THE CREDAL INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

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.

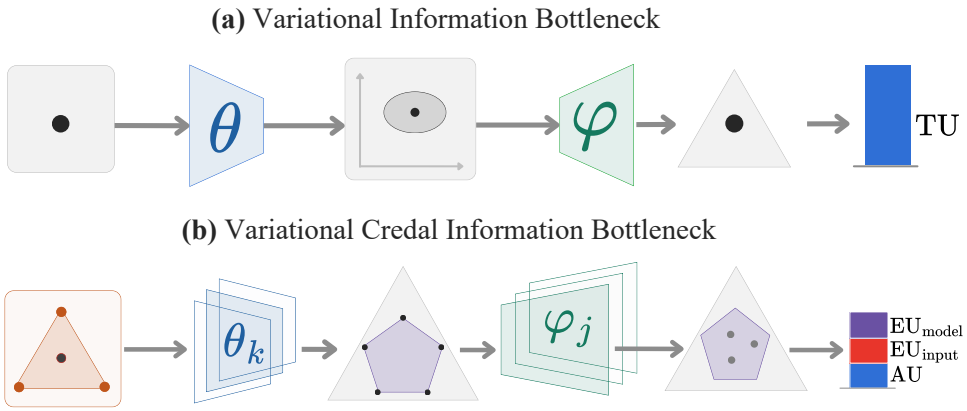


Figure 1: The IB compresses a point-valued input by optimizing a single pipeline of processing functions based on a compression–prediction tradeoff under one pooled law. The proposed CIB, instead, uses credal sets to represent the uncertainty in the input distribution and the processing functions. To 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.

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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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.

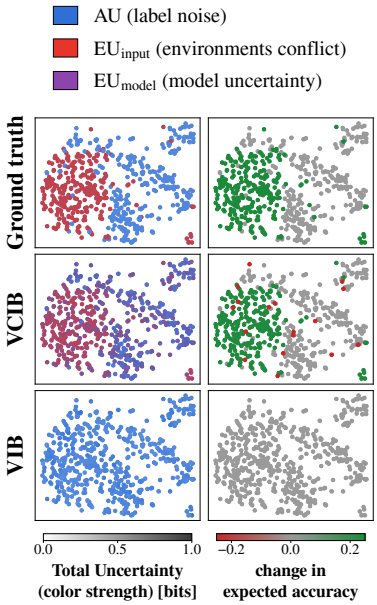


Figure 2: (**left**) Disentanglement of AU, input EU and processing EU at the task output using VIB (bottom) and VCIB (middle) against the ground-truth (top); (**right**) expected accuracy gain when accounting for the EU in making the decision.

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 $Z\sim e_k(\cdot\,|\,X)$. 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 $\mathcal{L}^{\tau}_{\mathrm{VCIB}}$ with temperature $\tau > 0$:

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

Since $\mathcal{L}_{\mathrm{VCIB}} \leq \mathcal{L}^{\tau}_{\mathrm{VCIB}} \leq \mathcal{L}_{\mathrm{VCIB}} + K\tau \ln m$, 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:

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

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).

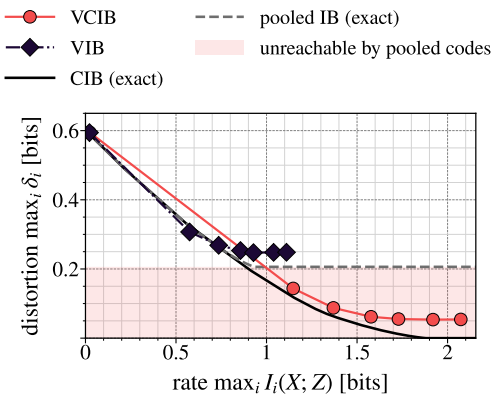


Figure 3: Rate–distortion plane on C-MNIST obtained by sweeping $\beta \in [2, 16]$. VCIB breaks the distortion lower bound of pooled bottlenecks.

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 $\mathcal{D}$ 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.

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	$1.960_{\pm.033}$	$3.291_{\pm.069}$	$0.165_{\pm.007}$	$\mathbf{0.773}_{\pm.007}$	$\mathbf{0.834}_{\pm.005}$	$\mathbf{0.068}_{\pm.004}$	$0.123_{\pm.003}$

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 $1.5\times$ to $3.4\times$. 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. EU$_{\text{input}}$ 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 VCIB$_{\text{mod}}$ 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.

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.