Reviewer #1:

**Point 1.1.** The authors investigate the magnitude of spatial attentional modulation with fMRI, measuring responses to faces presented at different eccentricities while participants perform either a central letter task, a dot task on the face stimuli or an identity task on the face stimuli. Additive effects of spatial attention are quite often observed in fMRI studies that vary stimulus strength. This may reflect the fact that top-down attentional feedback leads to increased synaptic activity in lower cortical visual areas, which could lead to a uniform increase in activity level when measured with fMRI, whereas studies of single-unit spiking activity usually find shifts in response gain or contrast gain. An observation of additive shifts in fMRI response amplitude (or raw enhancement) in this study would not be surprising.

An unusual aspect of the fMRI data is that in the face identity task, attentional modulation is greater for more eccentric faces, even though the overall response to faces is weaker as a function of eccentricity. (It should be noted that in the replication experiment, Figure 5, the dot task seems to lead to an additive fMRI shift when compared to the central fixation task.) The authors describe the notion of a flexible-attention framework, stating that there is a general pattern of large attentional enhancement at weak stimulus strength. This may be the case but what might be the functional reason for this?

We believe this is a good summary of the paper, and we thank the reviewer for the careful comments and queries.

**Point 1.2**. An important factor to consider, which is not discussed in the manuscript, is that of task difficulty. Greater attention may need to be directed to a stimulus when the task is more difficult. Several previous fMRI studies have found that the magnitude of attentional modulation tends to increase as a function of task difficulty. What is performance accuracy in the face identity task as a function of eccentricity? Is there a trend of poorer performance for eccentric faces? How about in the dot task? What would happen if the stimuli were modified or the magnitude of the discriminations were adjusted to control for task difficulty across eccentricities? Would the pattern of results shift to one of raw enhancement? It would also be useful to report the pattern of performance accuracy in the study by Kay and Yeatman; one would expect better performance better for high-contrast stimuli.

Here, the reviewer raises two general issues. One issue is whether or not ‘task difficulty’ is connected to the results we report in the paper. The second issue, related to the first, is whether the behavioral results from the experiments are informative with regards to task difficulty.



Above are behavioral results from the position study. The results show that subjects‘ performance (i.e., d-prime and reaction time) did not vary much across eccentricities in the dot task. In the face task, there was almost no variation in reaction time. (One likely reason is that the experimental stimuli are relatively fast-paced, and thus subjects are forced to respond fairly rapidly.) We observed a decline of d-prime only in the stimuli that appeared furthest (i.e., eccentricity = 4.24°). The estimation of d-prime and reaction time might not be very accurate at fovea (i.e., eccentricity = 0°) as we only have eight trials at fovea. Overall, these results indicate that there isn’t an obvious relationship between task difficulty (as indexed by reaction time and d-prime) and eccentricity. This is not to say that one does not exist, but rather that in the current set of measurements, task difficulty does not appear to be a simple explanation for the attentional effects.

Besides the behavioral results described above, we want to emphasize that we are in complete agreement with the reviewer’s point that task difficulty might be a relevant factor. Indeed, we discuss this issue in the paper (see lines 565-572). Our view is that task difficulty should not be viewed as a ‘confound’ for observed attentional enhancements. Rather, how a subject executes the task (and how hard it is for the subject to execute the task) is part of the neuroscientific phenomenon that we are trying to understand. As Kay & Yeatman 2017 has proposed, we might understand task difficulty (at least for certain types of tasks) as reflecting the duration required for a drift diffusion process, which might be instantiated in parietal cortex and fed back as a top-down enhancement to ventral temporal cortex.

**Action:** The manuscript now clarifies and discusses these issues at length (see lines 565-578). The behavioral results are now included in Supplementary Materials.

**Point 1.3**. Another factor to consider is that attention may lead to shifts in the cortical representation of visual space. Work by Dumoulin and colleagues has suggested that covert attention directed to peripheral locations can lead to a shift in the population receptive fields of voxels in the visual cortex, in the direction of the attended location, and that these shifts tend to be greater in higher visual areas. Might such a mechanism help explain the pattern of results found here?

We thank this reviewer for raising this issue. The effects of attention on the representation of visual space have been addressed in our previous paper ([Kay et al., 2015](#_ENREF_7)). In that paper, we show that attending to peripheral face stimuli leads to increased eccentricity, broadened size, and enhanced amplitude of voxel receptive fields. In this paper, we instead focus on region-level activation, which is consistent with conventional studies that explore attention effects on contrast response functions across different visual areas ([Boynton, 2009](#_ENREF_1); [Buracas and Boynton, 2007](#_ENREF_2); [Li et al., 2008](#_ENREF_9)). Thus, the two approaches are complementary and indeed attempt to analyze and interpret the same activity measurements, but they adopt very different theoretical frameworks. We believe it is not yet clear to the neuroscience field whether any framework is more “correct” than the other; hence, both ways of thinking about the data are valuable.

