We thank the reviewers for encouraging and insightful comments. We clarify the major points below:

Reviewer 3, Comment 3.1. Evaluating the confidence metric over more confusable out-of-distribution examples. A model trained on MNIST dataset was run on a more *confusing* dataset, `MNIST-rot-back-image`[1]. The accuracy and confidence of the model drops for the confusing dataset (see Figure 1 (Top Left)). Results in Fig. 1 (Top Right) show that the attribution-based confidence drops with increase in rotation angle (from 0 to 50 degrees) and decrease in accuracy.

Reviewer 3, Comment 3.2. Interpretability and qualitative analysis of the confidence metric. Figure 1 (Bottom) illustrates how confidence computed by attribution on examples from `MNIST-rot-back-image` reflects the perceived ambiguity and confusability of inputs.

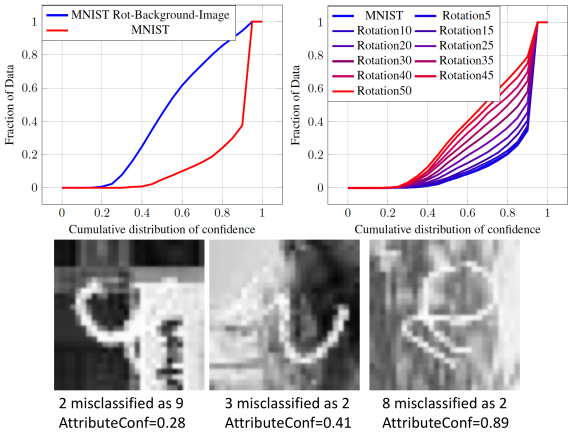

Figure 1: Confidence computed on new dataset `MNIST-rot-back-image` (Top Left) and rotated MNIST (Top Right) at different angles. Selected examples from `MNIST-rot-back-image` (Bottom).

Reviewer 4, Comment 4.1. Demo against Platt scaling/Calibrated predictor baseline. The comparison of attribution-based confidence metric with calibrated Platt scaling model is shown in Figure 2 (Left).

Reviewer 4, Comment 4.2. The sparseness of IG attribution maps. We present the distribution of attributions for ImageNet in Figure 2 (Right). As anticipated, attributions concentrate over a small number of high-attribution features.

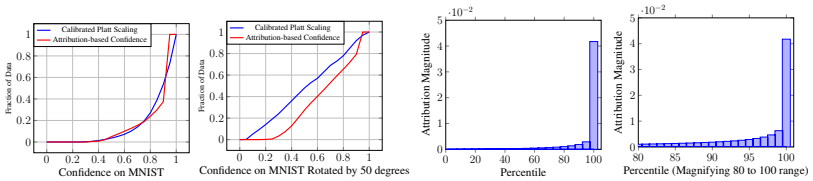

Figure 2: (Left 2) Comparison with calibrated Platt (temperature) scaling model. (Right 2) Concentration of Attributions over few features for ImageNet.

Reviewer 4, Comment 4.3. Less focus on adversarial examples. We will supplement our earlier reported results on out-of-distribution examples with additional new results (such as those presented here) to balance focus away from adversarial examples.

Reviewer 5, Comment 5.1. Theorem 1: How the first line of equation was arrived?; Reconcile with IG assigning high importance to saturating inputs. The first line of equation is derived by using the product rule for differentiation. that is, $\frac{d(fg)}{dx} = f.\frac{dg}{dx} + \frac{df}{dx}.g$. The IG attribution is $\mathcal{A}_j^i(\mathbf{x}) = (\mathbf{x}_j - \mathbf{x}_j^b) \times \int_{\alpha=0}^{1} \partial_j \mathcal{F}^i(\mathbf{x}^b + \alpha(\mathbf{x} - \mathbf{x}^b)) d\alpha$. By differentiating w.r.t x_j using product rule, we get $\int_{\alpha=0}^{1} \frac{\partial \mathcal{F}^i(\mathbf{x}^b + \alpha(\mathbf{x} - \mathbf{x}^b))}{\partial \mathbf{x}_j} d\alpha + (\mathbf{x}_j - \mathbf{x}_b) \frac{\partial}{\partial \mathbf{x}_j} \left(\int_{\alpha=0}^{1} \frac{\partial \mathcal{F}^i(\mathbf{x}^b + \alpha(\mathbf{x} - \mathbf{x}^b))}{\partial \mathbf{x}_j} d\alpha \right)$ where $\mathbf{x}_b$ is set to 0. The IG attribution is non-zero even when the partial derivative is zero (enabling better measurement for saturating features) but IG attribution also saturates, albeit to a non-zero value. Consequently, we avoid the effect of saturation while using the change in attribution for importance sampling by approximating the rate of change of attribution in Eqn. 7 as a linear variation over the change in features. We will include this discussion and associated intuition with an illustrative example.

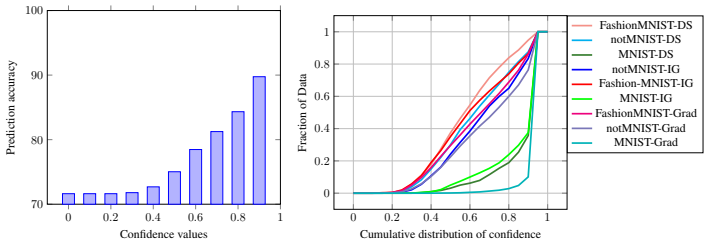

Figure 3: (Left) Accuracy improvement after abstention on `MNIST-rot-back-image`. (Right) Comparing different implementations of attribution-based confidence using Gradients (Grad), Integrated Gradient (IG), and DeepSHAP (DS). For out of distribution examples (FashionMNIST and notMNIST), results with DeepShap are slightly better than IG (which is better than Gradient).

Reviewer 5, Comment 5.2. Platt scaling comparison. Please see Figure 2 (Left).

Reviewer 5, Comment 5.3. Improvement after abstention. Please see Figure 3 (Left).

Reviewer 5, Comment 5.4-5.5. Using DeepShap and Gradients. Please see Figure 3 (Right) for results on MNIST, notMNIST, and FashionMNIST. More detailed evaluation with DeepShap/Gradients will be included in the full paper.

[1]Public dataset from U.Montreal (link omitted because response must not have external links). Dataset has MNIST images randomly rotated by 0 to 2π, and with a randomly selected black/white background image. See examples in Fig. 1 (Bottom).