear rate–distortion upper bound. Evaluating F_β directly at $T(X)$ gives $F_\beta(T) = R_T$, which bounds near-minimizers by $\max\{\mathcal{R}(e), \beta\mathcal{D}(e)\} \le F_\beta(e) \le R_T + \varepsilon$.

Part (d): Combining the lower bound $\mathcal{D}(e) \ge \mathcal{D}_{\boldsymbol{\lambda}}^{\text{pool}}$ from Part (a) with the explicit distortion bound for CIB minimizers in Part (c) immediately establishes the strict separation when $\beta > (R_T + \varepsilon)/\mathcal{D}_{\boldsymbol{\lambda}}^{\text{pool}}$. □

Remark 12 (Scope and implications)**.** (i) *Independence of rate metric:* Because the floor $\mathcal{D}_{\boldsymbol{\lambda}}^{\text{pool}}$ holds for any pooled-efficient code regardless of rate, the separation persists even if one measures the pooled encoder with respect to alternative, environment-agnostic rates.

(ii) *Geometric origins of the floor:* The floor $\mathcal{D}_{\boldsymbol{\lambda}}^{\text{pool}}$ is strictly positive if and only if, in at least one environment i, $p_i(\cdot \mid X)$ is not μ_i-almost surely a measurable function of $\bar{T}_{\boldsymbol{\lambda}}(X)$. This arises naturally in classification tasks with discrete features, demographic shifts, or multi-source sensor domains.

C.5 Consequence for uncertainty quantification

Let $\mathcal{C} \subset \Delta(\mathcal{Y})$ denote a closed, convex set of probability measures over labels. Following the uncertainty disaggregation framework introduced by Abellán et al. (2006); Hüllermeier & Waegeman (2021), total uncertainty, aleatoric uncertainty, and the resulting epistemic spread are given by:

$$\overline{H}(\mathcal{C}) = \max_{p\in\mathcal{C}} H(p), \qquad \underline{H}(\mathcal{C}) = \min_{p\in\mathcal{C}} H(p), \qquad W(\mathcal{C}) = \overline{H}(\mathcal{C}) - \underline{H}(\mathcal{C}). \tag{48}$$

When an ensemble of K hypotheses is deployed over m environments, the true credal set at input x is $\mathcal{C}^\star(x) = \text{conv}\{p_1(\cdot \mid x), \ldots, p_m(\cdot \mid x)\}$, with corresponding true uncertainty components denoted $U^\star(x) \in \{\text{TU}^\star(x), \text{AU}^\star(x), \text{EU}^\star(x)\}$.

Corollary 13 (Uncertainty estimation under pooled representations)**.** *Assume $\lambda_i > 0$ for every i, and let $\{e_k\}_{k=1}^K$ be pooled-efficient encoders. For fixed decoders q_{ik}, define*

$$\hat{p}_{ik}(y \mid x) = \int q_{ik}(y \mid z)\, e_k(dz \mid x) \tag{49}$$

Let $U(X)$ be any bounded measurable function of these readouts. Then

$$\mathbb{E}_{\boldsymbol{\lambda}}\big|U(X) - U^\star(X)\big| \ge \mathbb{E}_{\boldsymbol{\lambda}}\big|U^\star(X) - \text{med}_{\boldsymbol{\lambda}}(U^\star(X) \mid \bar{T}_{\boldsymbol{\lambda}}(X))\big|. \tag{50}$$

The right-hand side is strictly positive if and only if $U^\star(X)$ is not $\mu_{\boldsymbol{\lambda}}$-almost surely a measurable function of $\bar{T}_{\boldsymbol{\lambda}}(X)$.

If the input marginals are mutually absolutely continuous, a credally sufficient encoder with its environment-specific Bayes readouts recovers the true credal set and therefore its uncertainty summaries almost surely.

