We thank the reviewers for their thoughtful comments and are glad with the overall positive response.

Reviewer #1

1. Experiment results: linguistic explanation. We will add a linguistic explanation in Section 6 where we highlight that *ECCCo* produces plausible counterfactuals iff the classifier itself has learned plausible explanations for the data. It thus avoids the risk of generating plausible but potentially misleading explanations for models that are highly susceptible to implausible explanations.

2. Core innovation: more visualizations. Figure 1 shows the relationship between implausibility and the energy constraint for MNIST data. As expected, this relationship is positive and the size of the relationship depends positively on the model's generative property (the observed relationships are stronger for joint energy models). We will add such images for all datasets to the appendix. We note that our final benchmark results involve around 1.5 million counterfactuals per dataset (not including grid searches).

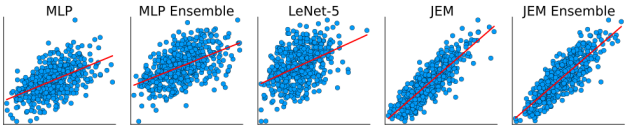

Figure 1: The L2 distance of randomly drawn MNIST images with Gaussian perturbations from unperturbed images in the target class (horizontal axis) plotted against their energy-constrained score, i.e. target logit (vertical axis).

3. Structural clarity. To facilitate comprehension, we will follow the reviewer's advice and add a systematic flowchart either in the appendix or in place of Figure 2.

Reviewer #2

4. Why use an embedding? There are two main reasons for using a low-dimensional latent embedding: firstly, to help with plausibility and, secondly, to reduce computational costs. The latter is not currently made explicit in the paper and we will add this in Section 5. The former is discussed in the context of the results for *ECCCo+* in Section 6.3, but we will highlight the following rationale.

There is indeed a tradeoff between plausibility and faithfulness through the introduction of bias: plausibility is improved because counterfactuals are insensitive to variation captured by higher-order principal components. Intuitively, the generated counterfactuals are therefore less noisy. We think that the bias introduced by PCA may be acceptable, precisely because it 'will not add any information on the input distribution' as the reviewer correctly points out. To maintain faithfulness, we want to avoid adding any information through surrogate models as much as possible.

5. What is 'epsilon' and 's'? From the paper: '[...] the step-size ϵ_j is typically polynomially decayed.' Intuitively, ϵ_j determines the size of gradient updates and random noise in each iteration of SGLD.

Regarding $s(\cdot)$, this was an oversight. In the appendix we explain that '[the calibration dataset] is then used to compute so-called nonconformity scores: $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ where $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function*.' We will add this in Section 4.2 of the paper.

6. Euclidean distance. As we mentioned in the additional author response, we investigated different distance metrics and found that the overall qualitative results were largely independent of the choice of metric. For image data, we still decided to report the results for a dissimilarity metric that is more appropriate in this context. All of our distance-based metrics are computed in the feature space. This is because we would indeed expect certain discrepancies between distances evaluated in the feature space and distances evaluated in the latent space of a VAE, for example. In cases where high dimensionality leads to prohibitive computational costs, we suggest working in a lower-dimensional subspace that is as uninformative as possible (such as PCA).

7. Model fails to learn plausible explanations. In these cases, *ECCCo* generally achieves lower plausibility while maintaining faithfulness (see also points 1 and 9).

8. Faithfulness metric: is it fair? We have taken measures to not unfairly bias our generator for the unfaithfulness metric: instead of penalizing the unfaithfulness metric directly, we penalize model energy in our preferred implementation. In contrast, *Wachter* penalizes the closeness criterion directly and hence does particularly well in this regard. In the absence of other established faithfulness metrics, we can only point out that *ECCCo* achieves strong performance for other commonly used metrics as well. For *validity*, which corresponds to *fidelity*, *ECCCo* performs strongly.

Joint energy models (JEM) are indeed explicitly trained to model $\mathcal{X}|y$, but the faithfulness metric is not computed for samples generated by JEMs. It is computed for counterfactuals generated by constraining model energy and hence there is no obvious source of bias. Our empirical findings support this argument: firstly, *ECCCo* achieves high faithfulness also for classifiers that have not been trained to model $\mathcal{X}|y$; secondly, our additional results in the appendix for *ECCCo-L1* show that if we do indeed explicitly penalize the unfaithfulness metric, we achieve even better results in this regard (also for models not trained to model $\mathcal{X}|y$).

9. Add unreliable models. We would argue that the simple multi-layer perceptrons (MLP) are unreliable, especially compared to ensembles, joint energy models and convolutional neural networks. Simple MLPs are generally more vulnerable to adversarial attacks, which makes them susceptible to implausible counterfactual explanations as we point out in Section 3. Our results support this notion, in that the quality of counterfactuals produced by *ECCCo* is higher for more reliable models. Consistent with the reviewer's idea, we originally considered introducing 'poisoned' VAEs to illustrate what we identify as the key vulnerability of *REVISE*: if the underlying VAE is misspecified, this will adversely affect counterfactual outcomes as well. We discarded this idea due to limited scope and because we decided that Section 3 sufficiently illustrates our line of thinking.