We thank the reviewers for insightful remarks and comments that help to considerably improve our manuscript. We address the most important ones in detail below. Before doing so, we highlight a comment from R3 in order to make an important clarification about the scope of our contribution. "*It is well known that an attention mechanism would reduce gradient vanishing. It feels trivial to me as there is a direct connection for gradients to pass. [...] it's not hard to see that having fewer things in the softmax (i.e. sparse attention) would increase the attention weights, thus improved gradient flow.* We are in complete agreement and recognize that the very mechanism of (self-)attention was designed to improve gradient propagation over long sequences, and that sparsity is a good way to keep complexity costs low. However, to the best of our knowledge, there is currently limited understanding of gradient scaling properties in the presence of attention. Much like work from the '90s established formal results for gradient exploding/vanishing in deep/recurrent networks, we believe it is crucial to establish similar theoretical tools for attention mechanisms, as these methods are under intense development where scalability and complexity are important issues. Our main aim is to contribute to this direction with a thorough analysis of gradient propagation in self-attentive systems, outlined in clear theorem statements which precisely quantify the intuition expressed by R3. These results are not trivial (see proofs in appendices), and offer valuable guarantees for attention mechanism development. The proposed relevancy mechanism and accompanying experiments, building on established work, are meant to illustrate how our theorems can be concretely exploited. We chose simple tasks for their ease of interpretation, and their variety of computational demands (memorization, prediction, RL, etc.). As is clearly indicated in the text, it is not our goal to propose this method "as is" in a race for state-of-the-art. Rather, we think it achieves its goal of providing a firm theoretical footing for a wide class of self-attention methods as well as suggesting future direction of methods based on concrete scaling guarantees and a balance between sparsity and attention complexity. We recognize that reviewers may have based their evaluation as they would have in a method paper, and we kindly invite them to reconsider the value of our experiments in the broader context of our theoretical contributions. We also thank reviewers for their additional minor comments not explicitly addressed here and agree to implement them.

R1: **Q**"*The authors didn't spell out the relation between κ and d: higher κ tends to have smaller d. This relation is discussed in the paragraph of line-173 but not reflected in the formula of theorem-2.*" **A:** We thank R1 for this important remark. As it stands, Theorem 2 offers scaling relationship for any κ and d. However in practice, κ and d co-vary in ways that depend on the task's underlying relevancy structure, a point that is explained in detail in Appendix C (see Fig 3) which explores trade-offs between ρ $(= \kappa - \nu)$ and gradient propagation (implicitly depending on d). We will sharpen this discussion in the main text. **Q**"*In experiments, the authors mentioned the proposed model is "faster" to train but didn't give any quantitative results.*." **A:** We will clarify this in the text and specifically point to Fig 4 in the Appendix D (or move it to the main text space permitting) which shows RelLSTM/RelRNN learns the copy and denoise tasks with significantly fewer number of updates as compared to other baselines.

R2: **Q**"*Line 145, how can Theorem 1 be related to the early attention mechanism [1]? As the attention weights are computed adaptively, it is unlikely that they are uniform.*" **A:** We recognize that uniform attention weights are unlikely in practice. This theorem offers a form of "worst case" guarantees which is applicable in practice for two reasons. (1) Typically, attention weights are initialized uniformally. (2) We experimentally verified that gradient propagation remains stable throughout training for a fully self-attentive RNN, see Fig 2 (Section 6). We will further clarify this in the text. **Q**"*What is the advantage of the proposed method over MANNs? / how are MANNs related to the Theorem 2? How are the paper's theoretical findings different from [5], wherein gradient norm of self-attentive RNN is also quantified?* **A:** We thank R2 for this insightful remark and acknowledge the lack of discussion about the MANN model class. Most instances operate on involved memory addressing/retrieval mechanisms that keep complexity costs low but, to our knowledge, do not offer gradient propagation guarantees like Theorem 2. Some instances, such as that proposed in [5], offer more similarities to our approach, and future adaptations of our Theorem 2 is an exciting future direction. Our contribution complements that of [5] in two subtle but important ways. (1) The proof of Theorem 2 in [5] describes an approximation of information propagation via recurrence, with "skip connections" contributions accounted only once in isolation. In self-attention, gradient propagates via a mix of skip and recurrent connections, leading to multiple gradient paths. Enumerating these paths and incorporating their contributions to gradient norms is the crux of our results, yielding complete and precise expansions of gradient terms applicable to a wide range of models, including Transformers. (2) The proposed method in [5] is an optimal memory writing schedule. In contrast, our method relies on relevancy for memory writing, allowing more flexibility in complexity scaling. We will incorporate this discussion in the revised text.

R3: **Q**"*augmenting RNNs with sparse attention to prevent gradient vanishing is not novel in itself [19]. The only novel part of the model is relevancy screening, but why that approach is better than other sparsity methods has no grounding in the theoretical analysis in the paper [...].*" **A:** We agree with R3 that sparse attention is not novel. However, quantifying how much sparsity contributes to gradient norm is new. The theorems we provide borrow notation from [19], but apply to a large class of self-attentive mechanisms as well. To see why theorem 2 directly applies to the relevancy screening mechanism, it suffices to take $\kappa = \rho + \nu$, as at each time step we are attending to the ρ states from the relevant set and the ν states from the short term buffer. To highlight how our relevancy screening is grounded in the theory, see Fig 3 in Appendix C, where we perform an experimental trade-off analysis between κ and d by tweaking ρ and ν for our relevancy screening mechanism. This will be clarified in the text. We would like to emphasize that theorem 2 applies to any κ-sparse self-attention recurrent model, but we harness the structural understanding derived from the theoretical framework to propose a more scalable form of sparse self-attentive RNN.

Q"*Experiments: I think the experiments in the paper are quite weak [...].* **A:** We appreciate this remark and refer to this rebuttal's opening statement about scope. We also point out that our methods appreciably improves generalization to out-of-distribution samples over baseline. This is a promising avenue and will be further discussed in the text. We also want to thank R3 for pointing out the Transformer-XL reference, which we will make sure to include in the main text.