# The Effects of Attention on Neural Representations in Working Memory

By

Andrew D. Sheldon

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Neuroscience)

at the

UNIVERSITY OF WISCONSIN-MADISON

2017

Date of the final oral examination: 6/1/2017

The dissertation is approved by the following members of the Final Oral Committee:

Bradley Postle, Professor, Psychology

Xin Huang, Associate Professor, Neuroscience

Scott Reeder, Professor, Radiology

Yuri Saalmann, Assistant Professor, Psychology

Bas Rokers, Associate Professor, Psychology

Matt Jones, Associate Professor, Neuroscience

**Acknowledgements**

Thank you Brad, for challenging me and believing in me the whole way through. Thank you for all that you taught me, in how to be a good researcher, student, mentor, and member of the scientific committee. Thank you for the incredible faith you had in me, and the loyalty you showed. Thanks for always fighting for me.

Thank you to my committee members, Xin, Bas, Yuri, Matt, and Scott for putting up with all of my last minute scheduling changes, and for providing excellent feedback and guidance along the way. Thank you for challenging me.

Thank you mom, dad, and Donnie for all the support you gave me. Thank you for always being there when I needed to talk. Thanks for the inspiration, words of encouragement and for allowing me to take my mind off things when I needed a break.

Thank you to all my friends, who I don't have enough space to name. You know who you are and I hope I got a chance to tell you what you mean to me and how much help you provided. Thank you for the long nights of conversation and companionship. Thank you for judging me only when I needed it, and putting up with my BS. Thank you for inspiring me.

Thank you to my fellow lab mates: Elyana, Muhammet, and Sawyer, for your incredible contributions to the current work. Thank you Jason, Josh, Mike, Nate, Emma, Adam, Qing, Ying, and Jackie for all your help in teaching me what I know today. Thank you to all my collaborators and teachers including those at the HTC and classmates.

Thank you to the NTP and MSTP for all your help with paperwork and keeping me on top of my graduation requirements!

# Contents

**Abstract**

What mechanisms underlie the prioritization of neural representations of the surrounding environment in order to effectively guide behavior? Selective attention is the ability to preferentially process incoming sensory information that is relevant for the task at hand, maximizing the deployment of finite cognitive resources. In situations where the relevant information is non-temporally coincident, the suite of cognitive functions collectively referred to as Working Memory enable the temporary maintenance of relevant sensory representations in the absence of sustained input. For primarily historical reasons, these two cognitive phenomena have been treated separately by the field, though recent theoretical formulations of working memory have recognized the parsimonious appeal of a common mechanism underlying both the prioritization of sensory representations of information still present in the environment and from the recent past. Here, the results of three studies addressing the commonality of mechanism between tasks of visual selective attention and working memory are reported, with a specific focus on the effects of task-dependent prioritization on neural population-level representations.

# Chapter One:
# General Introduction

*"Chaos is the law of Nature, Order is the dream of man."*

*-Henry Adams*

We live in a world of perpetual stimulation. Evolutionary selection pressures led to the development and subsequent refinement of sensing apparatuses which are continuously bombarded with the traces of events occurring in the surrounding world, many of which require responsive action to ensure continued survival. This abundance of information about the external world, while helpful when effectively utilized, brought along with it two new selection pressures. First, those organisms that could best prioritize the information coming in would possess a distinct advantage over their peers by utilizing their limited resources most efficiently. The resulting prioritization map is, by definition, the phenomenon of "attention" (Posner, 1980), and the mechanisms regulating its allocation are part of a larger suite of cognitive control functions that appear to be uniquely well-developed in humans.

Second, the most advantageous response to a particular event often depends on the presence or absence of other, related events. For example, were one to see a lion 50 feet away, one's response will be quite different if one were camping alone in the Serengeti than if one were spending the afternoon at the local zoo. While it is sometimes the case that these related, "contextual" events co-occur temporally with the putative event-of-interest, this is not always

true. Thus, behaviors that take into account related events separated in time require the existence of mechanisms for maintaining information about those events over the necessary time-frame. The short-term retention (STR) of information is believed to be achieved via specific neural mechanisms operating over time scales on the order of seconds or minutes, distinguishing them from those governing the long-term retention (LTR) of information (operating over time-scales on the order of days or longer).

Taken together, the intersection of the aforementioned suite of constructs of cognitive control and the STR of information forms the basis for "Working Memory (WM)", a cognitive ability whose integrity is believed to be important for human health[1]. Despite this, the neural mechanisms underpinning WM remain poorly understood. The intended purpose of the current work is to propose a series of experiments designed to elucidate these neural mechanisms via the domain of visual motion, one that possesses a variety of properties that make it amendable to novel quantitative analyses. But before presenting the specifics of the proposed course of study, it will be beneficial to review the currently accepted theoretical framework surrounding WM for visual motion, as well as the state-of-the art methods that hold the potential to answer outstanding questions in the field.

*"Short-Term Memory (STM)" vs "Long-Term Memory (LTM)"*

Taking a step back to discuss in more detail the phenomena of the STR and LTR of information, there is no intrinsic reason why separate mechanisms need be in place that operate over short (~seconds to minutes) compared to long (~ days or longer) timescales. Nevertheless, experimental evidence suggests that this may, in fact, be the case. The strongest evidence for

separate mechanisms governing memory over short timescales vs long timescales comes from lesion studies involving patients with damage to the integrity of the medial-temporal lobe diencephalic memory system (MTL). These patients, possessing a clinically-defined pure anterograde amnesia (an inability to perform above chance on tests requiring the formation of new long-term declarative memories) will nevertheless typically perform normally on tests of STM. To fully interpret these results, it is important to consider how tests of LTM and STM are performed. In both cases, a to-be-remembered item or set of items (the "sample") is presented. This is followed by a "delay" period of variable length, where information regarding the identity of the sample is no longer present in the environment. After this, a test "probe" appears requiring the participant to either freely recall the sample (a "recall" probe), or to compare the sample item to some new exemplar, usually by making a match or non-match judgement (a "recognition" probe). Tests of STM usually consist of delay periods that last on the order of seconds or minutes, while tests of LTM usually consist of delays spanning multiple hours, days, or even longer. Because patients with damage to the MTL perform well on tests of STM but not LTM, it implies that the two systems are independent of one another. However, it is still possible for information to be held simultaneously in both systems (Ranganath and Blumenfeld, 2005) or even for information held in LTM prior to testing to support information currently held in STM (Postle, 2007; Lewis-Peacock and Postle, 2008). Indeed, the interaction of attention and other executive control mechanisms with STM gives rise to another concept that will be useful to discuss here: that of "working memory"

*The "Central Executive" and Working Memory*

The concept of working memory as an operational construct first arose in the context of a task designed by Baddeley and Hitch (1974) that purported to test, simultaneously, the STM for two distinct domains of information: phonological and visuospatial. In doing so, they found that when participants maintained information from both domains in STM, performance was unimpaired when compared to trials when participants held only one domain in STM, despite the fact that the amount of information stored appeared to be twice as large. This led them to propose the existence of independent STM buffers for each domain (a "visuospatial sketchpad" and a "phonological loop"), as well as a "Central Executive (CE)" which could recruit and coordinate multiple buffers simultaneously, as needed (Baddeley, 1986). They dubbed this framework the "multiple component model", and it should be noted that this concept of the CE was not intended to be limited to tasks involving memory. The concept of "Working Memory (WM)" then can be formulated as tasks requiring both the STR of information and the engagement of the CE to perform the necessary cognitive control operations, including but not limited to: coordinating multiple STM buffers for item storage; shifting attention between representations held in STM; or manipulating the stored information in some way, such as in tasks requiring "mental arithmetic" or "mental reordering" of stimuli for subsequent report. The formulation of WM as STM+CE is a useful, if simplistic operational definition; for a more expanded review, see Postle and Pasternak (2009). The last concept related to working memory that needs to be covered before diving in to the search for neural mechanisms is that of capacity limits.

*A Brief Consideration of WM Capacity*

Unlike long-term memory, in which the capacity is believed to be so large that it is functionally infinite, tests of short-term memory show limits in the amount of information that can be maintained across a delay period (Miller, 1956). From pioneering work by George Miller who reported the capacity to be "the magical number seven, plus or minus two", to Cowan's magic number four (Cowan, 2001), many models seek to define this capacity in the context of a number of items. Taken to an extreme, this is the theoretical framework of the "slots" model, (Pashler, 1988; Luck and Vogel, 1997; Zhang and Luck, 2008) which proposes that working memory is limited only by the number of items, and not the complexity of individual items. This is in direct contrast to a resources model (Wilken and Ma, 2004; Bays and Husain, 2008; Bays et al., 2009), wherein an unspecified cognitive resource is allocated among items in working memory in order to represent them. Thus, in this model, one can store many items, but each one will have low representational fidelity (as the resources get divided amongst the many items). The substrate of this resource is not specified, but some theories postulate that each item beyond the first interferes with other items, creating the appearance of a resource being allocated (Barrouillet, 2004). This last model is somewhat unique among working memory capacity models in that it proposes a specific role for attention in the setting of capacity. Fundamental questions yet remain about the mechanisms governing working memory capacity. A greater exploration of behavioral capacity limits explored over a fixed feature space coupled with neural estimates of mnemonic precision is needed to help distinguish among these proposed models.

*Persistent activity and the role of pre-frontal cortex (PFC)*

Any aspiring comprehensive theory of WM must account for three things. First, how is information stored? And what accounts for observed capacity limits? Second, where in the brain is that information stored? Is it localized to a particular area, and if so, what area? Or is it distributed widely across the brain? Finally, how is the act of manipulating information stored in working memory (including, but not limited to, operations such as "mental arithmetic" or even simply the shifting of the allocation of attention) achieved? An influential early model attempting to answer the first of these questions centered around the idea of persistent neural activity (PNA) first proposed by Lorente de Nó (1933). Under this model, information about a stimulus was encoded in neural activity that persisted after the initiating stimulus had disappeared from the environment. In 1949, Donald Hebb refined this concept further, postulating that a "transient memory" for a stimulus could be sustained via the "reverberatory" activity of cell assemblies. Around this same time, lesion studies had found that intact PFC function was integral to successful performance of WM tasks (Pribram et al, 1964). Importantly, the investigators of these studies were careful to note that this did not necessarily imply that PFC was the site of information storage. In other words, PFC lesions could cause deficits in WM performance through the loss of inhibitory control over Proactive Interference, a phenomenon wherein information from previous trials interferes with information to be remembered on the current trial. This was demonstrated nicely in a study by Milner (1964) — incidentally published in the very same volume (Warren and Akert, 1964) as that of Pribram et al. (1964).

In 1971, however, the development of electrophysiological recording techniques enabled for the first time the measurement of activity of individual neurons in primate cortex. Influenced by the ideas of Hebb and Nó, as well as the lesion studies of Milner, Pribram and many others, two groups recorded the activity of neurons in PFC during tasks of WM. Fuster and Alexander (1971) used a delayed-recognition paradigm, while Kubota and Niki (1971) used one of delayed-alternation. Both groups found neurons that modulated their firing rate during the delay period of the task compared to sample presentation. At first glance, these neurons appeared to be the "reverberatory" persistently active units proposed by Hebb and Lorente de Nó. Indeed, subsequent work in the monkey found evidence for PNA in a variety of brain areas including posterior parietal cortex (PPC; Gnadt and Andersen, 1988) and inferior temporal (IT) cortex (Miller et al., 1993). When Funahashi et al. (1989, 1990) found that these persistently active PFC neurons modulated their response output based on the retinotopic position of a saccade target during an oculomotor match to sample WM task, this was heralded as proof that this delay-period activity was encoding the identity of the sample stimulus. A major proponent of this theory was Patricia Goldman-Rakic (1987) who saw in these data the experimental evidence supporting the proposed STM buffers of Baddeley and Hitch (1974). Furthermore, with the advent of functional magnetic resonance imaging (fMRI), able to measure the blood oxygen level dependent (BOLD) signal (an indirect correlate of neural activity), evidence in support of the PNA model in humans was inferred from the observation that BOLD signal during the delay period of WM tasks was elevated in PFC, but not in posterior sensory areas. This was the reverse of the pattern observed while the sample was still present in the environment (reviewed in Curtis and D'esposito, 2003).