**Action:** We added the references of the work by Dumoulin and colleagues and specifically discuss this point in lines 652-676.

**Point 1.4**. Finally, the regions of interest considered here, such as face-responsive regions, tend to over-represent the foveal representation. Could this be an important factor to consider? Would similar results be observed for house stimuli presented at different eccentricities when measuring voxel responses from the parahippocampal cortex, which tends to over-represent the periphery?

We see the possibility that the attentional modulation reported here might depend on the retinotopic coverage of a brain area. It is certainly possible and interesting. In fact, it is arguable that the fact that responses are weak to faces in the periphery can be described in two equivalent ways: one way is to fit population receptive fields as a method to quantify the spatial representation and establish that the representation is largely foveal; another way is to simply compute the regional-average activity and observe that responses are quite weak in the periphery (as we do in the current paper). In our view, both types of descriptions are basically equivalent.

Unfortunately, our current data do not allow us to systematically examine these issues in parahippocampal cortex, since face stimuli do not provide sufficiently strong responses in place-selective regions. This might be a good direction for future studies

**Action:** we add lines 640-645 to discuss this possibility and suggest that exploring this effect on more tasks and brain regions should be an important future direction.

Reviewer #2:

**Point 2.1.** In the manuscript "Flexible top-down modulation in human ventral temporal cortex" the authors propose a new method for characterizing attentional response that takes into account the stimulus characteristics being attended. Specifically, they demonstrate that the weaker the BOLD response of a stimulus, the greater the attentional modulation, in terms of eccentricity, contrast and phase. The manuscript is well written and may demonstrate an important advance in the field of visual attention.

We thank this reviewer for this summary of our manuscript.

However, I have a few concerns that would need to be addressed before I endorse publication. Primarily, there is no report of behavioral performance, limiting the ability to assess the model.

**Point 2.2**. My primary concern is that performance is not reported. It is difficult to make assumptions about visual attention without taking into account attention's effect on behavior. For instance, it is likely that performance of tasks is poorer in the periphery, suggesting that more attention is deployed to stimuli in the periphery. Was difficulty titrated across conditions? Was performance worse for stimuli in the periphery? Can the authors discount the claim that they are showing attentional modulation is stronger when the task is harder? If behavioral data from inside the scanner is not available, can a control study be run outside the scanner to demonstrate that the results do not simply reflect task difficulty?

Here, the reviewer raises the issues of ‘task difficulty’ and ‘behavior’, similar to Point 1.1 raised by the first reviewer. We agree that these are important issues that are relevant to the topics of the paper, and we have made some revisions to address these issues.



Above are behavioral results from the position study. The results show that subjects‘ performance (i.e., d-prime and reaction time) did not systematically vary much across eccentricities in the dot task. In the face task, there was almost no variation in reaction time. We observed a decline of d-prime only in the stimuli that appeared furthest (i.e., eccentricity = 4.24°). The estimation of d-prime and reaction time might not be very accurate at fovea (i.e., eccentricity = 0°) as we only have eight trials at fovea. Overall, these results indicate that there isn’t an obvious systematic relationship between task difficulty (as indexed by reaction time and d-prime) and eccentricity. This is not to say that one does not exist in general, but rather that in the current set of measurements, task difficulty does not appear to be a simple explanation for the attentional effects.

In addition to revising the paper to add the behavioral results, we would like to make clear here that we do not view task difficulty as a confound. Rather, task difficulty may indeed be one of the factors that can explain the attentional enhancements. However, we believe it is important to not be overly optimistic about the explanatory power of the concept of task difficulty. One, it may be very difficult to accurately quantify in a given experiment based solely on behavioral responses (e.g. it is not clear that simply using reaction time will accurately track difficulty in any arbitrary cognitive task). Two, it is at least plausible that the nature of the task itself might also matter to whether attentional enhancements arise in a given region. For example, during the face task, enhancements are substantially stronger in FFA (see Fig. 4) compared to the dot task, and it is plausible that asking the observer to make judgments about face identity somehow provides additional enhancement beyond the mere difficulty of the task.

Another good example comes from Point 2.6 raised by the reviewer. Specifically, the face task involves a memory component and if that is relevant to the attentional modulations, this is not fully subsumed by the concept of ‘task difficulty’.

**Action**: we revised the lines 565-578 to emphasize our argument here. The behavioral results are now included in Supplementary Materials.