Proof. By Theorem 11(a), each encoder factors through $\bar{T}_{\boldsymbol{\lambda}}(X)$ almost surely. Every readout $\hat{p}_{ik}(\cdot \mid X)$, and hence $U(X)$, is therefore a measurable function of $\bar{T}_{\boldsymbol{\lambda}}(X)$. The conditional median minimizes conditional expected absolute error, proving Equation (50). Its expected absolute deviation is zero exactly when $U^\star(X)$ is almost surely a measurable function of $\bar{T}_{\boldsymbol{\lambda}}(X)$. The final claim follows from Proposition 7 using Bayes readouts and mutual absolute continuity. □

C.6 VARIATIONAL RELAXATION

Theorem 3 (Variational credal bound)**.** *For $i \in [m]$ and $k \in [K]$, define the variational rate $\widehat{R}_{ik}$ and the average decoder cross-entropy $\bar{L}_{ik}$ as*

$$\widehat{R}_{ik} \coloneqq \mathbb{E}_{\mu_i}\big[D_{\mathrm{KL}}\big(e_k(\cdot \mid X) \parallel r_{ik}\big)\big], \qquad \bar{L}_{ik} \coloneqq \frac{1}{J}\sum_{j=1}^{J} \mathbb{E}_{P_i,e_k}\big[-\log_2 q_{ij}(Y \mid Z)\big], \tag{10}$$

and the resulting bracket as

$$\widehat{b}_{ik} \coloneqq \frac{\widehat{R}_{ik}}{\beta} + \bar{L}_{ik} - L_i^{\mathrm{ref}}, \tag{11}$$

where $\mathbb{E}_{P_i,e_k}$ is the expectation over $(X,Y)\sim P_i$ and $Z\sim e_k(\cdot\,|\,X)$. If these quantities are finite, then

$$\beta\,\widehat{b}_{ik} \geq I_{ik}(X;Z) + \beta\,\delta_{ik}(e_k) - \beta\,\varepsilon_i^{\mathrm{ref}}, \qquad \varepsilon_i^{\mathrm{ref}} \coloneqq L_i^{\mathrm{ref}} - H_i(Y \mid X), \tag{12}$$

with equality if and only if r_{ik} equals the code marginal $\int \mu_i(dx)\, e_k(\cdot \mid x)$ and every $q_{ij}(\cdot \mid z)$ equals the Bayes decoder of environment i under e_k for almost every z.

Proof. For fixed (i,k,j), expanding the cross-entropy under $P_{ik}(x,y,z) = \mu_i(x)p_i(y \mid x)e_k(z \mid x)$ gives:

$$\mathbb{E}_{P_{ik}}[-\log_2 q_{ij}(Y \mid Z)] = \mathbb{E}_{P_{ik}}[-\log_2 p_{ik}(Y \mid Z)] + \mathbb{E}_{p_{ik}(z)}[D_{\mathrm{KL}}(p_{ik}(Y \mid z) \parallel q_{ij}(Y \mid z))] \tag{51}$$

$$= H_{ik}(Y \mid Z) + \mathbb{E}_{p_{ik}(z)}[D_{\mathrm{KL}}(p_{ik}(Y \mid z) \parallel q_{ij}(Y \mid z))]\,. \tag{52}$$

Subtracting $H_i(Y \mid X)$ and using $H_{ik}(Y \mid Z) - H_i(Y \mid X) = \delta_{ik}(e_k)$ gives:

$$\bar{L}_{ik} - H_i(Y \mid X) = \delta_{ik}(e_k) + \frac{1}{J}\sum_{j=1}^{J} \mathbb{E}_{p_{ik}(z)}[D_{\mathrm{KL}}(p_{ik}(Y \mid z) \parallel q_{ij}(Y \mid z))] \geq \delta_{ik}(e_k)\,. \tag{53}$$

Substituting $L_i^{\mathrm{ref}} = H_i(Y \mid X) + \varepsilon_i^{\mathrm{ref}}$, multiplying by β, and adding the variational rate identity $\widehat{R}_{ik} = I_{ik}(X;Z) + D_{\mathrm{KL}}(p_{ik} \parallel r_{ik}) \geq I_{ik}(X;Z)$ yields the result. The gap in the stated inequality is

$$D_{\mathrm{KL}}(p_{ik} \parallel r_{ik}) + \frac{\beta}{J}\sum_{j=1}^{J} \mathbb{E}_{p_{ik}(z)} D_{\mathrm{KL}}\big(p_{ik}(\cdot \mid z) \parallel q_{ij}(\cdot \mid z)\big). \tag{54}$$

Under the finiteness assumptions, equality holds exactly when all these nonnegative terms vanish. □

