We thank the reviewers for their substantive and constructive feedback, and appreciate their assessment of our explanation framework as "simple yet effective" **(R1)** with convincing experiments and clear presentation. We particularly appreciate **(R4)**'s comment that our "detailed and thoughtful approach presents a great example of high-quality reserch." The main reviewer concerns were narrative-related; we have addressed these concerns (and other specific suggestions) in revision.

1: Technical misconceptions in reviews.

- **(R1)** (Weakness 3) Section 4 illustrates and tests the performance of our method in a controlled setting; here we define the data distribution to be Gaussian, but outside of this section we do *not* make any assumptions on the data distribution. Also, our objective does *not* attempt to minimize $I(\beta; \widehat{Y})$ (but see Appendix A for discussion of potential objective variants).
- **(R1)** (Weakness 4) We do *not* require the training data originally used to train the classifier, which we agree would be problematic. Our explanations are based on a data distribution, but this need not be the training distribution.
- **(R1)** **(R2)** The validity of our DAG stems from the independence of (α, β), which we impose with an isotropic normal prior. (We have clarified this point in Sections 3.3, 4, and 5). Although MI is indeed a correlative metric in general, the independence of (α, β) allows us to show that in our framework it quantifies information flow, a well-founded node-based metric for causal influence (see refs 7, 55). This argument is indeed very simple (as pointed out by **(R1)** **(R2)**), but it depends crucially on our modeling decisions. Note that in the generative modeling literature the assumption of latent factor independence is common. One might alternately allow dependencies between the latent factors, jointly learning their causal structure with the generative network. However, without labeled side-information that would give these features semantic meaning, in our view our independence-by-construction approach is more useful to generate parsimonious explanations.

2: Changes to storyline. The most consistent reviewer concerns were narrative-related: (1) **(R1)** **(R2)** **(R4)** commented that the role of causality in our framework was overstated or unnecessary (since by enforcing the (α, β) to be independent our metric for causal influence simply reduces to MI), and (2) **(R1)** **(R2)** **(R3)** **(R4)** commented that our narrative did not engage with the disentangled representation literature. We thank the reviewers for these suggestions and have addressed them with careful changes to our storyline and a major addition to our related work section:

- As suggested by **(R1)** **(R4)**, we have adjusted our storyline to change the perspective from which we approach causality, instead highlighting the aspects of the model that allow the simple MI metric to be interpreted causally **(R2)**. We also clarified that the prior enforcing independence of (α, β) leads to the validity of our DAG **(R1)**. We agree with **(R4)** (and related work) that MI on its own can serve as a valid theoretical justification for explanation, and have reworked our discussions of MI-based methods to remove the impression of downplaying, emphasizing instead that our framework provides the *complementary* benefits of causal and information-theoretic interpretations.
- We believe that the disentanglement framing suggested by **(R1)** **(R2)** **(R3)** **(R4)** is exciting and could open new research directions. Our method can indeed be thought of as an information-based disentanglement procedure. Unlike techniques such as InfoGAN, our method (1) uses a classifier as side information; (2) separates classifier-relevant from classifier-irrelevant features, making the framework suitable for explanation; and (3) allows the MI metric to be interpreted as a measure of causal influence of disentangled features on the classifier output. We have adjusted the storyline throughout the paper to add this perspective, and have added a paragraph in the background to discuss the relation to specific disentanglement techniques (including all suggestions from **(R1)** **(R2)** **(R3)** **(R4)**) in more detail.

3: Additional experiments and discussion.

- **(R4)** We agree that computational cost is a concern. We have added a sentence in the main text mentioning this cost, and have expanded discussion in the broader impacts section of the potential drawbacks of using complex, uninterpretable models for explanation.
- **(R3)** **(R4)** The question of how mismatched explainer model capacity affects results is indeed both interesting and important. We address this in two ways. First, we are currently performing an experiment in which latent space dimension and VAE architecture complexity are swept for a fixed classifier complexity. We will add these results (presented both qualitatively with sample explanations, and quantitatively as in Figure 5(a-b) and Supplement Figure 11) along with a brief discussion in the main text. Second, we can use results from Feder & Merhav (*IEEE Trans. Info. Theory*, 1994) to bound the capacity mismatch of our explainer (i.e., explainer error in predicting classifier outputs) with the $I(\alpha; \widehat{Y})$ part of our objective. In practice, this result means that a sufficiently large value of $I(\alpha; \widehat{Y})$ serves as a certificate that the explainer complexity is sufficient to explain the classifier. We have added an appendix containing details and implications of this analysis and a brief discussion to the conclusion of the main paper.