**Point 2.3**. Another concern is the definition of areas in early visual cortex across eccentricities. The reversed pattern of attentional effects in early visual cortex as compared to higher order cortex may be due to the fine retinotopic specificity in early visual cortex. The authors report "We noticed that voxels with population receptive fields in the periphery often exhibit negative BOLD responses when face stimuli appear in the central visual field. To avoid this complication, we positively rectified all voxel responses". Suppression is a well documented property of visual attention when stimuli appear in neighboring eccentricities and models of visual attention need to account for suppression. Two things here might be warranted. First, the authors might consider dividing the early visual areas into different eccentricity bins and measuring the response of stimuli only for that bin. Higher level visual cortical areas have less pronounced divisions based upon eccentricity, so these methods are probably not necessary. Secondly, the authors could use an index that includes the attended condition in the denominator, for instance (Rdot/face-Rdigit)/ (Rdot/face+Rdigit).

We thank this reviewer for raising these concerns. We agree that the observation of negative BOLD may be related to attentional suppression. However, such effects are not in the scope of the present study (and hence, this is why we remove such effects through rectification). In general, we would like to point out that compared to high-level visual areas, the attentional effects that we have measured in early visual areas are fairly modest. In particular, we investigated the eccentricity-dependent effects in the position study (([Kay et al., 2015](#_ENREF_7)), see figure below). Overall the attentional effects are minimal in early visual cortex. We quantified the pRFs as reflected in positive evoked BOLD responses and the results show that the pRFs of V1 voxels do not substantially change across tasks. As you can see the scatter plots below, the tuning preference of voxels across different eccentricity bins was not significantly altered by attention. This result is also consistent with the minimal attentional effects in our region-level analysis here.

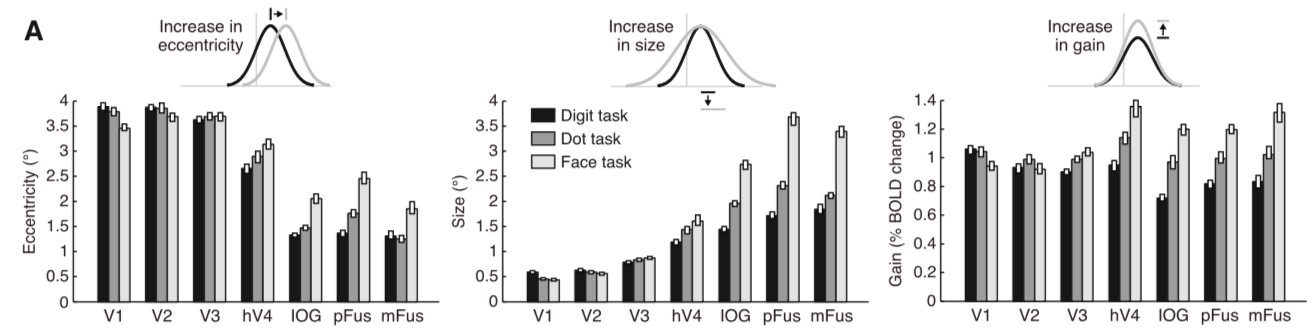


Fig. 3A in ([Kay et al., 2015](#_ENREF_7))



Fig. S3C in ([Kay et al., 2015](#_ENREF_7)). The voxel preferences in early visual cortex (highlighted by the red rectangle) did not substantially change across tasks.

With regard to the second suggestion (alternative attentional index), we calculated this index as below. The results are very similar to the results of percent enhancement. One drawback of this index is that it is somewhat hard to interpret its physical meaning. As such, we do not include this figure in the manuscript.



**Action:** we have revised the lines 196-199 and 668-676 accordingly in the manuscript to reflect our answer to this point.

**Point 2.4**. There are certain analyses that do not appear to support the authors' model. For instance, word and face phase coherence. It appears that peak attentional modulation occurs at the second or third lowest coherence. Can the authors account for this discrepancy?

Yes, the phase coherence data do not look like monotonically decreasing functions. We speculate that this may be due to the fact that 0% phase coherence images contain pure noise and thus it may be much easier to perform the one-back task on them compared to ‘mixture’ images. Ideally, future studies would develop a formal quantitative model of this type of explanation.

**Action:** we strengthen the explanations in lines 398-405 and 585-594.

**Point 2.5**. The low sample size is a concern for proposing a novel model of attentional processing. This is tempered by the number of tasks each participant has run, but it would be useful to see that this effect transfers to at least another couple of participants.