Proposition 14 (Environment-Specific Priors vs. Spurious Invariance)**.** *Let $S \sim \boldsymbol{\lambda} \in \Delta_m$ be an environment indicator, and let $p_{\boldsymbol{\lambda}}(z) = \sum_{i=1}^m \lambda_i p_i(z)$ be the pooled representation marginal. For any shared prior r on $\mathcal{Z}$, the mixture rate bound decomposes as:*

$$\sum_{i=1}^{m} \lambda_i\, \mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e(\cdot \mid X) \parallel r)] = \sum_{i=1}^{m} \lambda_i I_i(X;Z) + I_{\boldsymbol{\lambda}}(S;Z) + D_{\mathrm{KL}}(p_{\boldsymbol{\lambda}} \parallel r)\ . \tag{55}$$

In particular,

$$\max_{i\in[m]} \mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e(\cdot \mid X) \parallel r)] \geq \sup_{\boldsymbol{\lambda}\in\Delta_m} \left[\sum_{i=1}^{m} \lambda_i I_i(X;Z) + I_{\boldsymbol{\lambda}}(S;Z)\right]. \tag{56}$$

Proof. By the variational rate expansion applied to environment i,

$$\mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e(\cdot \mid X) \parallel r)] = I_i(X;Z) + D_{\mathrm{KL}}(p_i \parallel r)\ . \tag{57}$$

Weighting by λ_i and summing over $i \in [m]$ yields:

$$\sum_{i=1}^{m} \lambda_i\, \mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e(\cdot \mid X) \parallel r)] = \sum_{i=1}^{m} \lambda_i I_i(X;Z) + \sum_{i=1}^{m} \lambda_i D_{\mathrm{KL}}(p_i \parallel r)\ . \tag{58}$$

By the compensation identity for relative entropy against a mixture $p_{\boldsymbol{\lambda}} = \sum_i \lambda_i p_i$,

$$\sum_{i=1}^{m} \lambda_i D_{\mathrm{KL}}(p_i \,\|\, r) = \sum_{i=1}^{m} \lambda_i D_{\mathrm{KL}}(p_i \,\|\, p_{\boldsymbol{\lambda}}) + D_{\mathrm{KL}}(p_{\boldsymbol{\lambda}} \,\|\, r) = I_{\boldsymbol{\lambda}}(S;Z) + D_{\mathrm{KL}}(p_{\boldsymbol{\lambda}} \,\|\, r)\ , \quad (59)$$

where $I_{\boldsymbol{\lambda}}(S;Z) = \sum_i \lambda_i D_{\mathrm{KL}}(p_i \,\|\, p_{\boldsymbol{\lambda}})$ is the mutual information between the environment index S and the representation Z under the joint law $P(S = i, Z \in dz) = \lambda_i p_i(z)dz$. Set $A_i = \mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e(\cdot \mid X) \,\|\, r)]$. If $\max_i A_i = \infty$, the claimed inequality is immediate. Otherwise, for every $\boldsymbol{\lambda} \in \Delta_m$,

$$\begin{aligned} \max_i A_i &\geq \sum_i \lambda_i A_i \\ &= \sum_i \lambda_i I_i(X;Z) + I_{\boldsymbol{\lambda}}(S;Z) + D_{\mathrm{KL}}(p_{\boldsymbol{\lambda}} \,\|\, r) \\ &\geq \sum_i \lambda_i I_i(X;Z) + I_{\boldsymbol{\lambda}}(S;Z). \end{aligned} \quad (60)$$

The left-hand side does not depend on $\boldsymbol{\lambda}$. Taking the supremum over Δ_m proves the inequality. □

Proposition 14 shows that adopting a single shared prior $r = \mathcal{N}(0, I)$ across environments, as standard VIB does, implicitly penalizes $I_{\boldsymbol{\lambda}}(S;Z)$, driving representations from different environments to overlap. While domain-invariant models enforce $I_{\boldsymbol{\lambda}}(S;Z) \to 0$, credal sufficiency (Definition 1) does not require environment invariance; forcing it collapses distinguishing environment features into apparent label noise. We therefore assign a separate, learnable Gaussian-mixture prior r_{ik} to each pair (i,k), fitted on sampled latents $z \sim e_k(\cdot \mid x)$ via maximum likelihood, minimizing $D_{\mathrm{KL}}(p_{ik} \,\|\, r_{ik})$ without the explicit shared-prior penalty $I_{\boldsymbol{\lambda}}(S;Z)$ in this decomposition.

