We thank all four reviewers for their thoughtful and valuable feedback. We agree with all suggested edits and have incorporated the feedback to improve the content and clarity of our paper. Key changes to the paper include:

1. Switched the code repository from private to public.
2. Added a Supplementary Information section, which includes:

   - Implementation details for training and analyzing the RNN. We provided details many reviewers had asked for, including the exact loss function, when the RNN state was reset (only at the beginning of a gradient step), how gradients were backpropagated (over all RNN steps i.e. across multiple trials and multiple blocks), how the feedback signal was computed (at the end of each trial).

   - Distillation implementation details for our Representation and Dynamics Distillation (RADD) technique

   - Results for networks of different sizes (25, 100, 150, 250). Although we didn't perform statistical analyses comparing different networks, the results we discuss in the main body all hold.
3. Added a Related Work section to the main paper, citing additional related work including the references noted by the reviewers, and clarifying the relationship of that body of work to ours.
4. Added a comparison of RADD to traditional distillation, including a comparison of training efficiencies (wall-clock time) for model compression. We also compared the original network's timescales of integration/decay and eigenvalues of the recurrent Jacobian to those of the distilled network.
5. Clarified, added, removed or otherwise edited parts which were specifically mentioned in the detailed review comments.
6. Changed figure color schemes to be accessible to color-deficient readers.

A concern voiced by one reviewer was that, though of good quality, this work might not be a good fit for NeurIPS. However, quoting from NeurIPS's website, “The purpose of the Neural Information Processing Systems annual meeting is to foster the exchange of research on neural information processing systems in their biological, technological, mathematical, and theoretical aspects," which we feel directly concerns this work. Additionally, this paper is similar to recent NeurIPS papers (e.g. [1, 2]), and three of our four reviewers (who we hope are a reasonable proxy for our target audience) seemed to find the work interesting for NeurIPS.

References

[1] Ingmar Kanitscheider and Ila Fiete. “Training recurrent networks to generate hypotheses about how the brain solves hard navigation problems”. In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 4529–4538.

[2] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. “Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics”. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 15696–15705.