There are, however, three aspects inherent in the designs of these studies (both with fMRI in humans and single unit recordings in monkeys) that make this interpretation of the results problematic. First, they possess methodological factors that confound the action of attention, motor planning and the STR of information, as expertly demonstrated in a series of modified WM tasks designed by Akiko Ikkai and Clay Curtis (2011). Second, by treating each neuron (in the case of the monkey work) or each voxel (volumetric pixel in fMRI) as independent units and only looking at the information contained in the activity of that individual unit, these studies miss out on the ability to decode information contained at the level of networks of neurons (Cox and Savoy, 2003). Third, the fundamental assumption that informative units in PFC should act as "feature detectors" for representations stored in WM, analogous to the orientation-selective cells in primary visual cortex famously characterized in the seminal experiments of Hubel and Wiesel (1968), has recently begun to be called into question (Rigotti et al., 2013).

Even so, models proposing PNA, particularly in PFC, as a mechanism for the storage of information in WM remain popular, though recent evidence obtained through multivariate analyses has seen the creation of models challenging this framework.

*Multivariate Pattern Analysis(MVPA) and the Sensory Recruitment Hypothesis*

The recent explosion in the use of fMRI as a research tool to study cognitive processes, including visual WM, has seen with it a commensurate rise in the sophistication of analysis methods used to process the data. Early univariate methods have given way to the more powerful (in the parlance of testing, *sensitive* and *specific*) multivariate variety. The literature on MVPA is becoming quite extensive; for a review of multivariate pattern classification see Haynes and Rees

(2006), Kriegeskorte et al. (2006), Norman et al. (2006), and Pereira et al. (2009). By considering the activity of multiple units simultaneously, these methods achieve greater sensitivity by being able to pick up information carried in networks of neurons, even if that information is carried by unit signals that would otherwise not exceed the statistical threshold for significance when performing univariate analyses. They achieve greater specificity by not being beholden to a number of assumptions that often go into univariate analysis to get around the problem of multiple comparisons. These can include assumptions of neural mechanistic smoothness (i.e. that units in a given location represent information in similar ways) and locality (units that perform a given function are all in the same spot, thereby excluding the possibility of distributed networks). For a more in depth treatment of the specific assumptions made by univariate analyses compared with those of MVPA, see Cox et al. (2014).

These methods, when applied to the study of WM and STM (for a review, see Postle, 2015a and Sreenivasan et al., 2014), have yielded results appearing to contradict models advocating persistent activity in PFC as the mechanism and site of STR of information. MVPA of delay period activity during visual WM tasks has shown that the identity of the stimulus can be decoded from early visual areas despite the fact that these areas do not show elevated delay period activity. This has been done for a number of distinct sample stimuli, including the orientation and color of target stimuli in primary visual cortex (V1; Harrison and Tong, 2009; Serences et al., 2009); the identity of complex visuospatial patterns in occipital and parietal cortex (Christophel et al., 2012); and the identity of familiar faces, houses, scenes, objects and body stimuli from ventral occipitotemporal cortex (Han et al., 2013; Lee et al., 2013; Nelissen et al., 2013; Sreenivasan et al., 2014). Finally, of particular relevance to this volume, the direction of motion of moving dot kinematograms can be decoded from patterns of delay-period signal

intensity in early visual areas including in human area MT+, as well as from medial calcarine and extracalcarine cortex (Riggall and Postle, 2012; Emrich and Riggall et al., 2013).

Many of these studies also reported the inability to decode these same stimuli from PFC, despite robustly replicating the classic univariate finding of elevated delay-period activity in this area. This failure to decode is unlikely to be due to some inherent property of PFC rendering it inscrutable to multivariate analysis, as Riggall and Postle (2012) showed by successfully decoding the trial *task instructions* (whether the to-be-reported feature was speed or direction) from PFC.

Taken together, these results suggest a model where the sites of storage for the STR of information are the early sensory areas corresponding to those regions that are believed to be involved in the early processing of a sample stimulus, with the role of PFC shifting to one analogous to the multiple-component model's "central executive". This has been dubbed the "sensory recruitment hypothesis" (e.g. Ester et al., 2013). Under this framework, it would be predicted that the fidelity of the representations stored in these areas would correlate with behavioral report accuracy. This has been shown to be the case, both in work by Steve Emrich (in Emrich and Riggall et al., 2013) using visual motion stimuli, and through the application of inverted encoding models (IEMs; Serences et al. 2009b) by Eddie Ester (Ester et al., 2013), and Tommy Sprague (Sprague et al., 2014) while participants remembered the orientation of gradients and the spatial location of small squares, respectively. In the first example, Emrich nicely demonstrated that as the mnemonic load increased, the precision with which respondents behaviorally recalled the sample stimuli decreased in a corresponding manner to that of the accuracy of a multivariate classifier trained to distinguish stimulus identity. In the latter two examples, both Ester and Sprague first obtained estimates of the precision of participants'

mnemonic representations from IEMs that model each voxel's response to a given stimulus as the weighted sum of hypothetical channels tuned to evenly spaced components of the possible stimulus feature space (Serences 2009b). They then correlated these estimates with the behavioral precision and also found a positive relationship, confirming a prediction of the "sensory recruitment" framework.

*Candidate neural mechanisms for the Sensory Recruitment Hypothesis*

So far in discussing the evidence supporting the sensory recruitment hypothesis, I have primarily focused on the proposed site for the STR of information. Unlike with the PNA framework, which began with a proposed underlying mechanism that subsequently led researchers to attempt to find brain areas whose activity fit the model, the development of the sensory recruitment hypothesis has proceeded in reverse, with initial evidence for a likely site of the STR of information now causing a search for a mechanism. Two broad candidates have been proposed, and the supporting evidence for each is detailed below.

One proposed candidate is that of distributed, spatiotemporally dynamic neural code (like that discussed in Meyers et al., 2012 or Stokes et al., 2013). Under this hypothesis, delay-period neural spiking in sensory areas is responsible for the STR of information (not unlike the PNA). Unlike the PNA, however, the network of neurons that engage in this spiking activity changes rapidly and widely, with the result that any one neuron is only transiently involved in the STR of information, accounting for the failure of single unit recordings to find evidence for consistent delay-period activity in these areas. A conceptually related proposal is that of a neural oscillatory

code (Jensen, 2006; Axmacher et al., 2006; Uhlhaas et al., 2009). Indirect evidence for this proposal comes from the fact that these oscillatory fluctuations of neuronal membrane potentials have been shown to contribute most to the generation of the fMRI BOLD signal (Logothetis et al., 2001). Relatedly, work by Adam Riggall (Riggall, 2014 *in preparation*) reported that when training a classifier on early delay period time points only, late period decoding was significantly less robust than early delay period decoding, and vice versa when training on late delay-period time points only, consistent with a dynamic neural code. Finally, Anderson et al. (2014) demonstrated via an IEM that the STR of orientation could be decoded from alpha band (8-15Hz) oscillations in primary visual cortex, as measured with electroencephalography (EEG).

A wholly different candidate mechanism is that of transient synaptic weights (Barak et al 2010). Evidence in support of this mechanism derives from dual-retrocue studies designed to disentangle the effects of attention from the mechanisms of the STR of information. By cuing attention to one of two representations in working memory, the experiments reviewed in Postle (2015b) allow for the attempted multivariate decoding of "unattended" memory items compared to "attended" memory items. Importantly, it is the presence of a second retrocue that prevents participants from "dumping" the unattended item altogether, as in half of trials ($p=0.5$), the second retrocue will indicate that the participant is to report the previously unattended memory item. Regardless of the provenance of the data fed to the decoder — be it EEG (LaRocque et al., 2013) or fMRI (Lewis-Peacock and Postle, 2012) — these experiments show an inability to decode an unattended stimulus, even in sensory areas where a robust representation of the attended stimulus can nonetheless be decoded. While one might postulate that these unattended items may be temporarily hi-jacking Long-Term memory mechanisms and thus be transiently entering and exiting the purview of WM, a study by LaRocque et al. (2015) showed that there

was no benefit to performance on a subsequent surprise Long-Term memory task for items that spent time as an unattended memory item compared to those that spent time only as an attended memory item. This result may imply a mechanism for the STR of information that does not rely on sustained neural activity. Transient rises in synaptic calcium levels (Mongillo et al., 2008), as well as short-term synaptic potentiation mediated through GluR-1 (Erickson et al., 2009) are two leading candidates for this mechanism of transient synaptic weights, though this remains an understudied possibility in the field.

*Multivariate challenges to the Sensory Recruitment Hypothesis*

Not all multivariate evidence supports the sensory recruitment hypothesis, however. In addition to the possibility that a dynamic neural code or pattern of synaptic weights exits in early sensory areas, the same mechanisms could account for the failure to decode stimulus-identity in PFC (Stokes 2015). Along these lines, Ester and colleagues (2015) used IEMs to decode representations of the orientation of a presented stimulus during the delay period of a one item test of STM. The key difference between their approach and previous attempts to decode the STR of information in these areas hinged on a searchlight feature selection step wherein voxels were selected for inclusion based on their ability to generate robust searchlight IEM outputs utilizing an independent test set, rather than selecting voxels based on delay-period activation ala Riggall and Postle (2012). Importantly in this study, however, the object to be remembered was always also in the focus of attention.

Another challenge to the sensory recruitment hypothesis is the recent results of Bettencourt and Xu (2016), who showed a consistent ability to decode target identity during a

WM delay period from superior intraparietal sulcus (sIPS), even in the face of delay period distractors. This pattern was not observed for early visual areas, where the presence of the distractor abolished successful delay period decoding, even when participants could not anticipate the presence of a distractor.

An influential recent study by Mendoza-Halliday and colleagues (2014) highlights an important theoretical alternative to the sensory recruitment hypothesis. In this study, the authors conducted a one-item delayed match-to-sample STM task for visual motion in monkeys, while recording from either middle temporal area (MT), the medial superior temporal area (MST) or dorsal-lateral PFC (dlPFC). They found an absence of delay period spiking for units in MT, but not for units in MST or dlPFC. They achieved robust delay-period decoding of the sample direction of motion from the spiking data in both MST and dlPFC. They also were able to successfully decode from the local-field potentials (LFP) (specifically low frequencies such as the alpha and beta bands) of MT and MST. Crucially, they also observed high spike-field coherence between the spiking in dlPFC and the beta band of the LFPs in MT. From this, the authors concluded that it was the spiking in PFC that was both maintaining the representation in STM of the sample motion across the delay period, and that was subsequently driving the oscillatory activity in the LFPs of MT. Thus, under this hypothesis, the reason that many BOLD fMRI studies are able to decode from early sensory areas (like MT), is that persistent spiking in PFC drives corresponding oscillatory dynamics in early visual areas, which are what is being picked up via MVPA decoding.

As discussed previously, however, a crucial limitation of many of these studies revolves around the fact that the focus of attention is not manipulated independently of the contents stored in WM. Instead, what is being decoded is always *both* being maintained in WM and inside the

focus of attention. This leaves open the possible confound that the observed persistent activity in PFC actually reflects mechanisms responsible for *allocating attention to* representations within WM (the phenomenon of "object-based attention") rather than the mechanisms responsible for *maintaining* those representations. Thus, before continuing further, it will be helpful to review the mechanisms of top-down selective attention, with a particular emphasis placed on the effects of attention on sensory cortices.

*Early Models of Attention*

How does the brain prioritize incoming sensory information to successfully guide behavior? A version of this question has existed for at least as long as the field of cognitive neuroscience itself (James, 1890), and as a result, a multitude of studies have attempted to answer it. Work in the second half of the 20th century (reviewed nicely in Driver, 2001) cleverly manipulated this prioritization (or "selective attention") and observed the consequences on human behavior. A popular early theoretical formulation was that of the attentional filter proposed by Donald Broadbent (1958). Here, incoming sensory information enters a (relatively) high capacity buffer (the "pre-attentive" stage). Once there, only a subset is selected, based upon it possessing characteristics relevant to the task at hand, for further processing. The cognitive processes underlying this selection make up the attentional filter. Anne Treisman (1960) expanded upon this model to suggest that rather than a strict pass/no pass filter, attentional selection *attenuated* the information in the buffer such that unselected information was still processed, though to a lesser extent than the selected information. This reformulation was necessary, in part, to explain the now famous cocktail party effect, whereby salient stimuli such as one's name could enter awareness and influence behavior despite being uttered in a conversation that was outside the current attentional focus. Further challenges to these "early-

selection" models of attention pointed to increased reaction times (Eriksen & Eriksen, 1974; Stroop and Ridley 1935) as distracting information increased in similarity to attended information, as well as negative priming effects (Tipper, 1985) to argue that attentional selection occurred later, after sensory information had already been relatively well-processed. Despite attempts to reconcile these two accounts (eg Lavie 1995), it would take the advent of methods that allowed for the collection of neural data to conclusively settle the debate in favor of "early selection". Nevertheless, important concepts introduced by the "late selection" side or out of attempts to reconcile the two sides would come to be integrated into current models of attentional selection, including ideas of "Automatic processing" (Deutsch and Deutsch, 1963; Duncan, 1980), the importance of visual salience in guiding attentional selection (Eriksen & Eriksen 1974), active competition and therefore suppression of unselected information (Tipper 1985), and the flexible allocation of attention to serve different task needs (Lavie 1995).