Extension to the entire credal sets. For $\boldsymbol{\lambda} \in \Delta_m$ and an encoder e, define the objective under the mixture $P_{\boldsymbol{\lambda}}$ and its worst case over the data credal set,

$$F_\beta^{\boldsymbol{\lambda}}(e) \coloneqq I_{\boldsymbol{\lambda}}(X;Z) + \beta\, I_{\boldsymbol{\lambda}}(X;Y \mid Z), \qquad F_\beta^{\mathcal{P}}(e) \coloneqq \sup_{\boldsymbol{\lambda} \in \Delta_m} F_\beta^{\boldsymbol{\lambda}}(e), \quad (61)$$

where $I_{\boldsymbol{\lambda}}$ is mutual information under $P_{\boldsymbol{\lambda}}(dx, y)\, e(dz \mid x)$. With $\mathbf{1}_i$ the i-th standard basis vector, $P_{\mathbf{1}_i} = P_i$ and $F_\beta^{\mathbf{1}_i}(e) = I_i(X;Z) + \beta\, \delta_i(e)$, so $F_\beta^{\mathcal{P}} \geq F_\beta$. Every posterior $\sum_k \alpha_k Q_k$ with $\boldsymbol{\alpha} \in \Delta_K$ in the model credal set induces the encoder $e_{\boldsymbol{\alpha}} = \sum_k \alpha_k e_k$. To bound $F_\beta^{\boldsymbol{\lambda}}(e_{\boldsymbol{\alpha}})$ with the trained priors and decoders, let $\bar{r}_l = \frac{1}{K}\sum_{k=1}^{K} r_{lk}$ and $\bar{q}_l = \frac{1}{J}\sum_{j=1}^{J} q_{lj}$, and define

$$r_k^{\boldsymbol{\lambda}} = \sum_{l=1}^{m} \lambda_l r_{lk}, \quad \pi_l^{\boldsymbol{\lambda}}(z) = \frac{\lambda_l \bar{r}_l(z)}{\sum_{l'=1}^{m} \lambda_{l'} \bar{r}_{l'}(z)}, \quad q^{\boldsymbol{\lambda}}(y \mid z) = \sum_{l=1}^{m} \pi_l^{\boldsymbol{\lambda}}(z)\, \bar{q}_l(y \mid z), \quad (62)$$

with $\pi_l^{\boldsymbol{\lambda}}(z) \coloneqq \lambda_l$ where the denominator vanishes. Here $r_k^{\boldsymbol{\lambda}}$ is the code prior of head k under $P_{\boldsymbol{\lambda}}$, $\pi_l^{\boldsymbol{\lambda}}(z)$ is the posterior probability of environment l given code z implied by the priors, and $q^{\boldsymbol{\lambda}}$ is the corresponding mixture decoder. The *mixture bracket* of head k is

$$\widehat{b}_k^{\boldsymbol{\lambda}} \coloneqq \sum_{i=1}^{m} \lambda_i \Big(\tfrac{1}{\beta}\, \mathbb{E}_{\mu_i}\big[D_{\mathrm{KL}}\big(e_k(\cdot \mid X) \,\|\, r_k^{\boldsymbol{\lambda}}\big)\big] + \mathbb{E}_{P_i, e_k}\big[-\log_2 q^{\boldsymbol{\lambda}}(Y \mid Z)\big] - L_i^{\mathrm{ref}} \Big). \quad (63)$$

Proposition 15 (Variational bounds over the model and data credal sets)**.** *Let the priors r_{ik} have densities with respect to a common reference measure on $\mathcal{Z}$ (Lebesgue measure for Gaussian codes), assume that all brackets and L_i^{ref} are finite, and let $\varepsilon_{\max}^{\mathrm{ref}} \coloneqq \max_{i \in [m]} \varepsilon_i^{\mathrm{ref}}$. Then:*

(i) for every $\boldsymbol{\alpha} \in \Delta_K$ and $\boldsymbol{\lambda} \in \Delta_m$, $F_\beta^{\boldsymbol{\lambda}}(e_{\boldsymbol{\alpha}}) \leq \beta \sum_{k=1}^{K} \alpha_k \widehat{b}_k^{\boldsymbol{\lambda}} + \beta \sum_{i=1}^{m} \lambda_i \varepsilon_i^{\mathrm{ref}}$;

(ii) $\widehat{b}_k^{\mathbf{1}_i} \leq \widehat{b}_{ik}$ for every $i \in [m]$ and $k \in [K]$;

