We want to thank the reviewers for their thorough comments and suggestions for improving the manuscript. We have inlined responses to the major points below, and will address all minor points in our next revision as well.

(R1) Assumptions While we cannot establish formal guarantees when constructing our "robust dataset" in Section 3.1 (which we presume this comment is referring to), our method follows a fairly well-motivated approach—-for each input in the original training set, we choose as a seed an image that is randomly selected (independent of label—to avoid introducing any feature-label correlation), and then modify this image to make it match the representation of the original input under our robust model. The resulting dataset thus matches the original in terms of the features used by the robust model, while preventing the re-introduction of features that robust model is invariant to (which are non-robust features).

Finally, it is important to note that, in the end, our result is of *existential* nature: i.e., for the first time, we managed to construct a dataset that results in models that can tackle a non-trivial task and are robust after just *standard* (ERM) training. This suggests that our overarching conceptual framework might be indeed predictive of the way the underlying phenomenon behaves.

(R1) By selecting ... features? Note that our definition of a feature defines it via its *generalization* performance. This makes it impossible to "overfit" to a feature in the traditional sense.

(R1) In your proposed method ... not robust? This is correct—fortunately, they all resemble the target images.

(R1) Why distance in robust feature space? Our goal is to create a training set which does not contain the features that a robust model is invariant to (i.e., non-robust features). Optimizing distance in robust feature space is a clean way to induce this invariance while still matching the features that *are* important for robust classification.

(R2) Clarity We want to thank the reviewer for their suggestions regarding clarity of our presentation. In addition to adding examples/exposition around our methods, we will also: make sure to move the related work into the main body (in order to better position our work), and include algorithms for generating the four datasets constructed in the main body of the paper.

(R2) definition of "feature" Our formal definition of robust (and non-robust) features in Section 2 is designed to be a high-level guiding framework for the design and analysis of our experiments. As such, there are some nuances/complicated scenarios not captured by our simple definitions (as the reviewer points out), but we viewed it as fully sufficient to describe and predict the results of our experiments. Nonetheless, we view coming up with a more nuanced/fine-grained definition of features as an important direction for future work. As far as our manuscript goes, we will update it to reflect that view (and highlight the corresponding line of future work).

(R2) Section 4 The goal of S4, as the reviewer points out, is to provide an illustrative example of how misalignment between the "feature metric" in the data and the "adversary metric" (Euclidean distance) can lead to adversarial vulnerability—-and how robust training can "fix" this misalignment. To this end, we settled on the simplest possible setting (e.g. convexity, to ensure a closed-form solution exists even for the robust problem), so that robustness and robust optimization could be studied as rigorously as possible. (It turns out that even in this simple setting analysis is not completely straightforward.) We very much agree with the reviewer that similar analyses for more complicated settings and classifiers would be an important direction for future work.

Nonetheless, our preliminary empirical and theoretical work indicates that the results do extend beyond the simple setting presented here (for example, one can show that for linear models, non-robustness arises from misalignment not only in the case where the data is Gaussian but for any distribution with bounded second moment). While we are happy to include these extensions in the next revision, any more substantial extensions (such as moving beyond linear models, analyzing robust training for different distributions) might warrant separate work.

(R3) Looking ... original dataset? While we didn't notice a significant decrease in diversity (the four random samples do look somewhat "prototypical" but this seems to be mostly by chance—we can include a larger selection of random samples in our final version). It's possible that there is a slight decrease in diversity (maybe that's also why *standard* accuracy on the $\mathcal{D}_R$ dataset is very slightly worse than that of the original robust network). It would be interesting to see if different methods of constructing $\mathcal{D}_R$ (for example, starting from a bunch of different random images per training set image, etc.) would be effective at introducing more diversity into the training samples.

(R3) I'm intrigued ... generation process? While we did not perform a formal study on this, we noticed that the accuracy increases from 0 (when $\epsilon = 0$, clearly, since at this point it is just training with mislabeled data) and then just plateaus at a reasonable ϵ, a bit higher than the one we used for generating the dataset. (Note that we didn't tune the ϵ parameter at all to obtain these results, and as a sanity check our results are stable over a reasonable range of ϵ values.)