*Feature-Based Attention vs Object-Based Attention*

Anne Treisman would also contribute feature integration theory (Treisman and Gelade, 1980), arguing that features were processed during the pre-attentive stage, only to be bound into objects once reaching the second, "focused-attention" stage. This concept of emergent hierarchical structure in the cognitive representation of external stimuli would prove hugely influential in the attentional selection field and beyond, and efforts to test its predictions would reveal the relationship between this representational structure and the associated deployment of attention. Early on, this was illustrated in the contrast between tests of spatial vs. object-based attention. A common metaphor for visual attention is that of a spotlight, highlighting a specific area of the visual field for preferential processing (Posner, 1980). This spotlight can be "covertly" allocated to a portion of the visual field without making an eye-movement, though the

act of covertly shifting spatial attention and planning motor actions to/involving the same location appear to be intimately linked (Snyder et al 1997). Evidence for spatial location as a fundamental axis along which attention can select can be found in case studies of patients with hemi-spatial neglect (Bisiach and Luzzatti 1978), where damage to the (often right-sided) parietal lobe results in patients "ignoring" the entire contralateral hemisphere. In addition to spatial location, attention can be allocated along axes of time ("temporal attention"; Coull and Nobre 1998) as well as sensory modality (see Duncan et al 1997, but also Arnell and Jolicoeur 1999). Attention can even be allocated to specific features within a modality ("feature-based attention"; Itti and Koch, 2000; Mounsell and Treue, 2006). Intriguingly, though semantically, spatial location and temporal incidence can be formulated as features analogous to color, orientation, or auditory pitch, neural evidence indicates different circuitry underlies attentional allocation along spatio-temporal and non-spatiotemporal lines (Giesbrecht et al, 2003).

Reflecting the hypothesized neural structure codified by feature-integration theory, in addition to being able to be allocated at the basal level across space, time, sensory modality or feature alone, attention also can be deployed at the level of objects, or specific combinations of those features. Work by Egly, Driver and Rafal (1994) found evidence for this "object-based" form of attention by demonstrating that when participants focused their attention to one location of an object, they were faster at detecting cues presented at another location on the same object than at detecting cues presented at a location an equal distance away on a *different* object. When an object is selected for attentional prioritization, all of its associated features receive the benefit (O'Craven et al, 1999). For a review of the myriad, subtle ways in which attention can be deployed and the neural and behavioral consequences observed, see Chun et al (2011).

*Neural Mechanisms of Attention*

Of course, to fully answer the question of how the brain prioritizes incoming information requires the exploration of more than just the consequences for behavior. With the advent of neural recording techniques, both invasively with implanted electrodes in rodents and non-human primates, along with non-invasive imaging in humans via electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), it has become possible to address the neural mechanisms underpinning attentional selection. Broadly, observed effects of attentional modulation of neural activity tend to fall under two distinct and non-mutually exclusive general mechanistic categories. The first is enhancement of neural representation through changes in neural response properties within a region, and the second is enhancement of information transfer between regions by modulating signal transfer efficacy.

*Attention changes neural response properties in sensory cortex*

In the first category, the top-down allocation of attention has a wide variety of effects on the response profiles of individual neurons at each layer of the visual hierarchy, reflecting the wide variety of dimensions along which attention can be deployed (reviewed in Carrasco, 2011; Noudoost et al 2010; Reynolds and Chelazzi, 2004). Neurons whose receptive fields are in the spotlight of spatial attention fire more than those whose receptive fields fall outside, and while the magnitude of this effect depends on contrast, it is not clear whether the pattern is more consistent with a multiplicative response gain (with higher gains at higher contrasts; Reynolds et al, 2000) or a leftward shift in the contrast-response curve (Martínez-Trujillo and Treue, 2002). Attentional shifts to a specific feature-dimension show increases in neural response (as measured by Positron Emission Tomography) in brain regions known to be sensitive to the attended feature-dimension (Corbetta et al, 1990), that has since been replicated in fMRI (Clark et al, 1997) and EEG (Torriente et al 1999). Finally, attending to a specific exemplar within a feature

space modulates neural firing rates as a function of the distance between the attended exemplar and a given neuron's preferred exemplar (defined by its tuning function), an effect known as the "feature-similarity gain principle". Specifically, those neurons that are tuned closest to the attended exemplar display increased neural firing rates and those that are tuned furthest away show a smaller increase or even a decrease in firing rate. For instance, when attending to 90˚ motion, neurons which are tuned to 90˚ increase their firing rates, while neurons which are tuned to 270˚ motion decrease their firing rates (Martínez-Trujillo and Treue, 2004).

Relatedly, when two stimuli are presented in a single neuron's receptive field, attention has the effect of biasing the competition between them such that the observed neural response is nearly identical to the attended stimulus is presented alone (Desimone, 1998). This has been observed in IT cortex (Chelazzi et al, 1998), but also in earlier ventral stream visual areas such as V2 and V4 (Luck et al, 1997), as well as in dorsal stream areas MT and MST (Recanzone et al, 1997).

Complicating the picture further is the fact that spatial receptive fields and feature tuning profiles are themselves modulated by attention. Attention to a specific location in space shifts receptive fields towards the attended location (Womelsdorf et al, 2006; Klein et al, 2014) in a manner that increases both the spatial discriminability and representational fidelity of spatial location in visual cortex (Vo, Sprague and Serences, 2017). Analogously, attending to a specific exemplar in feature-space shifts tuning profiles in that feature space towards the attended exemplar (David et al, 2008).

It is still an open question as to how, mechanistically, these phenomena (feature-gain similarity, biased competition, RF shifts, etc) arise. Reynolds and Heeger (2009) elegantly demonstrate that simply implementing a canonical neural computational principle termed

"divisive normalization" that operates at the inputs to the visual cortical hierarchy can account for all of these effects. Recently, however, Miconi and VanRullen (2016) have argued that this model cannot account for the backward progression of attentional effects (Buffalo et al, 2010), whereby effects in higher order areas occur prior to those in hierarchically lower ones. Instead, they propose a model where top-down attention acts at higher levels of visual cortex (IT and V4 in the Ventral stream, MT and MST in the dorsal stream) and then "cascades" down the hierarchy, rather than acting at the inputs of each level independently.

*Attention enhances information transfer between brain areas*

In addition to attentional modulation to neural response within a region, attention has been shown to modulate neural activity between areas in a manner consistent with enhancing the transfer of attended information. Specifically, synchronization in spiking activity between neurons encoding attended information has been proposed as a mechanism to preferentially enhance the effects of those neurons on activity in downstream areas (Salinas and Sejnowski, 2001; Tiesinga et al, 2008). Along these lines, top-down attention has been shown to locally synchronize neural activity in both the gamma-band (~40-70Hz; Saalmann et al, 2007; Gregoriou et al, 2009) and beta-band (Buschman and Miller, 2007); an effect which has been linked to improved behavior (Womelsdorf et al 2006). Additionally, neural synchrony *between* areas is also modulated by attention (Saalmann et al, 2007; Gregoriou et al, 2009), ostensibly forming a common temporal window for selectively gating attended information (Fries, 2009); and is coordinated by the pulvinar (Saalmann et al, 2012). Consistent with this hypothesis, Ruff and Cohen (2016) elegantly demonstrated increased spike transfer efficacy between visual areas as a function of attention.

Finally, attention reduces the variability of neuronal responses to a particular stimulus, an effect which is stronger for interneurons (Mitchell et al, 2007). This in turn has been shown (Cohen and Mounsell, 2009) to lead to increased reliability of neural population representations by decreasing correlated trial-by-trial fluctuations between neurons (ie correlated noise). This decorrelation has been linked to the suppression of low-frequency, alpha-band (8-15Hz) oscillatory activity in visual cortex (Fries et al, 2008; Thut et al, 2006) providing a mechanistic explanation for a phenomenon long observed in human EEG.

*Sources of Attention*

So far, this introduction has considered the *consequences* of attention, with a commensurate focus on sensory cortex (particularly visual cortex), a common site of attentional allocation. The other half of the equation, of course, are sources of top-down attention: those areas responsible for allocating attention to guide behavior. As hinted at previously, evidence from wide variety of studies points towards frontal and parietal regions as sources. A comprehensive review is beyond the scope of the current work (for that, see Baluch and Itti, 2011), but relevant points are highlighted below.

Broadly, lesions in parietal cortex often accompany the clinical presentation of spatial neglect and deficits in covert attention shifts (Posner et al, 1984). More specifically, stimulating neurons in the Lateral Intraparietal Cortex (LIP) produced either saccades or attentional shifts in a spatial cuing task (as indexed by increased reaction times to a cued hemifield) depending on the strength of the stimulation used (Cutrell and Marrocco, 2002). Interestingly, stimulation during trials where spatial attention was not cued produced a decrease in reaction times, hinting at this region's role in the salience-driven, bottom-up, exogenous capture of attention. Indeed, modern formulations of LIP, and the parietal lobe in general cast it as a priority/salience map

(Bisley and Goldberg, 2010), coordinating with other regions to guide spatial attention and plan motor actions (Gottlieb, 2007).

In frontal areas, micro-stimulation of the Frontal Eye Fields (FEF) in the macaque (the human homologue of which is thought to be near superior frontal sulcus and precentral sulcus [SFS]) produces saccades at high stimulation intensities, and covert attention shifts at lower intensities ("sub-threshold"), similar to that observed for LIP but with greater spatial specificity (Moore and Fallah, 2001). Sub-threshold stimulation of FEFs produces effects in a variety of visual areas consistent with those observed with top-down attention, including biasing of neural responses (Moore and Armstrong, 2003; Armstrong et al 2006), increasing gamma-band synchrony (Gregoriou et al, 2009), and enhancing the visual activation of retinotopically corresponding foci (Stanton et al, 1995). Findings consistent with both frontal and parietal areas acting as sources of attention have also been observed in the fMRI literature (for a review see Yantis and Serences, 2003).

*Selective Attention of Representations in Working Memory*

As discussed earlier, fundamental questions remain regarding the functions of occipital, parietal and frontal cortices in successfully completing tasks of working memory. One framework that is beginning to coalesce is, in a sense, a unification of selective attention and working memory. Under this framework, Frontal and Parietal Cortices subserve similar roles in tasks of working memory as in selective attention, namely allocating attention appropriately along specific task demands, and planning motor actions to generate successful behavioral outputs. Meanwhile, just as sensory cortex was the site of attentional action in tasks of selective attention, representing attended information at the expense of unattended information, in

working memory it serves much the same role, representing information that is no longer present in the environment (Gazzaley and Nobre 2016; Myers Stokes and Nobre, 2017).

The relationship between the mechanisms underlying shifts of attention to internal representations of information still present in the external environment ("external representations"; tasks of selective attention) and the mechanisms underlying shifts of attention to internal representations of information that is no longer present in the environment ("internal representations held in working memory"; tasks of working memory) will be the primary focus of this text. Specifically, while the effects of attentional shifts on individual neurons has been relatively well-characterized, the effects on population level representations is less well known. The second chapter of this text will address this point. Chapter Three will address the effects of various non-spatial attentional shifts on population level neural representations of items held in visual working memory, and Chapter Four will investigate the ability of attention to gate information into working memory stores as a function of task demands.

# Chapter Two:
# The Effects of Attention on External Representations

## Introduction

What mechanisms underlie the prioritization of neural sensory representations to effectively guide behavior? By selectively processing sensory information which is most likely to be relevant for determining appropriate motor output, the human brain maximizes the use of finite cognitive resources. Models of selective attention (Treisman, 1964; Desimone 1998; Fries, 2009; Reynolds and Heeger, 2009; Miconi and VanRullen, 2016), have implemented this prioritization via distinct (though sometimes non-mutually exclusive) mechanisms, and each makes specific predictions about the consequences of this prioritization for neural activity in sensory cortex, both at the single neuron and population level.

