We thank the reviewers for their insightful feedback. We are pleased that the reviewers found our paper well-written [R1,R2,R4], thorough [R1,R3], and recognized the strength of our theoretical results [R3,R4]. We are especially encouraged by the comments of R3 and R4 concerning the possible impact of our work beyond the information bottleneck (IB). We first address two recurrent questions and proceed by addressing reviewer-specific comments.

G.1. Clarifying the definition of $\mathrm{Dec}(\mathrm{X}, y)$ [R1,R2,R4]. We simplified the notation and added an example to better illustrate this concept. We summarize these changes here. Suppose we are classifying cats versus dogs, then the y (say, "cat") decomposition of images X, corresponds to taking all "cats" and labeling them with every possible binary labels. Formally, let X_y be a r.v. s.t. $P_{\mathrm{X}_y} = P_{\mathrm{X}|y}$ (the "cat" r.v.). Then $\mathrm{Dec}(\mathrm{X}, y) := \{\mathrm{N} | \exists t' : \mathcal{X} \to \mathcal{Y} \; s.t. \; \mathrm{N} = t'(\mathrm{X}_y)\}$. We emphasize that: (i) N is a "labeling" of X_y so it takes value in $\mathcal{Y}$ ({"cat","dog"}) so we can predict it using a binary classifier in $\mathcal{V}$; (ii) the formal definition does not assume that the underlying labeling of Y is deterministic.

G.2. Practicality of our Decodable IB (DIB) [R1,R2]. We extended the discussion about practical optimization (summary of Appx. D6, D7) and estimation (summary of Appx. D4) in the main paper. We emphasize that: (i) the min-max problem can be well optimized using joint gradient descent ascent (c.f. Fig. 10 and L. 907), it does not require unrolling steps; (ii) we get good performance by only predicting a few (4) "labels" $\mathrm{N} \in \mathrm{Dec}(\mathrm{X}, y)$ (c.f. Appx. D6). As a result, training a network with DIB is approximately as computationally efficient as training a standard network or the variational IB — and much more efficient than IB. We will **release our code** to help practitioners use DIB.

R.1. We would like to thank you for the helpful review and useful comments which we adressed in the revised manuscript. We are pleased that you found the idea interesting, well-motivated, and the analysis thorough.

- "*Can we prove such representation exists?*": **Yes**, they always exist because they are defined as maximizers/minimizers of some quantity in a finite set. This is a very good point, which we now discuss in the paper.
- "*[...] I found part 3.2 unclear [...]*": Please refer to G.1. for $\mathrm{Dec}(\mathrm{X}, y)$ and N. We also added an algorithm on the optimization of $\mathrm{I}_{\mathcal{V}}[\mathrm{Z} \to \mathrm{Dec}(\mathrm{X}, \mathrm{Y})]$. For cat-dog classification: (i) assign each cat with a binary label; (ii) optimize the encoder such that classifiers in $\mathcal{V}$ *cannot* predict these labels, i.e., distinguish between cats; (iii) repeat for dogs.
- "*[...] Does the method induce additional complexity [...]?*": **No**, please refer to G.2.
- "*Consistency between [...] proposition 1. [...] Fig. 3 [...]*": We would like to clarify that Prop. 1 only concerns the *existence* of an optimal predictor. The generalization of *any* ERMs requires $\mathcal{V}$-minimality (Theorem 1). The best loss in Fig. 3a occurs at approximate $\mathcal{V}$-minimality (Fig. 3b) suggesting that ensuring generalization is more valuable than improving the best possible performance. We thus do not believe that Fig. 3 contradicts the theory.
- "*Usefulness of the method: [...] traditional IB [...]?*": Due to space constraints, we decided to highlight results that validate the theory, and thank the reviewer for acknowledging the importance of such results. We nevertheless emphasize that many of the results asked by the reviewer are in the appendices. Table 3 (Appx. D10) evaluates DIB as a regularizer on CIFAR10MNIST (CIFAR10 with overlaid MNIST images to evaluate robustness to "nuisances") and MNIST (to replicate the results from VIB's paper). We see that DIB outperforms VIB on both tasks, i.e., it gives rise to a better classifier and is more robust to nuisances. We added a small discussion about this in the main text.

R.2. We thank you for the kind and helpful review, we are pleased that you found the paper well-written and the approach elegant. We have put significant effort into incorporating both your suggestions into the revised manuscript.

- "*[...] the main text could benefit [...] practicality of DIB compared to IB [...]*": Please refer to G.2.
- "*[...] more clear, maybe through an example, is in explaining equation 4. [...]*": Please refer to G.1.

R.3. We thank the reviewer for reading the paper thoroughly and for providing a kind and detailed assessment. We are encouraged that you found the idea novel, the theoretical results strong, and the empirical evaluation thorough.

- "*Although Theorem 1 is a strong result, it assumes [...]*": We believe that our assumptions **do not significantly limit the applicability of our theory**: (i) Piece-wise universality: we now **relaxed this assumption** so that the proof holds for any model using a softmax layer with biases on that layer; (ii) Finite sample spaces: although the continuous setting is theoretically interesting, it (arguably) is less important in practice as everything must be discretized on a computer; (iii) Deterministic labeling: examples are usually seen once per dataset, i.e., no two same images are given different labels. In such cases the labeling is deterministic. We acknowledge that this is not necessarily true in the real world and hope to extend our proof in future work. We emphasized this assumption in the revised paper.

R.4. We thank you for your helpful review, and are pleased you think the approach could constitute a significant advancement for IB and be broadly relevant for methods based on information theory (e.g. variants of VAE and GAN).

- "*The choice of the appropriate predictive family $\mathcal{V}$ [...]?*": There is a trade-off. A larger $\mathcal{V}$ means a smaller best achievable loss, but a more complex $\mathcal{V}$ makes the estimation of $\mathcal{L}_{\mathrm{DIB}}$ provably harder. This is a very good point, which we now discuss in the revised paper.
- "*[...] the decomposition of the input X [...] might need a revision*": Please refer to G.1.
- "*What is [...] odd/even CIFAR100?*": Classifying the parity of the class index. This "CIFAR2" is easier to visualize.