(iii) the objective of Equation (13) bounds the worst case over the model credal set,

$$\sup_{\boldsymbol{\alpha} \in \Delta_K} F_\beta(e_{\boldsymbol{\alpha}}) \leq \beta \sum_{k=1}^{K} \max_{i \in [m]} \widehat{b}_{ik} + K\beta\, \varepsilon_{\max}^{\mathrm{ref}}, \quad (64)$$

and the worst case over both credal sets satisfies

$$\sup_{\boldsymbol{\alpha}\in\Delta_K} F_\beta^{\mathcal{P}}(e_{\boldsymbol{\alpha}}) \le \beta \sum_{k=1}^{K} \sup_{\boldsymbol{\lambda}\in\Delta_m} \widehat{b}_k^{\boldsymbol{\lambda}} + K\beta\,\varepsilon_{\max}^{\mathrm{ref}}. \tag{65}$$

Proof. Introduce an environment index S with $P(S=i)=\lambda_i$, $X \mid S=i \sim \mu_i$, $Y \mid (X, S=i) \sim p_i(\cdot \mid X)$ and $Z \mid X \sim e_{\boldsymbol{\alpha}}(\cdot \mid X)$, so that (X, Y, Z) has law $P_{\boldsymbol{\lambda}}(dx, y)\, e_{\boldsymbol{\alpha}}(dz \mid x)$.

Rate. For any probability density r on $\mathcal{Z}$, $\mathbb{E}_{\mu_{\boldsymbol{\lambda}}}[D_{\mathrm{KL}}(e_{\boldsymbol{\alpha}}(\cdot \mid X) \,\|\, r)] = I_{\boldsymbol{\lambda}}(X;Z) + D_{\mathrm{KL}}(p_{\boldsymbol{\lambda}} \,\|\, r)$, where $p_{\boldsymbol{\lambda}}$ is the code marginal. Choosing $r = \sum_k \alpha_k r_k^{\boldsymbol{\lambda}}$ and using joint convexity of the Kullback–Leibler divergence gives $I_{\boldsymbol{\lambda}}(X;Z) \le \sum_k \alpha_k \sum_i \lambda_i\, \mathbb{E}_{\mu_i}[D_{\mathrm{KL}}(e_k(\cdot \mid X) \,\|\, r_k^{\boldsymbol{\lambda}})]$.

Distortion. Since $\mathcal{Y}$ is finite and $Y \to X \to Z$ under $P_{\boldsymbol{\lambda}}$, $I_{\boldsymbol{\lambda}}(X;Y \mid Z) = H_{\boldsymbol{\lambda}}(Y \mid Z) - H_{\boldsymbol{\lambda}}(Y \mid X)$. The cross-entropy of any decoder upper-bounds the conditional entropy, and $q^{\boldsymbol{\lambda}}$ does not depend on k, so $H_{\boldsymbol{\lambda}}(Y \mid Z) \le \sum_k \alpha_k \sum_i \lambda_i\, \mathbb{E}_{P_i, e_k}[-\log_2 q^{\boldsymbol{\lambda}}(Y \mid Z)]$. Conditioning on S does not increase entropy, hence $H_{\boldsymbol{\lambda}}(Y \mid X) \ge H_{\boldsymbol{\lambda}}(Y \mid X, S) = \sum_i \lambda_i \big(L_i^{\mathrm{ref}} - \varepsilon_i^{\mathrm{ref}}\big)$. Adding the rate bound to β times the distortion bound proves (i).

Generators. At $\boldsymbol{\lambda} = \mathbf{1}_i$, $r_k^{\mathbf{1}_i} = r_{ik}$. Finiteness of $\widehat{R}_{ik}$ implies $e_k(\cdot \mid x) \ll r_{ik} \le K\bar{r}_i$, so $q^{\mathbf{1}_i} = \bar{q}_i$ almost surely, and convexity of $-\log_2$ gives $-\log_2 \bar{q}_i \le \frac{1}{J}\sum_j (-\log_2 q_{ij})$, which proves (ii).

