Experiment results: linguistic explanation of results Following the suggestion by the reviewer, we plan to add a the following linguistic explanation in a prominent place of Section 6:

"Overall, our findings demonstrate that *ECCCo* produces plausible counterfactuals if and only if the black-box model itself has learned plausible explanations for the data. Thus, *ECCCo* avoids the risk of generating plausible but potentially misleading explanations for models that are highly susceptible to implausible explanations. We, therefore, believe that *ECCCo* can help researchers and practitioners to generate explanations they can trust and discern unreliable from trustworthy models."

Elements of this explanation are already scattered across the paper, but we agree that it would be useful to highlight this notion in Section 6.

Core innovation: need more visualizations Following the reviewer's suggestion, we have plotted the distance of randomly generated MNIST images from images in the target class against their energy-constrained score. As expected, this relationship is positive: the higher the distance, the higher the corresponding generative loss. The size of this relationship appears to depend positively on the model's generative property: the observed relationships are stronger for joint energy models.

Structural clarity: add a flow chart Adding a systematic flowchart is a great idea. Due to the limited scope, may we suggest adding the following flowchart to the appendix? Alternatively, we may swap out Figure 2 for the flowchart.

Why use an embedding We agree that for any type of surrogate model, there is a risk of introducing bias. In exceptional cases, however, it may be necessary to accept some degree of bias in favour of plausibility. Our results for *ECCCo+* demonstrate this tradeoff as we discuss in Section 6.3. In the context of PCA, the introduced bias can be explained intuitively: by constraining the counterfactual search to the space spanned by the first n_z principal components, the search is sensitive only to the variation in the data explained by those components. In other words, we would expect counterfactuals to be less sensitive to small variations in features that do not typically vary much. It is therefore an intuitive finding, that *ECCCo+* tends to generate less noisy counterfactual images, for example (the same is true for *REVISE*). In our mind, restricting the search space to the first n_z components quite literally corresponds to denoising the search space and hence the resulting counterfactuals. We will highlight this rationale in Section 6.3.

We think that the bias introduced by PCA may be acceptable in some cases, 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 introducing additional information through surrogate models as much as possible. We will make this intuition clearer in Section 6.3.

Another argument in favour of using a lower-dimensional latent embedding is the reduction in computational costs, which can be prohibitive for high-dimensional input data. We will highlight this in Section 5.

What is "epsilon" and "s" From the paper: "$\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ_j is typically polynomially decayed. [...] To allow for faster sampling, we follow the common practice of choosing the step-size ϵ_j and the standard deviation of $\mathbf{r}_j$ separately." We go on to explain in the appendix that we use the following biased sampler

$$\hat{\mathbf{x}}_{j+1} \leftarrow \hat{\mathbf{x}}_j - \frac{\phi}{2}\mathcal{E}_\theta(\hat{\mathbf{x}}_j|\mathbf{y}^+) + \sigma\mathbf{r}_j, j = 1, ..., J$$

where "consistent with Grathwohl et al. (2020), we have specified $\phi = 2$ and $\sigma = 0.01$ as the default values for all of our experiments". Intuitively, ϵ_j determines the size of gradient updates and random noise in each iteration of SGLD.

Regarding $s(\cdot)$, this was an oversight, apologies. 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 main paper.

Euclidean distance As we mentioned in the additional author response, we investigated different distance metrics. We found that the overall qualitative results were largely independent of the exact metric. In the context of the high-dimensional 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 with respect to features, not latent features. This is because, as the reviewer correctly points out, we would expect certain discrepancies between distances evaluated in the feature space and distances evaluated in the latent space of the VAE, for example. Working in the feature space does come with higher computational costs, but the evaluation of counterfactuals was generally less costly than generating counterfactuals in the first place. In cases where high dimensionality leads to prohibitive computational costs, we would suggest either reducing the number of nearest neighbors or working in a lower-dimensional subspace that is independent of the underlying classifier itself (such as PCA).

Faithfulness metric: is it fair? We have taken measures to not unfairly bias our generator with respect to 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. That being said, *ECCCo* is of course designed to generate faithful explanations first and foremost and therefore has an advantage with respect to our faithfulness metric. In lieue of other established metrics to measure faithfulness, we can only point out that *ECCCo* achieves strong performance for other commonly used metrics as well. With respect to *validity*, for example, which as we have explained corresponds to *fidelity*, *ECCCo* typically outperforms *REVISE* and *Schut*.

Our joint energy models (JEM) are indeed explicitly trained to model $\mathcal{X}|y$ and the same quantity is used in our

proposed faithfulness metric. However, the faithfulness metric itself is not computed with respect to samples generated by our JEMs. It is computed with respect to counterfactuals generated by merely constraining model energy and we would therefore argue that it is not unfairly biased. 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.

Test with unreliable models We would argue that the simple multi-layer perceptrons (MLPs) are unreliable, especially compared to ensembles, joint energy models and convolutional neural networks for our image datasets. Simple neural networks have been shown to be 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 they demonstrate faithful model explanations only coincide with high plausibility if the model itself has been trained to be more reliable. Consistent with the idea proposed by the reviewer, we originally considered introducing "poisoned" VAEs as well, to illustrate what we identify as the key vulnerability of *REVISE*. If the underlying VAE is trained on poisoned data, this could be expected to adversely affect counterfactual outcomes as well. We ultimately discarded this idea due to limited scope and because we decided that Section 3 sufficiently illustrates our thinking.

References

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one.