We thank all reviewers for their careful reviews and many positive comments, including **R1**: "really valuable to the neuro community", "gives a roadmap for using NNs...to tell us how brains work", "I really liked this paper"; **R4**: "this paper is novel and significant," "well-written and relatively easy to understand"; **R3**: "could be an interesting contribution to the field." Even the most negative reviewer **R2** stated: "a highly original contribution with huge potential in the field," "the analysis...is of uniformly high quality, "with the supplement, the methods are clear and most model explanations make sense." We now address major reviewer concerns and clarify our contributions, as detailed below:

Extension to deeper CNNs without spatial invariances in stimuli (R2,R3,R4): While we demonstrated a novel application of attribution methods to model reduction of 1-hidden layer CNNs, specifically to validate deep CNNs in the retina where we had neurobiological ground truth, we can easily extend our method to deeper CNNs of depth D processing natural movies through a dynamic programming (DP) approach that works backwards from layer D to layer 1. First, note a natural movie of limited duration without spatial invariances is still well approximated by a low dimensional trajectory in both pixel space and every hidden layer. Let K be the max dimensionality for spatial input patterns for any channel in any layer. Then the basic idea is to attribute the response in layer D to the K dimensional space of inputs to each channel in layer $D-1$ using integrated gradients. We first find the important channels in layer $D-1$ using methods in our paper. Then we recursively iterate via the same method to layer $D-2$ and so on down to the pixel layer. Because of the DP-like nature of our algorithm, the computational complexity (after dimensionality reduction to K) is $O(DKC)$ where C is the max number of channels in a layer, and not exponential in D as **R3** worried. The end result is a set of important channels in each layer, along with, for each important channel $\leq K$ linear combinations of neurons that matter for generating the response in layer D. We are actively pursuing this method in deeper networks, but we will share pseudocode for this algorithm in a revised version before acceptance to NeurIPS. However, consistent with **R2,R3,R4** we feel completing this program is well beyond the scope of this paper, especially since neurobiological ground truth is missing for higher areas. But we hope our success in the retina and the extendability of our approach to deeper networks, will provide a great roadmap for neuroscience as recognized by **R1**.

Experimental evidence for our new model of omitted stimulus response (OSR) (R3): As shown quantitatively in [2], the model subunits match bipolar cells (BCs), and the 3 in the OSR correspond to fast OFF, fast ON and slow ON BCs, thus mapping directly to biological pathways. Furthermore, multiple BC types can connect to a ganglion cell (GC) (Asari and Meister 2012). Thus our new model is basically consistent with known anatomy. However, we leave *further* physiological validation of our model, beyond successfully generating the OSR, to future work, which would require painstaking experiments to perturb BC pathways and observe GC responses. We believe it is already a substantial contribution to show our approach *automatically* extracts validated models for 3 stimuli, and provides a new, experimentally testable model for a fourth (we will add suggested experiments to the paper). The main aim of our paper is to publish our new hypotheses in order to stimulate multiple retina labs worldwide to tackle the difficult neurophysiology experiments. In this manner, our theory could generate new experimental progress in future work.

Simpler approaches do not suffice (R3): A single linear receptive field (RF) plus a nonlinearity (LN model) cannot account for any of the 4 stimuli (indeed that is precisely why these stimuli are interesting). References from **R3** show that ON/OFF pathways differ in their threshold as well as timing, and optimized two-pathway LN models could partially capture the OSR [17] but *cannot* produce sufficient frequency-dependent shifting of the latency [18]. Thus the reported asymmetries cannot produce the observed OSR response, and our new finding is that three pathway LN models can.

Clarifying our contribution beyond previous work (R3): While building on a deep retina network from the authors of [2], that work did not provide conceptual understanding of *how* the network generated responses to 4 highly structured stimuli, and *whether* it generated those responses the same way the retina did. We provided such an explanation, showing only 3 of 8 channels were required to generate responses to all 4 stimuli in an interpretable manner, thereby demonstrating a single approach (natural scenes -> deep CNN -> model reduction) that can *simultaneously* discover what was previously only discovered piecemeal across ≥ 10 papers. We feel this yields a major advance in providing a "roadmap for neuroscience" (**R1**). Moreover, our method is primarily a *novel application* of attribution methods in [9,10] to model reduction in neuroscience with validation in a biological circuit. From the NeurIPS call for papers, such *application* papers are squarely within conference scope, and major advances in attribution methodology should not be required for acceptance since that direction is orthogonal to our application to model reduction in neuroscience. We will however revise to tone-down, discuss limitations, and clarify specific contributions (**R3,4**).

Revising the text (R2) We will follow **R2**'s excellent suggestions; we will shorten the intro, expand results, move info from Fig. 2 caption to text, and provide more background on integrated gradients in the main, using the extra page for the camera-ready. We note **R2** gave the lowest score (4 compared to 8 (**R1**) and 7 (**R4**)), despite being very positive (**R2**: "highly original contribution with huge potential," "with the supplement, the methods are clear"). We hope, given our restructuring, **R2** will be convinced that the revised version will be acceptable.

Other comments (R1-R4) Though we cannot address all remaining less major comments in the author response due to lack of space, we assure reviewers we can easily do so in the revision. We are grateful for your excellent suggestions.