For a variety of reasons, some of which are methodological, the effects of top-down selective attention on individual neurons in visual cortex have been markedly better-characterized. When attention is allocated spatially, neurons whose receptive fields fall within the focus of attention show increased firing rates to stimuli presented in their receptive fields, though how this increased sensitivity varies with stimulus contrast is still a matter of debate (Reynolds et al, 2000; Martínez-Trujillo and Treue, 2002). Attending to a specific feature dimension, independent of space, causes increased neural activity in regions known to possess neurons whose response varies as a function of that feature dimension (e.g. attending to orientation is associated with increased activity in V1; Corbetta et al, 1990; Clark et al, 1997; Torriente et al, 1999), whereas attending to a specific exemplar within a feature space (such as a

0 degree oriented bar) increases firing rates for neurons which respond maximally to that exemplar and suppresses firing rates of neurons minimally responsive to that exemplar (the so-called “feature-gain similarity principle”; Martínez-Trujillo and Treue, 2004). Finally, when two different stimuli are simultaneously presented within a neuron’s receptive field in the absence of attentional prioritization, that neuron’s response is intermediate between those for each stimulus presented alone. However, once attention prioritizes one of the stimuli, that neuron’s response to both presented simultaneously becomes more like its response to the attended stimulus alone, a phenomenon known as “biased competition”.

In addition to those effects of attention that modulate how specific neurons in visual cortex respond to stimuli, attention also increases the influence those responses have on downstream neurons. Attention increases local synchrony in spiking within task-relevant visual areas as indexed by increased spike-field coherence in both the gamma- (~40-70Hz; Saalmann et al, 2007; Gregoriou et al, 2009) and beta-band (Buschman and Miller, 2007); an effect which presumably reflects the coordinated aggregation of relevant neurons into a stimulus-specific representational ensemble. In support of this idea, the magnitude of these effects predict behavioral success (Womelsdorf et al 2006). Additionally, these same studies also observed an attention-dependent increase in spike-field coherence in the gamma-band between relevant visual processing areas and downstream regions, consistent with conditions predicted to maximize signal transfer efficacy by theoretical models (Fries, 2009). Critically, however, information gating models of attention predict not just an indiscriminant increase in connectivity between areas, but instead predict an increase in the transduction of signal and a decrease in the transduction of noise. Ruff and Cohen (2016) elegantly demonstrated this by showing that attention decreases stimulus-independent response correlations between visual areas (i.e.

correlated noise), while maintaining increased signal transfer efficacy (as indexed by the ability of microstimulation of primary visual cortex to drive spiking in the downstream Middle Temporal Region). This built upon previous work by Briggs et al (2013), who established increased signal transfer efficacy for thalamocortical projections.

However, while the manifestations of top-down attentional prioritization have been examined rather extensively at the single-, (or even paired-) unit level, population-level effects remain less well-characterized. Population-level attentional effects are of particular importance in interrogating object- or feature- based attention, where the putative substrate of attentional action is a specific representational ensemble population (Valdés-Sosa, 2014). Additionally, distributed neural codes are thought to become more prevalent as information travels up the cortical hierarchy (Yuste, 2011). One challenge facing population-level analyses is the question of how to measure the information content.

Multivariate approaches (Haxby et al 2001; vanBergen et al, 2015; Rigotti et al, 2013; Sprague et al, 2015; Huth et al 2016) take advantage of insights gleaned from the field of machine learning (Cortes and Vapnik, 1995). Namely, data recorded from a population of features, e.g. neuronal spike rates, voxel blood-oxygen-level-dependent (BOLD) signals, or electroencephalographic (EEG) electrode voltages, can be represented as points in high-dimensional feature space, which can then be regressed against the particular stimuli that gave rise to those patterns to uncover stimulus-driven patterns of activity. Importantly, because these methods preserve the high-dimensionality of the data, they are sensitive to population-level information in a way that considering each unit independently is not (Rigotti et al, 2013).

In non-human primates, an increasingly common practice is to perform repeated single-unit analyses, and aggregate the results together to make inferences about the population

(Martinez-Trujillo and Treue, 2004). This method has been used by researchers in the Desimone laboratory (Zhang et al., 2011) to extend their model of biased competition beyond a description of the modulatory effects of attention on the activity of single-units, into the realm of population-level neural representations. Researchers have begun to address similar questions surrounding population-level manifestations of selective attention in humans (Hou and Liu, 2012; Woolgar et al, 2015), though these mechanisms remain considerably underexplored.

In the present work, we employed multivariate pattern classification of neural activity assessed via functional magnetic resonance imaging (fMRI), collected while participants performed a cued attention task. Our results support our prediction that the effect of attention is to bias the strength of representations in visual cortex in favor of the cued stimulus. Furthermore, utilizing a multivariate metric of connectivity (Informational Connectivity; Coutanche and Thompson-Schill, 2013) we found evidence implicating frontal cortical areas as a putative source of this top-down biasing signal.

## Methods

*Participants*

10 healthy subjects (6 female, mean age = 22.4, SD = 5.08 years) with normal or corrected to normal vision participated in the study. All participants provided written informed consent and were monetarily rewarded. Participants were screened to rule out any neurological and psychiatric disorders or incompatibility with magnet resonance. The study was approved by the UW-Madison Health Sciences Institutional Review Board.

*Task*

Prior to the experiment all participants participated in training blocks in order to get acquainted with the behavioral task. The task consisted of two experiments (figure 1): Experiment 1, which we will refer to as the "training phase", was a one-item change detection task and Experiment 2: "the testing phase", was a visual search task. Participants underwent fMRI while performing both experiments. The order of Experiments was counterbalanced. Ten MRI acquisition runs were collected for each subject: 6 runs for Experiment 1 and 4 for Experiment 2.

Experiment 1, a change detection task, was broken into 6 runs of 16 trials, each of which lasted for 18 seconds. The trial began with a white fixation cross-presented in the middle of screen on a gray background for approximately 1 sec. A stimulus target (either a face, a doughnut, an abacus or a fixation cross) was foveally presented and began flickering with a cycle of 750 msec on/250 msec off (Brady et al, 2008). Unpredictably on 0,1, or 2 occasions the stimulus changed to a different image of the same object (i.e., a full doughnut, a half doughnut, a doughnut with a bite out etc…). This lasted 15 seconds, and participants were requested to report the number of "state changes" that the target underwent during the final 2 seconds of a trial. Trials alternated between the three images in addition to an equal number of fixation blocks. Each stimulus category was presented a total of 24 times across all runs (pseudorandomly ordered such that an equal number of each category appeared in each run with the specific order within a run randomized). Stimuli were selected from Brady et al (2008) to span a range of categories to optimize decoding. Figure 1A shows all 6 stimuli used.

Experiment 2 was a visual search task and consisted of 4 runs of 12 trials each (Figure 1B). Each trial began with the display of two potential search targets, one centered 5 degrees of

visual angle to the left of fixation and the other centered 5 degrees of visual angle to the right of fixation. Image size was 3 degrees of visual angle by 3 degrees of visual angle. These two potential search targets were displayed for 500msec (drawn from the 3 different stimulus types: face, doughnut or abacus), followed by a 7sec delay. Subsequently, in the first half of the trial, a "search array+ cue" comprising the same items with one designated a target (cued with a red box outlining the stimulus with a width of 0.5 degrees visual angle) and the other a non-target (no red outline) appeared for 500msec. When the cue disappeared, both stimuli started flickering with a cycle of 750 msec on/250 msec off for 7 sec. Subjects were instructed to report the number of "state changes" that the target underwent. The second half of the trial was identical to the first half with the exception that the other stimulus was now cued on 50% of the trials. Trials in which the same stimulus was cued twice are designated "stay" trials, and trials where the target changed midway through the trial are designated "switch" trials. There were 48 trials in total.

*Eye-tracking*

In order to ensure that participants fixated on the cross at all times, we recorded their eye movements using electrooculography. (BrainVision recording suite, Brain Products, http://www.brainproducts.com/index.php). A calibration run before scanning, where participants were instructed to make 10 saccades to each hemi-field was used determine the average EOG voltage deflection during a saccade for each participant. Participants were excluded from analysis if, during the visual search task, 10% or more trials showed EOG deflections consistent with saccades to either hemifield. No participants met this threshold, and thus all were included in the final analysis.

*fMRI Data Acquisition*

Whole brain images were acquired with the 3 T MRI scanner (Discovery MR750; GE Healthcare) at the Lane Neuroimaging Laboratory at the University of Wisconsin-Madison. High- resolution T1-weighted images were acquired for all subjects with an FSPGR sequence (8.132 ms time repetition (TR), 3.18 ms time echo (TE), 12° flip angle, 156 axial slices, 256 × 256 in- plane, 1.0 mm isotropic). Blood oxygen level-dependent (BOLD)- sensitive data were acquired using a gradient-echo, echoplanar sequence (2 s TR, 25 ms TE) within a 64 × 64 matrix (39 sagittal slices, 3.5mm isotropic).

*fMRI data analysis*

fMRI data analysis was performed using Analysis of Functional NeuroImages (AFNI) software package (http://afni.nimh.nih.gov; Cox, 1996). We excluded the first 3 volumes of each run to account for EPI-onset field inhomogeneity. All volumes were spatially realigned to the final volume of the final functional run using rigid-body realignment. The processing pipeline included slice time correction, de-trending of low-frequency signal drift, and conversion to percent signal change. Finally, spatial smoothing with a 4-mm FWHM Gaussian kernel was performed prior to conducting General Linear Model (GLM) analyses.

*Generation of ROIs*

In order to generate appropriate regions of interest (ROIs) for the MVPA, we first generated anatomical masks which broadly covered three separate regions of cortex: visual cortex (covering parts of both occipital and temporal cortex), parietal and frontal. We then selected a

subset of voxels from each anatomical mask which showed the greatest sensitivity to our behavioral stimuli, as indexed by a functional GLM.

*Anatomical ROI Generation*

Anatomical ROIs were generated using the Talraich anatomical atlas (TTatlas; https://sscc.nimh.nih.gov/afni/doc/misc/afni_ttatlas/). Briefly, coordinates for relevant gyri in the TTatlas were used to generate masks for each gyrus, which were then warped into an individual's original space, and aggregated based on broad anatomical location to create a regional mask. For instance, to create the "frontal" anatomical mask, TTatlas masks were generated of the inferior frontal gyrus, middle frontal gyrus, superior frontal gyrus, medial frontal gyrus and precentral gyrus, which were then warped into subject space and aggregated.

*Functional ROI Generation*

We performed a general linear model (GLM) analysis for each subject on the data from the one-item change-detection task in order to identify brain ROIs for the MVPA. A single regressor was included for stimulus onset along with covariates to control for motion and block-specific effects. Stimulus-onset was modeled as a boxcar of 8 seconds covering the full time each stimulus was present and flickering on the screen. All were convolved with a canonical hemodynamic response function. Each of these independent regressors was entered into a modified GLM for analysis using AFNI. In order to capture voxels which were likely to be useful for subsequent MVPA, for each subject, for each area (occipital/temporal, parietal and frontal) we extracted the top 400 voxels with the highest positive f-statistic associated with overall model fit. To generate hemisphere-specific masks, we also performed this unilaterally,

generating left hemisphere masks by selecting the top 200 voxels from each left-hemisphere broad anatomical area and generating right hemisphere masks by selecting selecting the top 200 voxels from each right-hemisphere broad anatomical area. A similar approach to ROI generation has been used in prior studies in our laboratory (Riggall and Postle 2012; Emrich et al. 2013) and has the advantage of accounting for individual differences in task-relevant neural activity.

*Pattern classification analyses*

Classification procedure was performed using the Princeton Multi-Voxel Pattern Analysis toolbox (www.pni.princeton.edu/mvpa/) and custom scripts in MATLAB. All MVPA analyses were conducted on neither smoothed nor time-shifted data. In order to examine the neural representations associated with visual selective attention, we trained pattern classifiers in experiment 1 to classify each pattern of the three different stimulus types during the training phase. We first validated classifier sensitivity using a leave-one-run-out approach. Classification was accomplished using L2-regularized logistic regression with a lambda penalty term of 25. All neural data were z-scored across trials, within runs, before computing MVPA.

