We thank the reviewers for encouraging and insightful comments. Below we respond to reviewers' major comments.

Question 1 (R1, R4, R5) *Why study robust attribution regularization? How does it correlate with human perception? (e.g., how about robust but totally flawed attributions?)*

We appreciate this question about the essential motivation of this work. To begin with, we believe that robust attribution is at least *necessary* for a machine learning model to be trustworthy. Model attributions are *facts* about a model's behaviors. It is true, as **R1** pointed out, that users may still want to be skeptical when interpreting attributions, even if they are robust. However in the opposite direction, one can never trust a model with *brittle* attributions.

Regarding human perception, a very intriguing recent paper [3] from Madry's group showed that, empirically, adversarial training (or robust prediction training) produces models whose attributions are *much more aligned with human perception*, and seems to learn salient features from data. Since robust prediction training is a special case of robust attribution training, their results form a basis for our belief that robust attribution regularization can lead to better correlation with human perception (which is however a somewhat subjective question). Indeed, our visualization results in the draft and appendices also corroborate their findings.

Note that, in the worst case, it is easy to construct models that have robust but totally flawed attributions – simply take models that always have "the same" behaviors. However, the situation becomes much more complex if one imposes the constraint that the model should also achieve small training error, which is what we did in this paper. Therefore, robust attributions can be thought of as imposing an inductive bias to encourage learning *invariant* features from data. Since human perception is essentially also related to recognizing invariances, small training error with invariant features intuitively implies alignment with human perception. Next version will incorporate the above points.

Question 2 (R4, R5) *How useful is evaluating IN and CC metrics?*

The use of IN and CC is aligned with previous literature in studying robustness of attributions (in particular, the work of Ghorbani et al. [1]). We agree with reviewers that these two metrics are only one form of brittleness, and are direct consequences of our objectives. We also agree with reviewers that it is hard to quantify how increasing these two metrics improves (directly) human perception. On the other hand, we think that these evaluations are still useful: (1) It answers the scientific questions raised in [1], (2) Perhaps more importantly, it corroborates our analysis that robust prediction training will robustify attributions as well. We will make these points explicit in the next version.

Question 3 (R1) *Section 3.2 seems redundant.*

We apologize for the confusion. In fact Section 3.2 fulfills an important theoretical purpose: Distributional robustness approach forms a different school towards robust prediction training (see [2]). The analysis here shows that generalizing these objectives to robust attributions (a much larger class) essentially still gives very similar objectives in two different robust optimization models, and thus we can "safely" stick to the formulation in Section 3.1, which is reassuring.

Question 4 (R1) *Is there any functional value of regularizing an intermediate layer?*

Proposition 3 proves that if one regularizes by the output layer it gives a natural surrogate loss of Madry et al.'s objective function, which to us makes even more sense as it directly bounds the *absolute difference* between x and x'. We believe that there is more to regularizing intermediate layers and we are actively researching it.

Question 5 (R5) *More datasets, MNIST is not great.*

Besides MNIST we have evaluated Flower and GTSRB (traffic signs). Both are more diverse than MNIST. GTSRB is practically motivated, and Flower is a high-resolution vision dataset well suited for studying attributions. We get similar results in terms of both metrics and visualizations. We are actively working on more datasets and plan to include more in the final version if this paper gets in.

Question 6 (R5) *More details on optimization difficulty and architectural properties.*

We will add more details. Roughly speaking, a main problem is *network depth*, where as depth increases we get very *unstable* trajectories of gradient descent, which seems to be related to the use of second order information during robust attribution optimization (due to summation approximation, we have first order terms in the training objectives).

References

[1] A. Ghorbani, A. Abid, and J. Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 3681–3688, 2019.

[2] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[3] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.