We thank the reviewers for their time and comments. Mutual information has emerged has an indispensable tool in representation learning and we share the reviewers' enthusiasm to explore further implications of our method, the first to provide an analytic expression for the compression rate for arbitrary input distributions. Some reviews express concern that we do not elucidate the breadth of potential applications for our approach, so we briefly discuss some of these ideas before addressing more detailed response points.

The supervised Information Bottleneck is a natural first extension, and we can indeed reference our concurrent work using Echo in the context of fairness and invariant representation learning. For example, we train using the IB functional to classify digits from an augmented MNIST dataset with rotation at angles $\{0, \pm 22.5, \pm 45\}$. We then train a classifier to predict rotation (a nuisance) from our representation, shown in Table 1. We find that Echo learns close to a task-minimal, invariant representation (since the classifier is simply random chance), and improves upon several baselines [2, 5]. We show similar results on the 3d Chairs dataset, with style as the label and yaw angle as the desired invariance.

Reviewer 1 references disentanglement as another desirable property in representation learning, although recent work has questioned current approaches and definitions of this term [4]. In the setting of [4], we calculate ELBOs and disentanglement scores ([3]) for Echo and several comparisons in Table 2. With no modifications of the objective and diagonal $S(x)$, Echo obtains competitive disentanglement scores with superior reconstruction. We plan to strengthen these results in the final version and extend our evaluation to the case when the ground truth factors of variation are dependent. In this scenario, Echo may have more flexibility to preserve disentanglement than independence-based methods [1, 3] given its lack of such assumptions.

While extending Echo to non-diagonal (e.g. triangular) $S(x)$ is an ongoing next step, we chose to first analyze the diagonal case where the architecture is directly comparable to a VAE (with the same structure and number of parameters). Further, the structure of $S(x)$ does not affect our ability to derive an analytic mutual information for arbitrary data, whereas common VAE assumptions only yield an exact rate in the unrealistic setting of Gaussian inputs. As we note in Sec. 2.1, Echo noise may still be dependent across dimensions with diagonal $S(x)$, and maintains tractability without prescribing a Gaussian form for the encoder or marginal.

With respect to the conditions of Lemma 2.3, bounding the activations $|f(x)| < M$ and $|s_j(x)| < r$ leads to a straightforward argument for convergence of the infinite sum defining Echo noise. If we truncate after d samples, we can bound the sum of remainder terms to be within machine precision by solving for r s.t. $M\, r^d/(1-r) < 2^{-23}$. We will clarify this reasoning in the main text, with detailed discussion of our implementation in the Appendix.

To address other minor comments, we found that Echo ran in just under 110% of the wall clock time for VAE on an NVIDIA Tesla V100 GPU, with a consistent ratio across datasets. While our distortion measure is referenced in line 163, we will emphasize this choice in Sec. 3.1. Our Appendix develops more detailed connections with classical rate-distortion and recent related works, which we hope will form a rich foundation for future developments.

Table 1: Invariant Classification Results
Label (higher is better) and Nuisance (lower is better)
Random chance : 0.20 for MNIST, 0.25 for Chairs

	MNIST-Rotated		Chairs	
Method	Label	Nuisance	Label	Nuisance
Echo IB	**0.98**	**0.20**	**0.84**	**0.25**
UAI [2]	0.98	0.34	0.74	0.34
VFAE [5]	0.95	0.38	0.72	0.37

Table 2: FactorVAE Disentanglement Scores: dSprites [3]

	ELBO	Disentanglement
Echo $\beta = 1$	**41.8**	.66
VAE $\beta = 1$	46.7	.61
Factor-VAE $\gamma = 10$ [3]	62.0	.72
Factor-VAE $\gamma = 50$ [3]	74.2	**.73**

References

[1] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[2] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. In *Advances in Neural Information Processing Systems*, pages 5092–5102, 2018.

[3] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[4] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[5] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.