We fully agree that replication is important to substantiate our claims. But we want to emphasize that this is exactly the reason why we included two independent studies. The two studies contain four experiments, five tasks, and three stimulus dimensions (eccentricity, contrast, phase), and several visual regions. Also, there is in general a trade-off between the number of stimuli/tasks to sample and the number of participants to scan. In the position study, we prioritized the former (i.e., more stimuli/tasks) rather than the latter aspect. We thus only scanned three subjects but collected nine sessions for each subject to establish the population receptive field model of face-selective regions. In the category study, we scanned nine subjects and sampled significantly more stimuli than most imaging studies. We believe that the strength of evidence provided here is on par or greater than most conventional neuroimaging studies.

**Point 2.6**. In Figure 1 each of the models has an equation to characterize the attentional response. Do the authors have a suggested equation for their model?

We do not have a specific mathematical equation to suggest here because one of our central claims is that attentional effects are highly task-dependent and the field as of yet lacks detailed models of variety of cognitive tasks. Cognitive tasks are remarkably diverse, and our belief is that using a fixed parametric form, as in conventional approaches, overly simplify the neural processing of a task. We thus do not suggest a single mathematical function here and conservatively called it a “flexible-attention framework”.

**Action:** we explain this issue in lines 533-536.

Other concerns:

**Point 2.7**. A one back task may increase memory demands in addition to attention.

We agree that the one-back (face) task involves more cognitive components (e.g., memory) than merely attention. Both the dot and the face tasks require subjects to attend to face stimuli, but only the face task imposes additional memory processing. The differential response patterns in the two tasks might reflect the potential effects of the memory component. How these components jointly modulate neural activity in visual cortex is an interesting topic that deserves further concerted study. As such we conservatively call it “a flexible top-down framework”.

**Action:** we explain this issue in lines 538-542.

**Point 2.8**. Why do the authors use a confidence interval of 68%.

68% confidence intervals correspond to plus-or-minus one-standard deviation of a Gaussian noise distribution (thus, they are akin to standard errors).

**Action:** We add the explanations for this choice in line 231.

**References**

Boynton, G.M., 2009. A framework for describing the effects of attention on visual responses. Vision Res 49, 1129-1143.

Buracas, G.T., Boynton, G.M., 2007. The effect of spatial attention on contrast response functions in human visual cortex. J Neurosci 27, 93-97.

Cutrone, E.K., Heeger, D.J., Carrasco, M., 2014. Attention enhances contrast appearance via increased input baseline of neural responses. Journal of Vision 14, 16-16.

Epstein, R.A., Baker, C.I., 2019. Scene Perception in the Human Brain. Annual Review of Vision Science 5, null.

Hopf, J.M., Boehler, C.N., Luck, S.J., Tsotsos, J.K., Heinze, H.J., Schoenfeld, M.A., 2006. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. Proc Natl Acad Sci U S A 103, 1053-1058.

Itthipuripat, S., Chang, K.-Y., Bong, A., Serences, J.T., 2016. Stimulus visibility controls the balance between attention-induced changes in contrast appearance and decision bias. bioRxiv, 091421.

Kay, K.N., Weiner, K.S., Grill-Spector, K., 2015. Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. Current Biology 25, 595-600.

Kay, K.N., Yeatman, J.D., 2017. Bottom-up and top-down computations in word-and face-selective cortex. eLife 6, e22341.

Li, X., Lu, Z.L., Tjan, B.S., Dosher, B.A., Chu, W., 2008. Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. Proc Natl Acad Sci U S A 105, 6202-6207.

Murray, S.O., 2008. The effects of spatial attention in early human visual cortex are stimulus independent. J Vis 8, 2 1-11.

Reynolds, J.H., Heeger, D.J., 2009. The normalization model of attention. Neuron 61, 168-185.

Silson, E.H., Chan, A.W., Reynolds, R.C., Kravitz, D.J., Baker, C.I., 2015. A Retinotopic Basis for the Division of High-Level Scene Processing between Lateral and Ventral Human Occipitotemporal Cortex. J Neurosci 35, 11921-11935.

Silson, E.H., Groen, II, Kravitz, D.J., Baker, C.I., 2016. Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. J Vis 16, 14.

Stormer, V.S., Alvarez, G.A., 2014. Feature-based attention elicits surround suppression in feature space. Curr Biol 24, 1985-1988.

Summerfield, C., Egner, T., 2016. Feature-Based Attention and Feature-Based Expectation. Trends in Cognitive Sciences 20, 401-404.

Williford, T., Maunsell, J.H., 2006. Effects of spatial attention on contrast response functions in macaque area V4. J Neurophysiol 96, 40-54.

Wyart, V., Nobre, A.C., Summerfield, C., 2012. Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. Proceedings of the National Academy of Sciences 109, 3593-3598.