We first used k-fold cross-validation on the data from the training phase experiment. We processed fMRI data for the 18 second trial blocked design. Each functional volume was acquired over a 2-s TR. We had three exemplars associated with an array of features corresponding to BOLD signal of every stimulus type. Each classifier was trained on data from time points associated with the middle to late portions of each trial (e.g., TRs 4-8 of 9 total) for 5 of the 6 runs, and then tested on the analogous time points in the left-out run. Implementing k-fold cross-validation (k = 6), we trained a classifier 6 times, leaving out a different run each time, and then averaged across folds. The classifier accuracy was computed by performing a one-

sample, one-tailed t-test comparing accuracy to chance performance (33%) to evaluate statistical significance. Once validated, we applied a classifier trained on data from Experiment 1 to all time-points in Experiment 2, sorting the evidence outputs by attentional state

*Decoding from Experiment 2, having trained on Experiment 1*

The classifiers were trained on the patterns extracted from the one-item change detection task and were then applied to data from each time-point in the visual search task with retro-cues. A measure of pattern similarity was computed between the voxel patterns of each time-point in the testing set and the learned pattern for each category extracted from the training phase. Using logistic regression, each category's pattern similarity score was then converted into a value between 0 and 1, analogous to an estimate of probability that the observed testing pattern was generated by that category. This measure of classifier "evidence" for each of the three stimuli at each time point in the trial was sorted according to whether that category was initially cued, initially un-cued or absent on each trial and then averaged together, generating an averaged trial time-course. Because the category that was cued on the second half of the trial changed for "switch" trials and stayed the same in "stay" trials, these trial types were averaged, and are presented, separately. Statistical significance of attended and unattended evidences for each time-point were computed by a subject-level permutation analysis between each condition and the baseline evidence for each subject.

*Informational Connectivity*

The estimates of classifier evidence from the second delay period (following the first retrocue) were also used to compute a measure of Informational Connectivity (IC) between the ROIs (Coutanche and Thompson-Schill, 2013) using a modified version of the freely available online

toolbox for matlab (http://lrdc.pitt.edu/coutanche/informationalconnectivity). Analogous to functional connectivity, informational connectivity correlates the experimental time-course of the evidence outputs from a classifier in two separate regions (rather than correlating the raw BOLD signal itself). This difference is important because it enables the case where two areas directly sharing information of interest to be distinguished from cases where two areas' activity is correlated but unrelated to the information of interest. In other words, it enables correlated signals to be distinguished from correlated noise. Lastly, because the cued stimulus is always from a different category than both the un-cued stimulus and the stimulus not present on that trial, informational connectivity can be computed separately for both attended information and unattended information. Specifically, classifier evidences for each stimulus category (Donut, Face, Abacus; relabeled to Attended, Unattended, Absent as determined by each trial) were obtained using the procedure outlined previously (training on Experiment 1 and testing on Experiment 2) for each area. This was done separately for each time-point for each trial, giving rise to an evidence "time course" in each area for each stimulus category (attended, unattended and absent) that was 144 time-points long (48 trials x 3 time-points in the second delay period of each trial following the first retro-cue). Because evidence for the absent category represents an empirical "baseline" by which to gauge classifier noise, as a normalization step, the absent evidence time-course was then subtracted from both the Attended and Unattended time-courses to yield an Attended discriminability time-course and Unattended discriminability time-course, respectively. Attended discriminability time-courses in one ROI are then correlated (using a non-parametric Spearman's rank method) with Attended discriminability time-courses in a different ROI to generate a measure of IC between those ROIs. The same is then done for Unattended discriminability time-courses.

## Results

*Behavior*

Participants attained a high degree of accuracy on both Experiment 1 (92.3%, SD=3.4) and Experiment 2 (90.5%, SD=4.1). Taken together, these results indicate that participants performed well on the task and complied with instructions. They also indicate that all three stimulus categories were of comparable difficulty, and that performing sequential visual search tasks did not increase the difficulty of the second search relative to the first.

*fMRI Classifier Training (Experiment 1)*

Across all participants, leave-one-run-out cross-validation analyses of brain data from Experiment 1 demonstrated extremely reliable classification of stimulus category in all ROIs (Fig. 2). Despite significant variance across ROIs, above chance decoding in every ROI motivated the use of this classifier, trained on Experiment 1 data, to decode brain activity related to each stimulus category in Experiment 2, where multiple stimuli were presented simultaneously under different attentional prioritization states.

*Bilateral fMRI Classifier Decoding Under Shifting States of Attentional Prioritization*

Group averaged results reveal successful classification of stimulus category in all three bilateral ROIs (Fig. 3). Early in the time course for Experiment 2, when both stimuli are presented on the screen but neither is yet attentionally-prioritized (~ 4 seconds), classifier evidences for both presented categories (red and blue traces) rise above evidence for the third category (black trace), which is not present and thus can be thought of as an empirical baseline.

In Parietal and Frontal ROIs (Fig. 3A and 3B), once one of the categories is cued (red trace, ~ 8 seconds), evidence in all three ROIs for that category becomes greatly elevated above baseline, while evidence for the un-cued category (blue trace) falls to baseline. On stay trials, where the same category is cued again for a second visual search, the pattern is the same. However, on switch trials, where the initially un-cued item is now cued for visual search (blue trace, ~ 16 seconds), classifier evidence for that category rises above baseline and the initially cued, now un-cued category (red trace) falls to baseline. The observed patterns of classifier evidence are consistent with an account whereby Frontal and Parietal ROIs represent only attended information, as at no point in the trial is evidence for an un-cued category significantly elevated above baseline. In contrast, for the Visual Cortex ROI (Fig. 3C), evidence for the un-cued category largely remains elevated above baseline, though at a significantly lower level than evidence for the cued category. Recall that prior to the cue onset, category evidences for both presented items were equally elevated above baseline, demonstrating that attentional prioritization has the effect of enhancing prioritized representations and/or suppressing un-prioritized ones.

*Unilateral fMRI Classifier Decoding Under Shifting States of Attentional Prioritization*

Experiment 2 presented both items in separate visual hemi-fields, in part to take advantage of a known property of early visual cortex: receptive fields in each hemisphere are isolated to contralateral space. To assess the impact of shifts in spatial attention on population representations in each hemifield, a classifier was trained to distinguish stimulus category using unilateral ROI voxel data generated during Experiment 1, where all stimuli were presented foveally, and then tested on voxel data from the same unilateral ROI, but generated from

Experiment 2, where stimuli were presented peripherally (one in each visual hemi-field). Figure 4 shows the results for each general region (frontal, parietal and visual), plotted separately for the ROI hemisphere *contralateral* to the initially cued stimulus (fig 4A) and for the ROI hemisphere *ipsilateral* to the initially cued stimulus (fig 4B). Note that because stimulus location and cue location were counterbalanced, each plot contains equal parts left and right hemispheres, though which hemisphere was contralateral changed for each trial (see methods). Importantly, consistent with previous literature, neither Frontal nor Parietal ROIs displayed an effect of laterality, in that classifier evidence was only ever observed above baseline levels for the cued category, regardless of hemisphere (Figs. 4 and 5).

Again, however, visual cortex (Fig. 6) displayed a markedly different pattern. Early in the time-course, before the cue onset (at ~ 4 seconds), the category of the stimulus in the contralateral hemifield can be fully distinguished from the baseline category (red trace in the contralateral cue plots, blue trace in the ipsilateral cue plots). Evidence for the ipsilateral stimulus category, in contrast (blue trace in the contralateral cue plots, red trace in the ipsilateral cue plots), is either weakly above baseline, or indistinguishable from baseline. This indicates that before the cue onset, both hemispheres are behaving similarly, which is expected, given that the cue has not yet spatially prioritized one hemi-field over the other. Once the first cue appears, however, differences emerge between the hemisphere ipsilateral to the cue and contralateral to the cue. For the hemisphere contralateral to the cue, the evidence for the cued category (red trace) is greatly elevated above baseline, while the evidence for the uncued category (blue trace) falls all the way to baseline. In contrast, for the hemisphere ipsilateral to the cue, both the cued and uncued categories show evidence above-baseline of nearly equal strength. For both stay (fig 6, left plots) and switch (fig 6, right plots) trials, this pattern persists for the second cue, as well.

Note that in switch trials, which hemisphere is contralateral to the cue naturally switches along with the second cue, so that the traces in the plot labeled "contralateral cue" (so named because the data is derived from hemispheres contralateral to the first cue) are actually hemispheres *ipsilateral* to the second cue, and vice versa.

*Connectivity between ROIs differs for Attended and Unattended information*

To examine whether attention influences the extent to which stimulus category specific information is shared between ROIs, an information connectivity analysis (fig. 7) was conducted (see methods). For all connectivity analyses in figure 7, bilateral frontal and parietal ROIs are used, because the unilateral masks showed no effects of laterality, but visual cortex is split into hemispheres contralateral to the cue and ipsilateral to the cue as in figure 6. A version of the analysis performed with all unilateral masks showed qualitatively similar results (and did not show differences in frontal or parietal ROIs as a function of laterality), so the first analysis is presented here. Figure 7A shows the connectivity strengths for attended (cued) information between each ROI. The Frontal ROI displays significant positive informational connectivity between itself and the Parietal ROI for attended information, but interestingly, no significant connectivity is observed between the Frontal ROI and either Visual ROI. The Parietal ROI, however, shows a positive informational connectivity with both the contralateral and ipsilateral visual ROI. Strong positive connectivity also exists between contralateral and ipsilateral visual ROIs. For unattended (uncued) information (Fig 7B), the picture is mostly the same: both Parietal-Frontal ROI connectivity and Ipsilateral-Contralateral Visual ROI connectivity remain strong, as does Parietal ROI connectivity with both Ipsilateral and Contralateral Visual ROIs. Crucially, however, the Frontal ROI is now significantly *negatively* connected with the

Contralateral Visual ROI. The modulatory effect of attention on connectivity is thus best demonstrated by figure 7C, which plots the *difference* in informational connectivity as a function of attentional state. Only the Frontal-Contralateral Visual ROI connectivity is significantly modulated by attention, such that the connectivity is greater with attention.

## Discussion

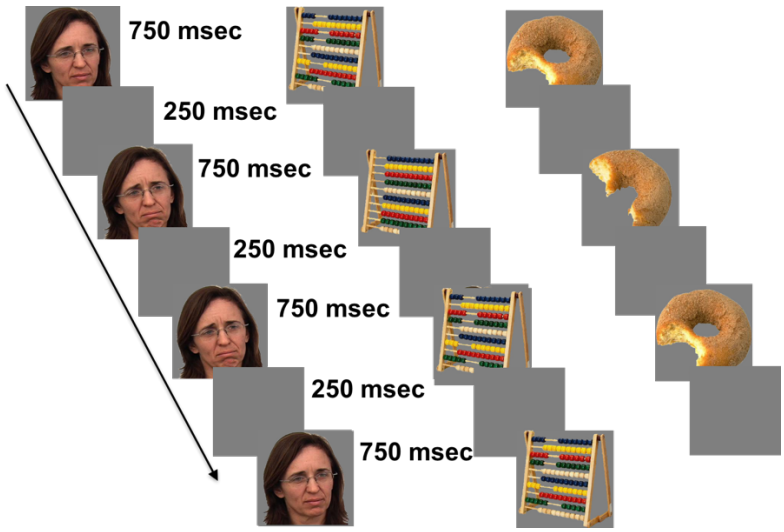
A fundamental concept at the heart of selective attention is the principle of biased competition. Classically (Desimone, 1998) at the level of single neurons in visual cortex, attention has the effect of making neuron's response to a pair of stimuli appear similar to that of the attended stimulus presented alone. More generally, attention is hypothesized to bias the competition of external stimuli for visual representation. We used multivariate analysis methods to quantify the effects of selective attention on population level neural representations of external stimuli in a visual search task. We observed that attention selectively enhanced the representational fidelity of the attended stimulus in both hemispheres of visual cortex and suppressed that of the unattended stimulus. This effect was most pronounced in the hemisphere contralateral to the attended stimulus, where the unattended representation became indistinguishable from baseline. In contrast, Parietal and Frontal Cortex showed above baseline evidence of representation only for the attended stimulus category, regardless of hemi-field. This is consistent with two non-mutually exclusive interpretations. One hypothesized function of

attention is the enhancement of communication between visual processing areas and downstream areas underlying executive functions such as motor planning. Therefore, one interpretation of the pattern of representational fidelity observed in frontal and parietal areas is that it reflects the "consequences" of attentional selection. Attention biases the competition between stimuli in visual cortex, and the one that wins gets passed downstream, where it becomes represented in frontal and parietal areas. Evidence in support of this view comes from Logethetis et al (2001) who showed that the fMRI BOLD signal is correlated most strongly with local field potentials (LFPs), suggesting that BOLD tracks the collective inputs to an area. A second interpretation is that this activity reflects the source of top-down attentional control. Activity in subregions of both areas is well known to drive shifts in spatial as well as object-based attention (Baluch and Itti, 2011).