Suprema. By (i) with $\boldsymbol{\lambda} = \mathbf{1}_i$ and (ii), $F_\beta^{\mathbf{1}_i}(e_{\boldsymbol{\alpha}}) \le \beta \max_k \widehat{b}_{ik} + \beta\, \varepsilon_i^{\mathrm{ref}}$, hence $F_\beta(e_{\boldsymbol{\alpha}}) \le \beta \max_k s_k + \beta\, \varepsilon_{\max}^{\mathrm{ref}}$ with $s_k := \max_i \widehat{b}_{ik}$. Applying (i) with $\boldsymbol{\alpha} = \mathbf{1}_k$, $\boldsymbol{\lambda} = \mathbf{1}_i$ and $F_\beta^{\mathbf{1}_i} \ge 0$ gives $\widehat{b}_{ik} \ge -\varepsilon_i^{\mathrm{ref}}$, so $s_k + \varepsilon_{\max}^{\mathrm{ref}} \ge 0$ for every k and therefore $\max_k s_k \le \sum_k s_k + (K-1)\,\varepsilon_{\max}^{\mathrm{ref}}$. This proves the first inequality in (iii). The second follows in the same way from (i), with $s_k := \sup_{\boldsymbol{\lambda}} \widehat{b}_k^{\boldsymbol{\lambda}} \ge \widehat{b}_k^{\mathbf{1}_i} \ge -\varepsilon_i^{\mathrm{ref}}$. □

The objective of Equation (13) minimizes the right-hand side of the first inequality in (iii), so it controls the worst case over the model credal set and over the generators of $\mathcal{P}$. It does not minimize the right-hand side of Equation (65), which instead provides a bound over the entire data credal set that can be evaluated on a trained model. The mixture decoder uses the head-averaged priors $\bar{r}_l$ so that it does not depend on the encoder head; this makes the cross-entropy linear in $\boldsymbol{\alpha}$.

D THE COLORED-MNIST DATASET AND UNCERTAINTY MAP

This section describes how Colored-MNIST is built, how its two environments are constructed, why the true uncertainty of every image is known exactly, and how the uncertainty map of Figure 2 is computed for the ground truth.

Construction. Colored-MNIST follows Arjovsky et al. (2019): every MNIST digit is turned into a binary classification example in which both the shape of the digit and its color predict the label, with a reliability that depends on the environment. For an image of digit $d \in \{0, \dots, 9\}$, three variables are sampled in turn:

1. the *shape label* is $s = \mathbf{1}\{d < 5\}$, a deterministic function of the digit;
2. the *label* Y equals s with probability 0.75 and $1 - s$ otherwise, so the shape predicts the label with 25% label noise in every environment;
3. the *color* $c \in \{0, 1\}$ equals Y with probability $a_i(d)$, and $1 - Y$ otherwise, where $a_i(d)$ is the color agreement of environment $i \in [m]$.

The 28×28 grayscale image is sub-sampled to 14×14 (every other row and column), scaled to $[0, 1]$ and written into channel c of a two-channel image whose other channel is zero; we draw channel 0 red and channel 1 green. The color is thus a feature that is generated from the label, as in the spurious-correlation benchmarks, but here it does not merely become less reliable in some environments: its relation to the label is reversed on part of the input space.

Environment characterization. The environments differ only in their color agreement:

$$a_1(d) = 0.9 \text{ for every } d, \qquad a_2(d) = \begin{cases} 0.9, & d \ge 5, \\ 0.1, & d < 5. \end{cases} \tag{66}$$

In environment 1 the color agrees with the label 90% of the time on every digit. In environment 2 it does so on digits 5–9 and is reversed on digits 0–4, where it agrees with the label only 10% of the time. The two environments therefore label digits 5–9 by the same law and disagree on digits 0–4, where a color that points one way in environment 1 points the other way in environment 2. Because the image reveals its digit d and its color c, Bayes' rule gives the labeling law of each environment in closed form,

$$p_i(Y = s \mid x) = \begin{cases} \dfrac{0.75\,a}{0.75\,a + 0.25(1-a)}, & c = s, \\ \dfrac{0.75(1-a)}{0.75(1-a) + 0.25\,a}, & c \neq s, \end{cases} \qquad a = a_i(d), \tag{67}$$

which is 0.964 or 0.250 for $a = 0.9$ and the reverse for $a = 0.1$. The color agrees with the shape label with probability $0.75a + 0.25(1-a)$, i.e. 0.7 for $a = 0.9$ and 0.3 for $a = 0.1$. Table 5 lists the four kinds of input that this produces. A predictor that ignores color and predicts the shape label is right 75% of the time in both environments, while a predictor that uses color can reach 90% in either environment alone, but only if it knows which one it is in. The conditional entropy is

$$H_i(Y \mid X) = 0.7\,h(0.964) + 0.3\,h(0.25) \approx 0.399\,, \tag{68}$$

where $h(\cdot)$ is the binary entropy function introduced in Equation (24). This corresponds to the floor L_i^{ref} of the objective. The closed form assumes that the digit can be read off the image; where it cannot (e.g., due to degradation of the image), the true conditional entropy can only increase due to the AU induced by the degradation. Thus the given floor is a conservative lower bound.

