Thank you for your thoughtful feedback. We will first discuss common themes and then specific reviewer comments.

Significance: Even though ExpO is "simple" (in that it connects existing concepts, albeit in a novel way), we believe that it is highly impactful because there is no other model-agnostic and domain-knowledge free method for improving the quality of local approximation explanations such as LIME (which is a seminal method in Interpretable ML).

Prior Work: The suggested related works (which we will cite in the revision) all solve different problems than the one we consider. We will add a discussion as outlined below.

- "Adversarial Robustness ..." by Qin et al does not consider interpretability at all. When adapted to consider interpretability, it uses a gradient based explanation and its regularizer is quite similar to SENN's. Consequently, it will have the same issues with flexibility, fidelity, and stability as gradient based explanations. See A.2 for details.
- Several methods rely on domain knowledge: "Learning credible ..." by Du et al, "Learning Deep ..." by Weinberger et al, "Interpretations are ..." by Rieger et al, and "Regional Tree ..." by Wu et al [2].
- "Beyond sparsity ..." by Wu et al [1] regularizes for global interpretability while ExpO regularizes for local interpretability. Despite the fact that they are globally interpretable, small decision trees are difficult to explain locally with explainers like LIME (see Figure 1 for an example). As a result, [1,2] do not solve the same problem as ExpO because making the model look more like a decision tree makes LIME less effective.

Reviewer 1. *Reproducibility*. The reviewer is correct that we are comparing MLPs trained using standard techniques to ones trained with ExpO. We will add a detailed discussion of the neural networks (structure, activations, widths, depths, etc), hyper-parameters (learning rate, optimizer, regularization), and selection procedures to the appendix so that the reader does not have to reference the code (which reproduces all of our results) to reproduce our results.

Reviewer 2.

"computational complexity ... cube ... not usable for higher-dimensional inputs." We introduce ExpO-1D-Fidelity to address this concern (line 160-167). Its complexity is independent of the data dimension and we show it scales well to datasets with ~ 100 features. We also note that related methods require expensive operations (FTSD and [1] both are non-differentiable; SENN and RRR both require differentiating through the model gradient).

"compare to RRR in this manner." To the best of our knowledge, it is not technically possible to encode fidelity/stability using RRR's regularizer.

Reviewer 3.

"difference between RRR, SENN is that a neighborhood ... is introduced ... not a huge difference." ExpO is the only method that is differentiable and model agnostic that does not require domain knowledge; the differences are not just in whether or not a neighborhood is used. See Table 1 for details.

"Algorithm 1 ... not very novel." Viewing the novelty of ExpO merely through the lens of Algorithm 1 sells it short; the novelty stems from its impactful connection to interpretability. It is common for algorithms designed in one area to be impactful when introduced to another area (eg, SENN/RRR are "just" regularizing the gradient which is a strategy at least as old as "Tangent prop-a formalism for specifying selected invariances in an adaptive network." NeurIPS92.)

*"results ... not very surprising ... idea is to *optimize* those metrics during learning."* Two small clarifications: the results are shown for points that were not regularized for during training and the results shown in the main paper were regularized only for fidelity, so the improvement in stability is not a given.

"why is the accuracy of SENN explanations...measured using a post-hoc explainers." While the reviewer is correct that SENN's Point-Fidelity (PF) is perfect by-design, its Neighborhood-Fidelity (NF) is not guaranteed; the setup of user study clearly motivates why NF can be preferable to PF (lines 220 - 223). Following the reviewer's suggestion, we computed NF and Stability for SENN explaining itself. While the results are better than using LIME, they corroborate the general message that ExpO is a more flexible solution than SENN for trading off between accuracy and interpretability. Specifically, SENN explaining itself has a NF of 3.1e-5 and a Stability of 2.1e-3; these numbers are generally comparable to LIME explaining the appropriate ExpO model. See A.1 and Table 5 for details.

"regularize neural nets ... behave similarly to decision trees, either globally or regionally ... expect tree-regularized models to work well together with LIME... for both linear explanations and tree-based explanations." As noted in the above discussion on [1], neither of these methods would improve LIME's explanation quality for linear explanations. Although we agree that exploring non-linear local explanations is an interesting direction, ExpO focuses on the setting where the explanation is linear because this is what LIME, MAPLE, and SENN all do.

Reviewer 4: *"experimental part is somewhat not convincing ... it is not surprising to see the results in user study: the regularized model achieves better interpretability than the normal model."* Fidelity/stability are the standard proxy metrics used to evaluate local approximations. However, as we emphasize in the paper (line 38-42), they are only proxies for some underlying notion of interpretability, and the goal of the user study is to directly study explanation usefulness. Consequently, it inconsistent to criticize the results on the metrics and then use those same results to criticize the results of the user study.