R1: ***"I am not entirely convinced that an amortized explanation model is a reasonable thing. To investigate outliers (...) I presume computation of feature importance on the ground truth system would be more appropriate."***

R2: ***"I would appreciate some clarification about what is gained by learning $\hat{A}$ and not just reporting Ω directly."***

We thank R1, R2 and R3 for their insightful feedback. In settings where we have access to samples with associated ground truth labels, we could indeed directly use Ω, as defined by Eq. (6), to explain the a predictive model without training a separate explanation model. As correctly pointed out by R1, this would be preferable, for example for debugging at model development time, because Ω can be computed without any uncertainty, and computational performance is not a major concern at that stage. However, Ω can *only* be computed given ground truth labels. For many use cases of explanation methods, ground truth labels are not available, and an explanation model that generalises beyond the training data is therefore necessary. Imagine, for example, a predictive model that indicates whether or not an individual is at high risk for heart failure based on her individual attributes x_i. Suppose now that this system indicates an increased risk of heart failure for a specific person. For the physician that receives this prediction, it would be paramount to know whether or not this prediction is caused by the patient's blood pressure reading or by their genomic information, as this would dramatically change which further clinical steps should be taken. In this setting, every model decision is explained, no ground truth labels are available, and explanations and their certainty are consequential. This setting is not unusual, since we would not need to train a predictive model if we readily had access to accurate labels.

R1: ***"(The objective) does not attribute importance to features that change how the model goes wrong (...)"***

The perturbed feature itself would receive the same importance, but, since all attributions $\hat{a}_i$ are conditional on all input features x_i, the overall distribution $\hat{A}$ of importance scores would change along with the model's reasoning. In which case the user would be informed that the perturbation dramatically changed *how* the model arrived at its decision. This very approach was used in [19] to show that the explanations used by models are not robust to small perturbations.

R1: ***"Why is the rank-based method necessary?"***

We chose the rank-based method to show that the uncertainty estimates reflect accuracy according to the log-odds metric that is widely accepted by the research community as a benchmark for feature importance estimation (e.g. [1, 6]). We believe this is a higher standard, and therefore stronger evidence, since the rank-based metric shows that the uncertainty estimates are accurate not just by our metric, but by the community's standard.

R2: ***"Additionally, can the authors clarify what is being averaged in the definition of the causal objective?"***

The causal objective is averaged over all N samples in the dataset. Every data point has an Ω. We originally omitted the data point indices for brevity, but we will make the dependence of Ω and $\hat{A}$ on the sample explicit in the next revision.

R2: ***"If the goal is to determine what might happen to our predictions if we change a particular feature slightly keeping all others fixed, I don't see any role for the explanation model – one can simply compute the new prediction.***

Our goal is not to estimate what would happen if a particular feature's value changed, but to provide a causal explanation for the prediction made by the model, i.e. which input features x_i causally influenced the prediction and to what degree.

R2: ***"Some additional clarity on why the authors are using a KL discrepancy is merited. Why not use, say, the euclidean distance between the vector Omega and the importance weights derived from the explanation model?***

The KL divergence has connections to Bayesian surprise and human attention (see Itti and Baldi, NIPS 2006), and is therefore a particularly suitable candidate for optimising the distribution $\hat{A}$ of importance attributions.

R3: ***"Masking one by one; this is essentially equivalent to assuming that feature contributions are additive."***

We do not define a feature's importance as its additive contribution to the model output, but as it's marginal reduction in prediction error. This subtle change in definition allows us to efficiently compute feature importance one by one. Non-linear interactions between model inputs and model outputs are possible within this definition, since the additivity constraint pertains to the marginal reduction in prediction error only (which holds in the general setting).

R3: ***"Replacing a masked value by a point-wise estimation can be very bad, especially when the classifiers output changes based on the masked feature. Why would the average value (or, even worse, zero) be meaningful?"***

R3 is absolutely correct. There is a range of imputation strategies that could be employed to mask the individual features x_i, and our work focused on the most straight-forward strategies. We will clarify this point in the next revision.

R3: ***"It would also be interesting to compare the proposed method with causal inference technique for SEMs."***

Recent work [29] has explored the use of SEMs for model attribution in deep learning. Compared to CXPlain, the main disadvantages of their approach were that (i) their method was limited to specific neural network architectures whereas CXPlain can explain any machine-learning model, and (ii) attribution time was considerably slower than CXPlain.

R3: ***"It seems to me that the chosen performance measure may correlate much more with the Granger-causal loss than with the objectives optimized by the other explainers."***

Related works, such as LIME, SHAP and gradient-based methods, compute attributions directly based on the change in the explained model's output as also measured by the log-odds metric. In contrast, the causal loss uses the marginal reduction in prediction error and, therefore, only indirectly models the change in model output.