Our multivariate connectivity analysis, termed informational connectivity or IC (Countanche and Thompson-Schill, 2013), revealed evidence for a little bit of both interpretations. Informational Connectivity between parietal cortex and both occipitotemporal (visual) hemispheres was significant and positive for both attended AND unattended information. This, in essence, indicates that the more visual cortex represented information of any kind, the more parietal cortex also represented this information. This pattern is consistent with parietal cortex as a downstream receiver of representations sent from visual cortex. In contrast, IC between frontal and visual cortex was significant only for unattended information, and was significantly negative. This demonstrates that the patterns observed for this task in the Frontal ROI represent a source of attentional control, implicating frontal areas as a source of suppression for representations of unattended information in visual cortex. In a recent study, Baldauf and Desimone (2014) found evidence for increased gamma synchrony between the

Inferior Frontal Junction (IFJ) and either the fusiform face area (FFA) or parahippocampal place area (PPA) in an object-based attention task, depending on whether faces or houses was attended, respectively. They found that gamma phase in the IFJ led that in the FFA or PPA, implicating the IFJ as a driver of attention-mediated sensory enhancement in these areas.

A




B

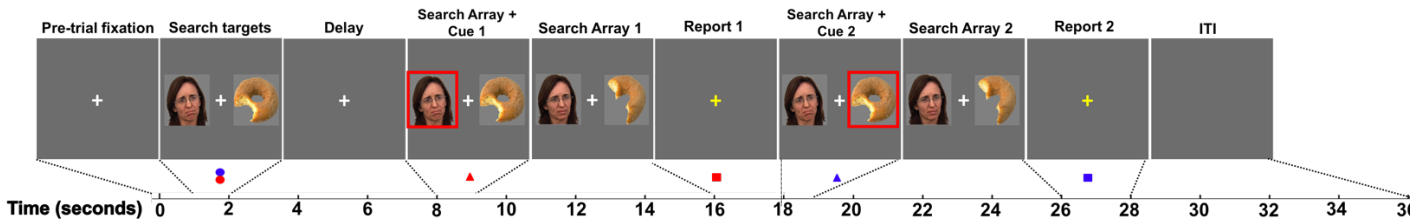



***Figure 1 – Task Diagram***

Participants completed two separate tasks, a one-item change detection task (A), and a Visual Search Task (B) while functional imaging was collected. In the first task, one of three possible stimuli appeared foveally and flashed for 16 seconds in a cycle of 750ms on followed by 750ms off. On each reappearance, the stimuli could switch between one of two distinct states (eg smiling or frowning face, organized or disheveled abacus, donut with a large or small bite removed). Participants counted the number of times the stimulus changed on a trial which could occur 0-3 times. Each block consisted of 4 trials for each stimulus category, plus 4 fixation trials, where participants were instructed to maintain fixation. In Task 2 (B), participants were instructed to fixate centrally while two of the three possible stimulus categories appeared simultaneously, one in each hemi-field, then disappeared. This informed participants which two of the three possible categories would be relevant on that trial. After a 6 second delay, both stimuli reappeared, along with a cue (red square) that indicated which stimulus was to be covertly monitored for changes. Both stimuli then began to flash as in the one-item change detection task (A), but each stimulus could change independently. After 8 seconds of this, participants were instructed to report the number of times the cued stimulus changed (again 0-3 times). Immediately following this probe, a second cue appeared that could either cue the same item again (a "stay" trial) or cue the item that was previously uncued (a "switch" trial). After another 8 seconds of flickering, participants again reported the number of times the newly cued item changed. Colored shapes in (B) symbolically indicate the phases of each trial and will be used in subsequent figures to aid in interpretation of timings.

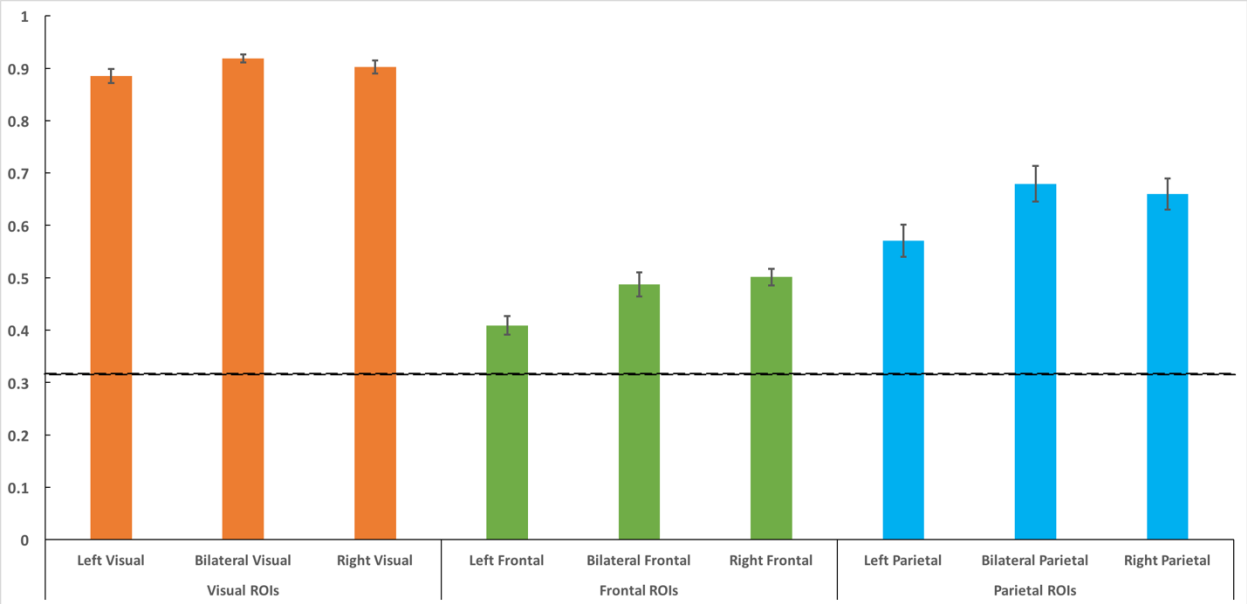



***Figure 2 – Experiment 1 Decoding Results***

Accuracies for each ROI, grouped by region. Classification was performed using a leave-one-run-out k-fold cross validation procedure (k=6, the number of runs). Data was used for time-points 4-8 (of 9 total per trial). The first 3 time-points of each trial and the final time-point were excluded to avoid probe-evoked responses. Fixation trials were excluded from classification, making chance performance 33% (black dashed line). All error bars SEM across subjects. Decoding was significantly above chance in all areas, and was strongest for Visual (Occipital/Temporal) ROIs, followed by Parietal, and lastly by Frontal.

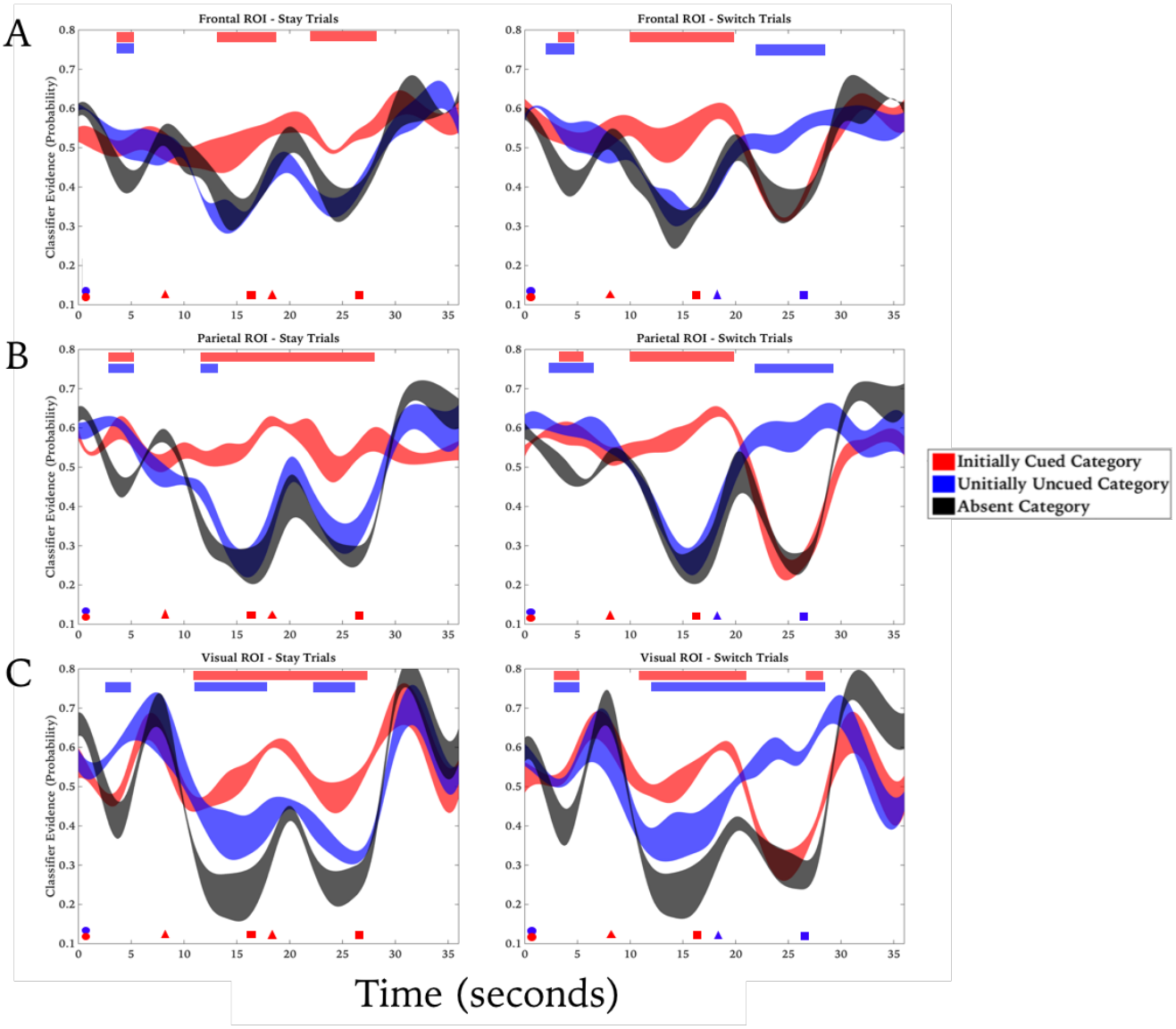



***Figure 3 – Decoding in Bilateral ROIs, Train Experiment 1-Test Experiment 2***

Trial and Subject Averaged Evidence Time-courses for the dual-retrocue visual search task. Evidences calculated from the output of a classifier trained on category in Experiment 1. Stimulus category counterbalanced across both hemi-field and cue condition, meaning each of the three traces contains evidence outputs for all 3 categories averaged together. Colored symbols at bottom of plots indicate sample, cue, and probe onsets (see task diagram for specifics). Ribbons display classifier evidence, interpolated across 19 time-points in each trial, for the cued category (red traces), uncued category (blue traces) and absent category (black traces) across trials, with the center representing the mean value and the edges denoting +/- 1 SEM. Significance bars at the top of the plot indicate those time-points for which evidence of the same-colored trace was significantly elevated above the Absent trace. Stay trials are plotted on the left and Switch trials plotted on the right, separately for Frontal (A), Parietal (B), and Visual (C) Bilateral ROIs. Evidence for the initially Cued Category is well above baseline (the "Absent" trace) in all three ROIs following Retrocue #1 (red triangle), and the same thing holds for the secondarily cued category (red trace in left plots, blue trace in right plots) following Retrocue #2 (second triangle, onset @18 seconds). In contrast, the uncued category shows significant evidence above baseline only in Visual ROIs, and this evidence level is intermediate between baseline and the evidence for the cued category. Just as with figure 2, decoding is strongest in Visual, Intermediate in Parietal, and Weak (though still significant) in Frontal ROIs.