Table 5: The four kinds of Colored-MNIST input: their frequency within the digit group and the probability that the label equals the shape label, in each environment.

		Environment 1		Environment 2	
Digits	Color	frequency	$p_1(Y = s \mid x)$	frequency	$p_2(Y = s \mid x)$
5–9	$c = s$	0.7	0.964	0.7	0.964
5–9	$c \neq s$	0.3	0.25	0.3	0.25
0–4	$c = s$	0.7	0.964	0.3	0.25
0–4	$c \neq s$	0.3	0.25	0.7	0.964

Splits. The $60{,}000$ official MNIST training images are split at random, disjointly and equally, between the two environments ($30{,}000$ each), and each half is colored by its environment's rule; 20% of each environment's training data ($6{,}000$ images) is held out as a validation split, used only for early stopping and for calibrating the label-shift baseline. The $10{,}000$ official test images are split and colored the same way ($5{,}000$ per environment). The labels and colors are sampled once per seed. During training only, each image is translated at random by up to 2 pixels (zero padding, both channels together); the labeling laws depend only on the digit and the color, which translation preserves, so the environments are unchanged.

Layout. Every panel of the map shows the same $10{,}000$ test images, the $5{,}000$ of each environment together, each at the same position; nothing is subsampled or selected. The positions come from a two-dimensional embedding of the raw pixels: each image is flattened into its $2 \times 14 \times 14 = 392$ pixel values, centered, and projected onto its leading 50 principal components, and t-SNE (van der Maaten & Hinton, 2008) with perplexity 40, initialized by PCA and seeded, embeds these in the plane. The layout sees only pixels: no label, environment or model output enters it, so it cannot favor one method. We use t-SNE because the map is meant to show *where* in the input space each kind of uncertainty occurs, and in Colored-MNIST where is determined by the digit and the color: all four quantities of Table 5, and hence the true uncertainty, are functions of (d, c). t-SNE preserves neighborhoods, so images of the same digit and color form one compact island, and each kind of input becomes a set of islands that can be read at a glance. Because a red and a green copy of the same digit share no nonzero pixel, every digit splits into two islands, one per color. A linear projection such as PCA alone preserves global variance instead, and overlaps the digits. As usual with t-SNE, only the grouping is meaningful; distances between islands and their sizes are not.

Ground truth. Since both environments' labeling laws are known in closed form, we evaluate $p_1(\cdot \mid x)$ and $p_2(\cdot \mid x)$ at every test image, whichever environment it was drawn in. The true credal set

at x is $\mathcal{C}^\star(x) = \mathrm{conv}(\{p_1(\cdot \mid x), p_2(\cdot \mid x)\})$, and the ground-truth panel reads it out by Definition 4 with the true laws as vertices: $\mathrm{TU}^\star(x)$ is the largest entropy over $\mathcal{C}^\star(x)$, which for two labels is 1 bit if the interval of $p(Y = 1 \mid x)$ contains $\frac{1}{2}$ and the larger endpoint entropy otherwise; $\mathrm{AU}^\star(x) = \min_i h(p_i(\cdot \mid x))$, attained at a vertex by concavity; $\mathrm{EU}^\star_{\mathrm{input}}(x) = \mathrm{TU}^\star(x) - \mathrm{AU}^\star(x)$; and $\mathrm{EU}^\star_{\mathrm{model}}(x) = 0$, since the laws are known and there is no model hypothesis to be uncertain about. By Table 5, the test images fall into three groups:

- digits 5–9 whose color agrees with the shape label: $p_1 = p_2$ with $p(Y = s \mid x) = 0.964$, so $\mathrm{AU}^\star = h(0.964) = 0.22$ bits and $\mathrm{EU}^\star_{\mathrm{input}} = 0$;
- digits 5–9 whose color disagrees: $p_1 = p_2$ with $p(Y = s \mid x) = 0.250$, so $\mathrm{AU}^\star = h(0.25) = 0.81$ bits and $\mathrm{EU}^\star_{\mathrm{input}} = 0$; this label noise cannot be reduced, even with the environment known;
- digits 0–4 of either color: the two environments assign $p(Y = s \mid x) = 0.964$ and 0.250, in an order that depends on the color, so $\mathcal{C}^\star(x) = [0.250, 0.964]$ contains $\frac{1}{2}$, $\mathrm{TU}^\star = 1$ bit, $\mathrm{AU}^\star = 0.22$ bits and $\mathrm{EU}^\star_{\mathrm{input}} = 0.78$ bits.

