We thank the reviewers for their time and thoughtful comments.

re: L1 regularization. (R1 and R2) Our implementation did not use L1 reg. We added this note to describe a variation of ROAR on models with explicit feature selection, e.g. L1 regularization in a linear model. Here we know which features are not used, hence we can prevent them from being used when retraining. This makes ROAR more reliable.

Reviewer 1 (R1) re: portrayal of human studies: R1 correctly points out our portrayal of human studies requires more nuance. We would be glad to correct this and will update the manuscript accordingly. **re: variance of the results:** The variance is very low across all datasets and estimators. The maximum variance observed for ImageNet was a variance of 1.32% using SG-SQ-GRAD at 70% of inputs removed. On Birdsnap the highest variance was 0.12% using VAR-GRAD at 90% removed. For food101 it was 1.52% using SG-SQ-GRAD at 70% removed. As the reviewer assumed correctly, the gap between estimators is far larger than the variance. **re: single ROAR metric using AUC:** This is something we will consider when benchmarking additional methods in the future. But as the reviewer points out, sometimes the curve itself provides additional information. **re: ROAR dataset generation:** The saliency maps for the datasets were pre-computed and stored on disk. These were then combined in the pre-processing pipeline to mask out the required part of the image.

Reviewer 2 (R2) re: What ROAR measures: At line 101, we discuss the nuances of using a set of ROAR models to evaluate the accuracy of an explanation produced on a different original model. We will update the manuscript to discus this earlier, so it is clearer to the reader what ROAR measures.
re: Decoy MNIST: In the Decoy MNIST training set, the corners of the image are modified to be predictive of the label. On the test set, these modifications are random. If the model uses these corners the test set accuracy will be low. Assuming the patches are the most important and the interpretability method detects this, the following happens using ROAR. As the corners are masked, test set performance increases. After this the accuracy would start degrading. A random baseline would be expected to exhibit a steady decrease. We believe that with a comparison to the random baseline the odd behavior can be detected (and explained).

Reviewer 4 (R4) R4 argues that "*The main drawback of this work is that it presents negative results*", however this manuscript results in a strong **positive recommendation** for the use of SmoothGrad-Squared and VarGrad. We find that these ensemble estimators *far* outperform all other methods; for example on ImageNet, there is a remarkable gap of 56.34% between the best performing ensemble method (VarGrad) and the best performing base methods (GRAD). The result is consistent across all 3 large scale datasets we evaluate.

The **significance** of the work is also supported by the other reviewers. **R1:** "*This evaluation could become popular, inspire future metrics, and inspire better importance estimators.*". **R2:** "*Surprisingly, most methods underperform random feature ablation, and also surprisingly smoothgrad-squared and a similar method far outperform the other methods. This finding raises interesting questions about both the failure of many traditional methods as well as how smoothgrad-squared works so well. This is a very interesting result that will lead to follow-on work, and it is a second significant contribution.*".

re: "*whether or not curves [...] with and without retraining generally match*". We note that this experiment is performed on toy data in **Fig. 2** and on Imagenet in **Fig. 3**, and the results strongly support our stated methodology of re-training. Without retraining, the 'removal' of pixels by replacing with a constant introduces new image statistics that were not seen by the model during training.

R4 suggest we evaluate on **more datasets and methods**. We note that we already present consistent results on both toy data and several large scale, natural image datasets such as ImageNet, BirdSnap and Food 101. In addition to two baselines (sobel edge detector and random), we compare 12 methods in the main paper (and 4 in suppl.). In total, we generate 540 large-scale modified image datasets in order to consider all experiment variants (180 new test/train for each original dataset). For each of these datasets, we independently train 5 ResNet-50 models from random initialization.

re: Fong and Veldadi, who propose a minimum perturbation area which preserves the model prediction. This minimum deletion area is identified by perturbing and evaluating the model output without retraining. The authors openly acknowledge the shortcomings of not re-training is the introduction of artefacts that could change the image to be out of distribution. This is *precisely* the reason we do insist upon retraining on the modified inputs, such that the model can learn that the constant we use to replace the pixels removed are uninformative.

re: considering dropping patches vs. individual pixels. We note that Liu et al. had a different stated motivation of image in-painting. However, we agree that evaluating rankings according to different units of importance would be valuable. Due to computational constraints, we leave this as the subject for future work.