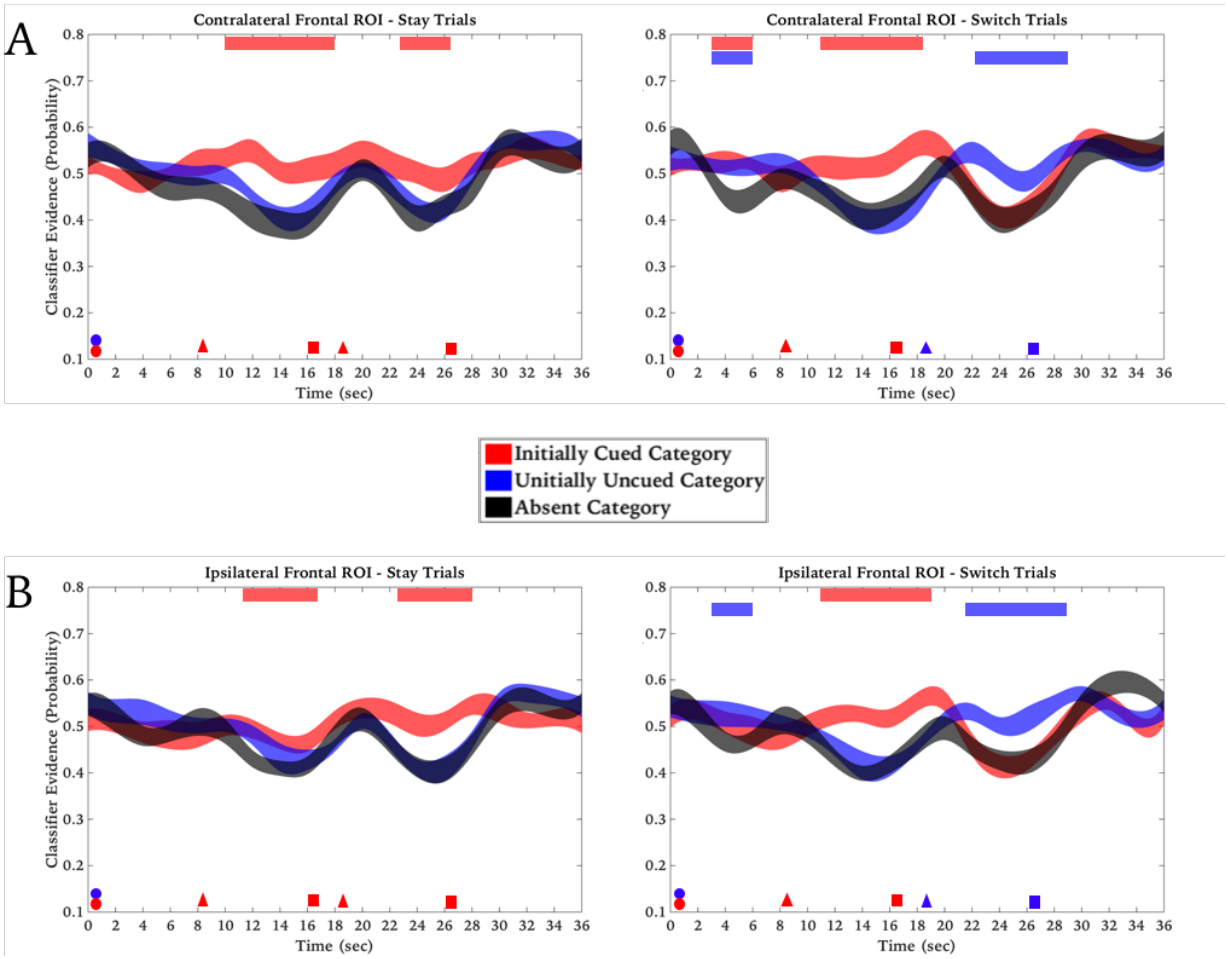



***Figure 4 – Decoding in Unilateral Frontal ROIs, Train Experiment 1-Test Experiment 2***
Conventions same as in Figure 3. All plots from Frontal ROIs. (A) Shows data from the contralateral ROI: i.e. the left-hemisphere Frontal ROI when the cue initially appears on the right side of the screen averaged with data from the right-hemisphere Frontal ROI on trials where the cue initially appears on the left side of the screen. Trials are broken down further by whether the second cue "stays" on the same side or "switches" (left and right columns, respectively). (B) Shows the analogous plots for ipsilateral Frontal ROIs (left hemisphere on cue left trials, averaged with right hemisphere evidences on cue right trials). Notably, just as in figures 2 and 3, decoding in Frontal ROIs is weak but still above chance. As with the bilateral ROI, unilateral Frontal ROIs seem to possess above baseline evidence only for cued stimuli, regardless of laterality.

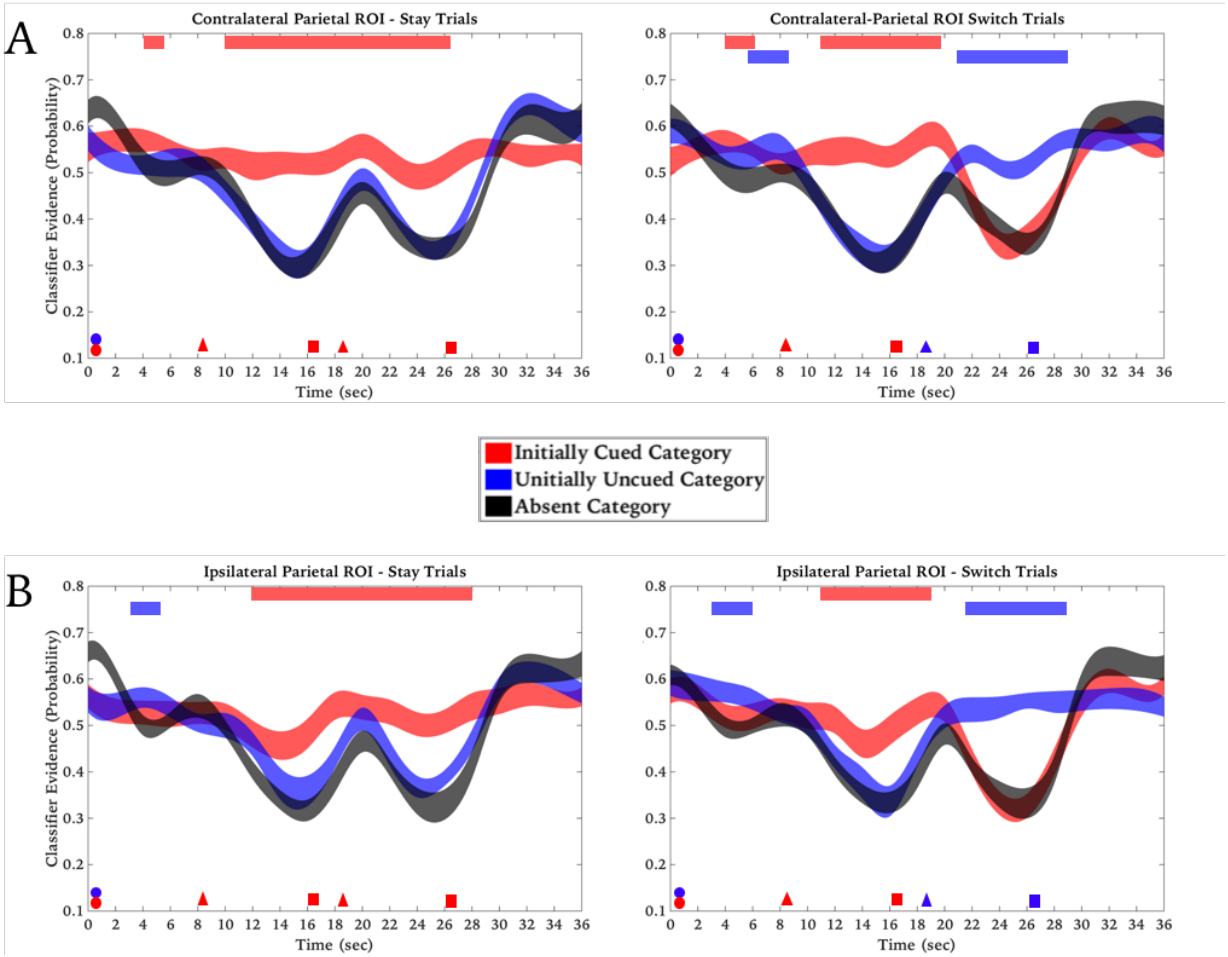



***Figure 5 – Decoding in Unilateral Parietal ROIs, Train Experiment 1-Test Experiment 2***

Conventions are the same as in figure 4. Parietal displays a pattern consistent with that observed in frontal ROIs (figure 4), namely that regardless of laterality, the cued stimulus (red trace throughout for left plots, red trace early followed by blue trace late for right plots) is the only trace with consistent above baseline evidence. Overall, evidence levels sit in between Frontal and Visual ROIs.

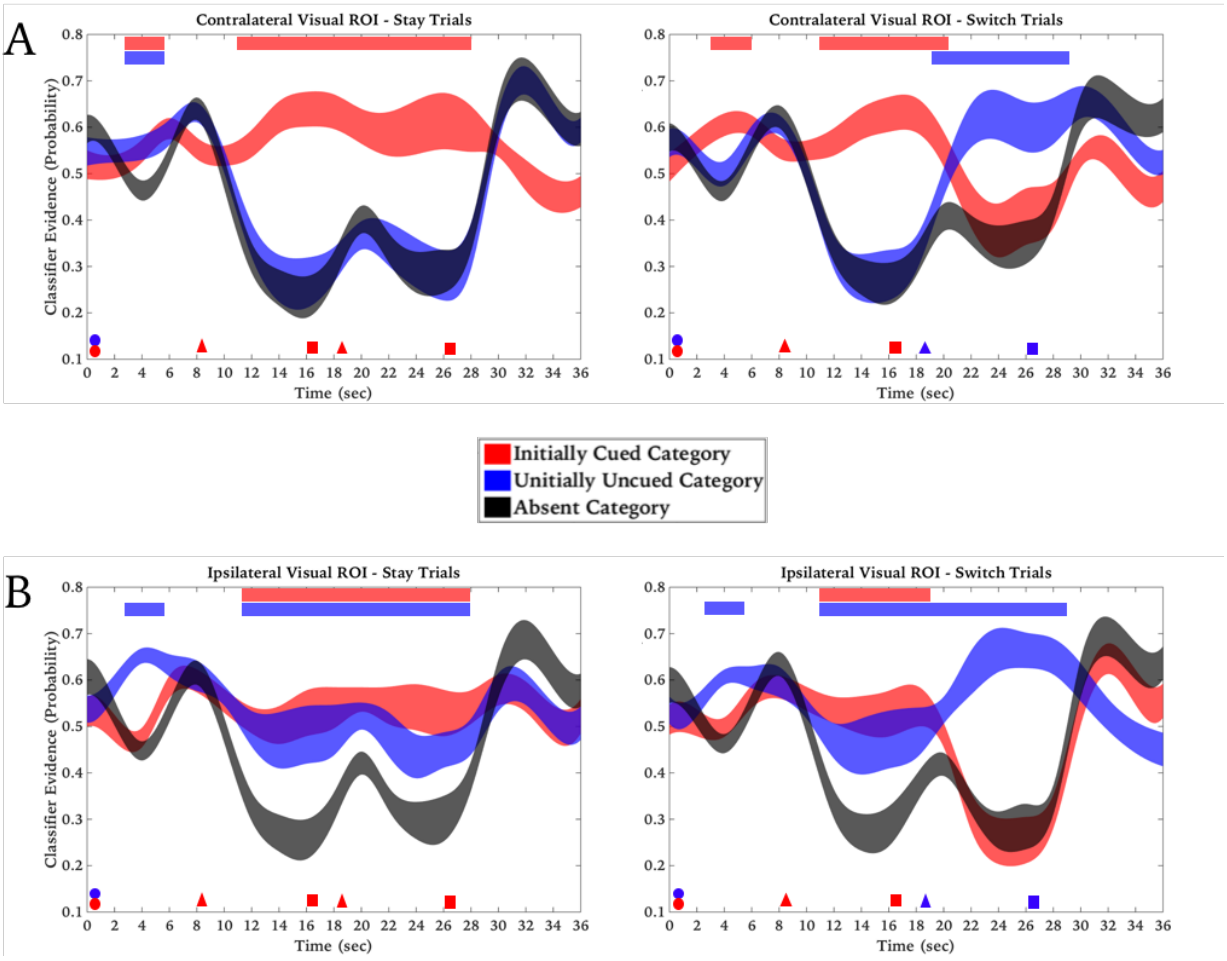



***Figure 6 – Decoding in Unilateral Visual ROIs, Train Experiment 1-Test Experiment 2***

Same conventions as in figures 4 and 5. Unlike with the Parietal and Frontal ROIs, laterality interacts with cue location. At trial onset, before the first cue (<8s), the category on the screen contralateral to the respective hemifield (red trace for top two plots, blue trace for bottom two plots) shows significant, but weak, evidence above baseline. The category on the screen ipsilateral to the unilateral visual ROI (blue trace on top plots, red trace on bottom plots) shows very weak, and only occasionally significant evidence above baseline. With cue onset, the cued stimulus shows significant above base-line evidence in both hemispheres (red trace in all plots, between 8-16 seconds). In Visual cortex Ipsilateral to the cue (bottom plots), the contralateral (uncued) stimulus also shows above baseline evidence, a pattern which does not hold for visual cortex contralateral to the cue. The same general pattern holds following Retrocue #2.

