We thank all reviewers for their insightful feedback and endorsements. We will incorporate all suggestions in our final version and have addressed the main comments and questions below:

**[R1 - Biological plausibility and rationale behind the soft-attention module]** The proposed attention module is inspired by a core property of information processing in the brain, namely, its flexible and selective nature in the face of limited neural resources. By incorporating the module within the encoding network, we can model this selective process to select certain portions of visual stimuli (the attention "spotlights") for subsequent processing at the expense of others. The use of multiplicative scaling factors to modulate feature maps has some biological grounding insofar as it is loosely inspired from the notion of gain modulation in existing studies on biological attention. A biologically plausible computational model of attention would capture both bottom-up as well as top-down influences of working memory and context as they may ultimately constrain which locations are selected. While this is beyond the scope of our present work, we believe the findings presented here provide further motivation for future work in this direction.

**[R1 - "standard" fMRI preprocessing]** We fully agree. We will report the data pre-processing operations following the guidelines recommended in the suggested study.

**[R1 - Evaluation]** We agree with the limitations. We wanted to isolate the stimulus-driven cortex in a data-driven manner rather than relying on pre-defined atlases or other task-based functional localizers to identify voxels of interest. We further identified these voxels solely based on the inter-group correlations within the training dataset and presented the performance of different models at varying thresholds of synchrony from very loose (0.15) to very strict (0.75). However, we do understand that this evaluation methodology may still induce biases and we will acknowledge the limitations in the discussion. The number of voxels varies from 160,900 to 8,804 as we vary the synchrony threshold from 0.15 to 0.75. We will also include this trend in our final version.

**[R1 - fixations during training and the last contribution]** Perhaps we were not clear in the statement. We here refer to the learned attention model that does not employ fixation data during training or testing. The attention network is trained on top of the representation network for the goal of neural response prediction. As a consequence, this network only requires stimulus-response pairs and no eye-tracking data.

**[R1 - modeling v/s not modeling]** We will rephrase the statement to be more specific. Here, we mean that modeling attention as re-weighting of stimulus representations based on spatial attention masks is beneficial in response prediction.

**[R3 - Attention vs gaze]** We agree with the reviewer's point. This is an important distinction and we will add the clarification in the introduction section to resolve this difference and shed light on the different types of attention.

**[R3 - Inter-subject variations in saliency maps]** This is an intriguing direction and is part of our ongoing work.

**[R3 - Future improvements for gaze prediction]** There are several ways in which saliency prediction can be improved with our method. Here, we focused on simplicity as a proof of concept. A straightforward extension would be to add the attention module on top of both low-level cues and high-level representations or to combine feature maps across layers before presenting to the attention network. Further, attention selects across space and time - here we focus on spatial selection of stimuli but it is likely that modeling temporal context can lead to substantive improvements. Context can also help in highlighting attentional targets that may be driven by "surprise". We will expand on this in the Discussion.

**[R1 - One particular type of attention]** Our study integrates a *spatial attention* module within a neural encoding model. However, the proposed approach is not restricted to this particular form of attention. For example, spatially global *feature-based attention* can also be studied within the context of neural encoding models as "channel-wise" attention-weights instead of spatial attention masks. We believe the observation that neural response prediction may be a useful supervision goal to study attentional deployment is particularly exciting and can be extended in novel ways. Promisingly, as pointed out by **R4**, our study highlights that attention can be studied even with the majority of naturalistic fMRI datasets with no eye tracking.

**[R1 - Relationship to prior work]** We apologize for missing any related prior work. Upon further literature search, we came across a somewhat related paper modeling mouse V1 spiking activity while explicitly accounting for gaze shifts [Sinz2018, NeurIPS], that we will include in our final references. However, the proposed approach is based on knowing fixation locations during training/testing and is restricted to predicting V1 responses.

**[R1 - "neural response is dominated by sensory signals at attended locations - what is the novelty?"]** We understand the reviewer's concern. While it is known that neural responses are dominated by sensory signals at attended locations, we here demonstrate that this property can indeed be leveraged in the neural response prediction task with the dual goal of (a) improving response prediction in later stages of the visual pathway and (b) learning attention policies employed by humans while viewing naturalistic scenes without employing any eye-tracking data.

**[R1 - "covert attention maps?"]** Since the learned attention model does not employ fixation data at all, we believe the model is equally well capable of modeling covert attention. In fact, the discrepancy between the predicted saliency and human fixation maps could partially also be explained by covert mechanisms of attention, although this speculation may be pre-mature given current analysis. Since the learned attention model is only trained to maximize neural response prediction accuracy, in principle, the attention sub-network therein should learn to focus on locations within the stimulus insofar as they dominate neural representations.

**[R4 - scale of heatmaps and code]** We apologize for the missing colorbars and will include them in the final version. We will also provide link to the code and trained models in the final version.