We thank the reviewers for their insightful comments on our work, and appreciate their help in making our manuscript better. Below, we address the major concerns that they have brought up.

Tone/Framing of the Paper: We had argued that there was a flaw in the SATNet authors' visual Sudoku experimental design, primarily because they claimed to have integrated logical reasoning into a deep neural network to enable "end-to-end" learning in a "minimally supervised fashion." We understand "end-to-end" learning to mean that there are no intermediate labels that need to be collected. For example, end-to-end speech recognition means that there is no need to collect intermediate phoneme labels to build an acoustic model, and end-to-end facial recognition means there is no need to collect intermediate labels for a nose/eyes/lips detector. We do not argue that SATNet has claimed to solve the symbol grounding problem in general, only that a necessary implication of their claim to end-to-end learning for visual Sudoku is that they must have solved it for this specific instance. We are concerned that many papers about logical reasoning in deep learning omit any discussion about symbol grounding, and we hope that our paper reminds the community of its importance. Ultimately, if neural networks are not able to ground perceptual phenomena into symbols, there is little need to integrate any logical reasoning into deep learning, since we can always build a system that has distinct machine learning and discrete solving modules, like the OpenAI Rubik's Cube solver.

That said, we appreciate Reviewers 1 and 2 for suggesting that there is an alternative interpretation of "end-to-end" learning that means being able to train a deep network with gradient descent by making every component in the network differentiable, which might well have been the SATNet authors' intended claim. This is something that we had not previously considered, which likely made our argument overly critical towards SATNet. To allay the reviewers' concerns, we will amend our title to Reviewer 1's proposal: "Assessing SATNet's Ability to Solve the Symbol Grounding Problem". We will further remove any reference to SATNet's use of intermediate labels as a "flaw" or a "bug", and provide a discussion about the two different interpretations of end-to-end learning. It is quite possible that many readers of the original SATNet paper might have understood it with the same interpretation as Reviewer 4 and us, and thus, we think that our work will clarify to the community the exact nature of SATNet's capabilities. We empathize with the reviewers' concerns that our work might come across as "sensational" or lacking "objectivity", and welcome further suggestions that can improve the quality of our manuscript.

Significance of our Contribution: Reviewer 3 was concerned about a lack of novelty in our work besides our critical review of SATNet's visual Sudoku experiment. One of the main contributions of our work is to highlight the difference between defining gradients for an architecture (the end-to-end learning interpretation of Reviewers 1 and 2) and actually being able to train it. This problem is similar to the one that the deep learning community had to solve for neural networks to enable the training of architectures with hundreds of layers, where gradients are well-defined but successful training is non-trivial and requires careful initialization, batch normalization, adaptive learning rates, etc. Our work identifies several factors that affect the learning dynamics of SATNet and provides practical suggestions for configuring SATNet to enable successful training. We reveal surprising complexities that are unique to SATNet and break standard deep learning norms. For example, using different learning rates for different layers in neural networks is not a common practice, since the use of Adam usually suffices. But for the case of SATNet, even when Adam is used, the backbone layer has to learn at a slower rate than the SATNet layer for successful training to occur. In addition, our findings go beyond verifying simple intuitions about SATNet. We believe that broad statements like setting m to "the total number of variables," as Reviewer 2 suggests, do not suffice to elucidate the complicated relationship between m and aux, and their effect on the intricate training dynamics in SATNet. For example, we see in Figure 4 that even when $m = 20$ is the total number of variables, performance can degrade for certain values of aux, but not when m is increased. Instead of relying on intuition, our work systematically studies the empirical effects of different SATNet configurations and highlights key steps required for successful SATNet training to occur.

Other Comments: We thank Reviewer 4 for pointing out relevant prior work from sub-symbolic reasoning, and will include them in our manuscript. Reviewer 2 suggests that SATNet's failure on some random seeds was due to the evaluation metric for visual Sudoku, but we showed that the failure also occurred on the much simpler MNIST mapping problem, where the metric is the percentage of test images classified accurately. It is clear that these cases should be identified as learning failures (resulting from the sensitivity of SATNet to the random initialization) and not dismissed as mere artifacts of the evaluation metric. As Reviewer 2 points out, m is strictly speaking the number of clauses that the SATNet layer *represents*; we stick to the same terminology used in the original SATNet paper and believe that most readers will not be confused by this.

Concluding Remarks: In general, we think that the differences between deep learning and logic mirror the ones between continuous and discrete optimization. These differences go far deeper than the superficial lack of derivatives in discrete optimization, and we believe true progress has to come from significantly tighter integrations between deep learning and logic. We are excited that our work brings these differences to the forefront and encourages the community to think more critically about how to go about integrating logical reasoning into deep learning.