In the deployment panel, the ground truth's default decision is the pooled Bayes rule $\hat{y}(x) = \arg\max_y \frac{1}{2} \sum_i p_i(y \mid x)$, which is the best environment-agnostic rule here up to randomization (worst-environment expected accuracy 0.816 against 0.817, the best environment-agnostic accuracy on the test split). After deployment in environment i it switches to that environment's Bayes rule $y^\star_i(x) = \arg\max_y p_i(y \mid x)$. This gains $+0.25$ in expected accuracy on every image of digits 0–4 (nothing in the environment whose rule the pooled rule already follows, $0.75 - 0.25 = 0.5$ in the other) and nothing on digits 5–9, where the environments agree. The worst-environment expected accuracy thus rises from 0.816 to 0.897, the known-environment ceiling of the test split.

Uncertainty panels (left). For VCIB, the predictive vertices $\hat{p}_{i\ell}(y \mid x) = \mathbb{E}_{z \sim e_k(\cdot \mid x)} q_{ij}(y \mid z)$, with $i \in \{1, 2\}$ and $\ell = (k, j)$ over $(K, J) = (5, 3)$, are estimated at every test image by averaging 32 joint draws of the network weights and the code, and decomposed by Equation (15) into TU, AU, $\mathrm{EU}_{\mathrm{input}}$ and $\mathrm{EU}_{\mathrm{model}}$. VIB has a single predictive distribution, whose entropy is its only uncertainty score (Section 5); we draw it as AU, which is the residual it would receive under the decomposition. Each image is colored by mixing the colors of AU (blue), $\mathrm{EU}_{\mathrm{input}}$ (red) and $\mathrm{EU}_{\mathrm{model}}$ (purple) in proportion to their shares of the total, then blending with white so that the color strength is $0.12 + 0.88 \cdot \min\{\mathrm{TU}(x), 1\}$; the floor keeps low-uncertainty images visible, and $\mathrm{TU} \leq 1$ bit for two labels. The images are drawn in random order, so that neither environment covers the other.

Deployment panels (right). Each image is shaded by its expected change in accuracy from the default decision to deployment without labels, averaged over the two environments it may be deployed in,

$$g(x) = \frac{1}{m} \sum_{i=1}^{m} \Big[\mathbb{E}_B \, p_i\big(\hat{y}_B(x) \mid x\big) - p_i\big(\hat{y}(x) \mid x\big) \Big], \tag{69}$$

where accuracies are expectations under the true law $p_i(\cdot \mid x)$ rather than realized labels, $\hat{y}$ is the method's default decision (the credal minimax rule with code-space environment weights for VCIB, the arg max for VIB; Appendix A.7) and $\hat{y}_B$ its decision after observing an unlabeled batch B drawn in environment i. For VCIB, B is a random batch of 100 test images of environment i (2,000 batches per environment). Its posterior π_B over environments is the softmax of the summed code log-likelihoods under the moment-matched Gaussian priors r^G_{ik} (Appendix A.6); the credal set at x collapses to $\{\sum_{i'} \pi_B(i') q_{i'} : q_{i'} \in \widehat{\mathcal{C}}_{i'}(x)\}$, and x is assigned the E-admissible label with the largest lower probability over it (Troffaes, 2007); the expectation over B runs over these random batches. This is the protocol of WC-$\mathrm{Acc}_{\mathrm{int}}$ (Appendix A.7), taken in expectation over batches where the metric partitions the test set once, so that each image receives a smooth expected gain rather than the outcome of one partition. VIB, which cannot infer the environment, receives the label-free adaptation of the metric: bias-corrected temperature scaling fitted on the pooled validation split, the class prior of environment i estimated by EM on its whole unlabeled test sample, and the calibrated predictive reweighted by the ratio of that prior to the training class prior; this decision does not depend on a batch. Gains are drawn on a diverging scale from -0.25 (red) through 0 (gray) to $+0.25$ (green), the ground truth's gain on a conflicting image, with the largest gains in absolute value drawn on top.