Response to Reviewer 1:

"It is not clear what terms eventually vanish and why." Terms multiplied by weight matrix W (red boxes) in equation after line 97 vanish if the largest eigenvalue of matrix W is less then 1. Multiplying it repeatedly causes the vanishing gradient. This reduces equation after line 97 to equation in line 100.

"Euclidean distance is penalized if saliency goes over 1." We use the normalized Euclidean distance as mentioned in line 157 so the distance will never go over 1.

"The fMRI dataset do not show anything more than what previously laid out" The goal of the dataset is to show a real-case application where we are interested in seeing how features importance change across time. Previous work [1] that has been done on this dataset gives a single interpretation for the entire time series and does not show how features change across time. To clarify our point we will add the off-task time to our experiments (time where subject is listening to instructions not preforming the actual task) and we will investigate the ability of our model to ignore this time and put the importance on the on-task time.

"The definition of self-attention is not up-to-date." We choose to use the definition of self-attention of Lin et al. (2017) since they are applying self-attention to a RNN, similar to our case. Also, we do not use the same attention function described by Vaswani et al. (2017).

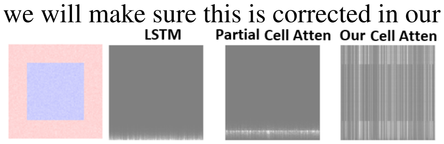
"Improved saliency comes from the attention done in the last time steps of the RNN." We disagree. If this was true then removing attention from first time steps would make no difference which is not the case. Figure 1a shows an experiment where we applied attention only to the last 10 time steps (referred to as partial cell attention) for middle box dataset, saliency still vanishes. Our proposed method improves saliency because, at each time step, cell-attention attends to different inputs from current or previous time steps preserving importance through time.

"Proposed approach ignores the temporal nature of the problem". We disagree. Our entire paper is based on time series, permuting data in time and producing different saliency is the entire purpose of our synthetic dataset. The moving box experiment in line 158 and figure 1 in supplementary material shows a clear example where the only difference between samples is their location in time, it is very clear from the experiments that we do *NOT* ignore the temporal nature of data.

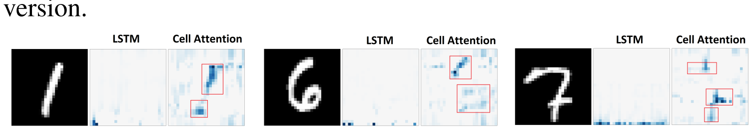
Minor comments. For weighted Jaccard, we compare the saliency with the absolute value of synthetic data sample, we will update this in the final version. We will add numbering to all equations, correct mentioned typos and fix coloring for figure 1 in supplementary material.

Response to Reviewer 2: Thank you for your comments. In the original version of the paper, we mentioned related work briefly in the introduction but we did not have an entire section dedicated to related work due to space limitation. In the revised draft, we will add a related work section and make sure we cite and explain all papers you listed in this section along with others. Our scope in this paper is on studying saliency of RNNs where we propose an approach to resolve the vanishing saliency that hinders the interpretation of such networks. Thus, we have kept the comparison with vanilla attention and other non-recurrent network architectures to our future work. Thank you for suggesting to add other standard benchmarks to our experiments. Upon your suggestion, we decided to add two new benchmarks. (1) We use MNIST dataset as a time series data where one dimension of the 2D images acts as the time axis (a 28×28 image is turned into a sequence of 28 time steps, each of which is a vector of 28 features). We choose MNIST because it offers an interpretable visualization. Figure 1b is an example of a saliency map produced for vanilla LSTM and our proposed LSTM + cell attention. (2) CMU Multimodal Opinion Sentiment Intensity (MOSI, Zadeh et. al 2016) , a dataset of opinion level sentiment intensity in online videos. In the final version, we will include these new experimental results. Finally, we will include distributions of weight matrices of the network, as suggested.

Response to Reviewer 3: Thank you for your comments. You are correct that the saliency is computed for each input captured and accumulated till the current time step. We will make sure to make this point more clear in the final manuscript. The accumulation effect is reduced by the approximation mentioned in the paragraph under line 121. We called our method cell attention because its attention is on the cell level rather than hidden layer level although we understand your concern about this name and how this might create some confusion. We may consider changing the name of the method to Recurrent Attention. For lines 118 and 119: A_t has dimensions $r \times t$ where t is the number of time steps in the current input, A_t has a weight for each time step; weight of all time steps sum up to 1. For lines 120 and 121 M is flattened to a vector of length $r * N$ and W_M is a matrix of $h \times (r * N)$. Thank you for pointing this out we will make sure this is corrected in our final version.



(a) Experiment showing that having cell-attention **only** in the last time steps (Partial Cell Atten.) still produces vanishing saliency.



(b) An example of saliency map produced for MNIST, when treated as a time series. Saliency vanishes for vanilla LSTM while our proposed model is able to detect important features.

[1] Thomas, Armin W., et al. "Interpretable LSTMs for whole-brain neuroimaging analyses."