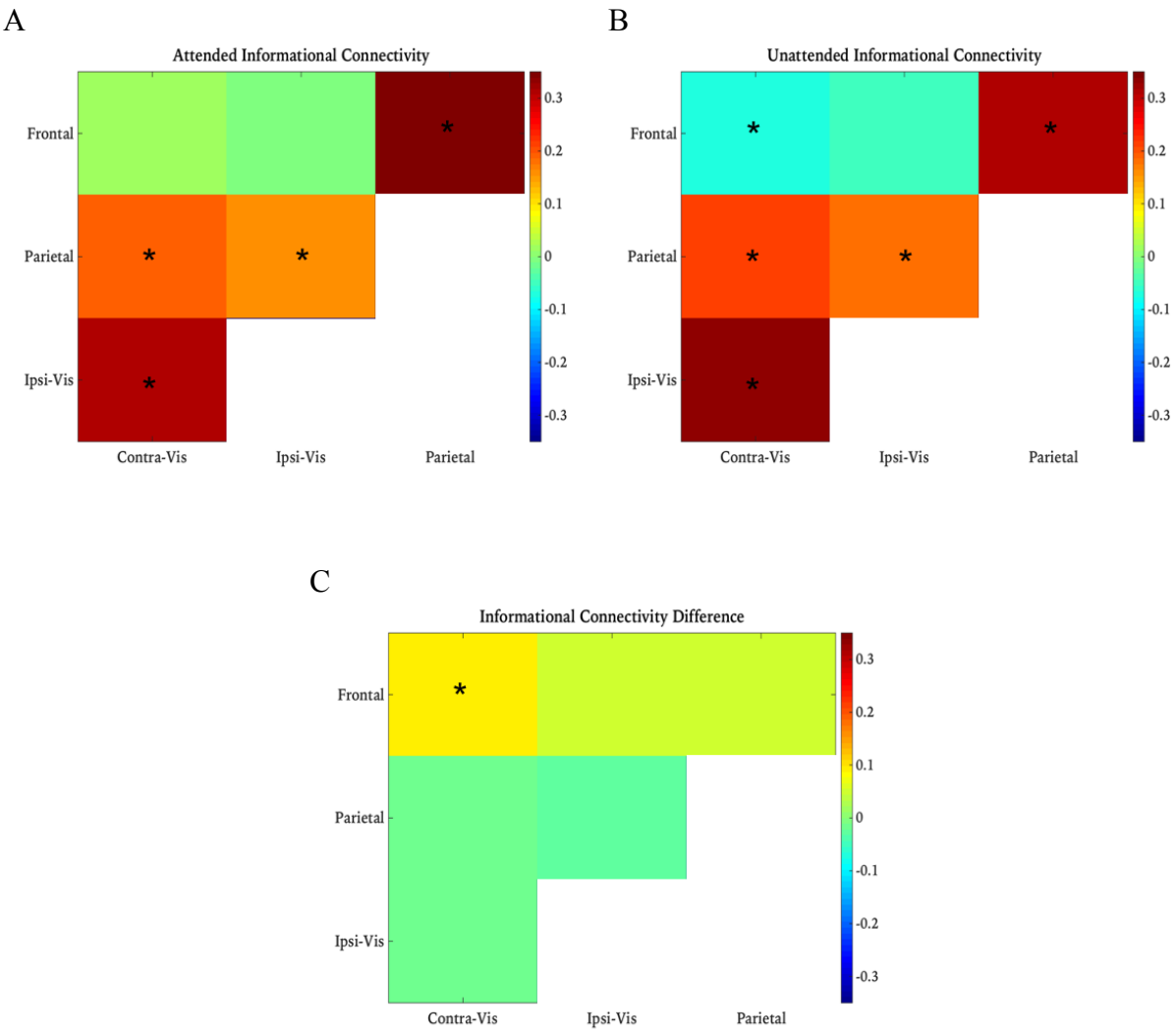



***Figure 7 – Informational Connectivity Results***

Color at each intersection represents informational connectivity (IC; see methods) for time-points in the first delay (8-16 seconds) between the ROIs listed for that intersection. A black asterisk indicates significance on a two-tailed t-test quantified from a bootstrap procedure across subjects against the null hypothesis that the informational connectivity is 0. The chart in (A) shows IC between ROIs for the attended (cued) information, (B) shows IC for unattended information, and (C) shows the difference between the two. So, taking the example trial shown in figure 1B, the attended information is quantified as the amount of Face evidence (cued) minus the amount of Abacus evidence (absent) at each time-point in the first delay (post Retrocue #1). This is computed separately for each ROI, and the associated time-course of this information metric is correlated between regions. Important to note is that "Contra-Vis" and "Ipsi-Vis" refer to the unilateral Visual ROI contralateral and ipsilateral to the cue (respectively). Therefore, in our example trial, the Contra-Vis ROI would be the right hemisphere, because the cue appears on the left for the first delay. The same thing is then done for the unattended information (in this case the evidence for the donut minus the evidence for the abacus) at each timepoint. The time-course for the unattended information is then correlated between ROIs. Again, "Contra-Vis" and "Ipsi-Vis" are relative to the cue, so even though the Donut appears on the right side of the screen in the example trial, "Contra-Vis" refers to the right Visual Cortex hemisphere.

# Chapter Three:

# The Effects of Attention on Internal Representations In Visual Working Memory

## Introduction

This work turns now to the question of how attention prioritizes representations of information no longer present in the environment. Current accounts of Working Memory (WM) suggest that attentional selection of relevant representations heavily relies on common attentional mechanisms engaged in processing perceptual information (D'Esposito & Postle, 2015; Gazzaley & Nobre, 2012; Kiyonaga & Egner, 2013; Oberauer & Hein, 2012; ). Accordingly, a growing body of research has shown that internally prioritizing mental representations is characterized by the same behavioral patterns and recruits the same neural systems as externally prioritizing visual information (Awh, Jonides, & Reuter-Lorenz, 1998; Emrich, Riggall, Larocque, & Postle, 2013; Nobre et al., 2004). Similar to prospectively cued visual information, retrocues guide the allocation of attentional resources to the cued memory information, improving its recall at the expense of uncued memory information (UMI, Griffin & Nobre, 2003; Lepsien & Nobre, 2007; Pertzov, Bays, Joseph, & Husain, 2013; Zokaei, Manohar, Husain, & Feredoes, 2013). This line of research has highlighted an important remaining question about the physiological consequences of selection in VWM: whereas attentional selection invariably strengthens the neural representation of cued information, what is its influence on the uncued information, which must nonetheless still be retained in VWM?

Research over the recent years has started to address the question pertaining to the fate of representations falling outside the focus of attention (FOA) by using information based multivariate pattern analysis (MVPA) of functional magnetic resonance imaging data (Lepsien & Nobre, 2007; J. a Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). In studies in which discrete objects are being held in VWM, evidence for the active representation of the UMI drops to baseline levels. For instance, Lewis-Peacock et al. (2012) decoded the time course of neural representations of discrete objects in VWM depending on their attentional state in a dual retro-cueing paradigm. In this task, after two sample stimuli were presented for encoding into WM, one of which was retrospectively cued during the subsequent delay period indicating which object would be relevant for the impending probe. While multivariate evidence for the active representation of the cued memory information was heightened, evidence of the UMI representations dropped to baseline levels even though participants knew that this unattended object had a 50% likelihood of being subsequently cued for a second probe. As discussed and demonstrated in Chapter 2, well-established principles of the biased-competition model are also reflected in population level representations, and thus have the potential to accommodate the finding that unattended representations do not have an elevated status in WM (Desimone & Duncan, 1995). More precisely, prioritization can be achieved by top-down excitatory signals that bias neurons representing the selected stimuli at the expense of unselected representations that lose competition as a consequence of lateral inhibition.

However, when multiple semantic frameworks of the same object were to be maintained in VWM, and retro-cues prioritized particular aspects of that object (e.g., its physical outline, its semantic category, the phonology of its verbal label), the UMI retains an intermediate level of elevated activity relative to the cued dimension and the baseline (Lewis-Peacock, Drysdale, & Postle, 2015). This raises the possibility that the well-characterized

"same-object" benefit in visual target detection, attributed to object-based attention, may also exist for VWM (Driver, 2001; Duncan, 1984; Vecera, Behrmann, & McGoldrick, 2000). The notion that object-based attention also applies to VWM was supported by behavioral benefits related to the recall of features from different dimensions (e.g. color, orientation) bound to the same object (Luck & Vogel, 1997; Luria & Vogel, 2011; Woodman & Vogel, 2008). Within this framework, Luck and Vogel (1979) showed that objects defined by a conjunction of four features can be maintained just as well as single-feature objects suggesting that the elementary units of VWM are integrated objects. However, others have argued that the elementary units of VWM are features that are simultaneously stored in dimension-specific storages, and that feature binding is achieved only when attention is exerted over the to-be-bound features (Bays, Wu, & Husain, 2011; Wheeler & Treisman, 2002). For instance, Wheeler and Treisman (2002) showed that while features from the same dimension (i.e. color-color) competed for attentional resources, features from different dimensions did not interfere with one another (i.e. color-orientation). Furthermore, they showed that same-object benefits were observed only when there were no other competing multi-feature objects during retrieval that would potentially disrupt sustained attentional control.

Although the same-object benefits can be explained by either sustained object- or feature-based attention (Driver, 2001; Vecera et al., 2000), these accounts pertain to the representational resolution with which stimuli are maintained, namely either as integrated multi-feature objects or as multiple independent features. However, what has not been shown yet is how individual features are selected from multi-feature objects and how this selection affects the unselected features. The findings of Lewis-Peacock and colleagues (2015) that unselected feature dimensions were marked by an intermediate level of representational activity relative to the baseline and attended feature dimension are indicative of how attentional selection mechanisms might operate at the level of features. The mechanism they

proposed was an interplay between the well-established principles of object- and feature-based attention and biased-competition. On the one-hand, object-based attention automatically prioritizes features that are integrated into objects themselves via spread of activation results in attentional benefits over features from unselected objects (Driver, 2001; Vecera et al., 2000). On the other, the selected feature dimension enjoys further prioritization to higher activation levels as a consequence of the biased-competition (Desimone & Duncan, 1995). It is important to note that this interpretation was inferred from data in which higher order abstract features (i.e. phonology and semantics) were bound to the same object. What has not been studied yet is whether these attentional selection mechanisms for abstract features also apply to low-level visual features (e.g. color and motion direction) that are bound into multi-dimensional object.

The aim of the current study was to put this interpretation to the test by examining the implications of attentional selection on features that were bound into the same objects in VWM. We carried out a multiple-step retrocuing paradigm in which two multidimensional objects (two patches of colored dots drifting coherently in one direction) were presented as memoranda, followed by an initial "general cue" that indicated whether one of the two objects (i.e. bound condition: color and direction of motion) or one of the two feature dimensions (i.e., unbound condition: the two patches of colors or the two patches of directions of motion) would be relevant for that trial; then, each of two serially occurring "feature cues" indicated which feature would be tested by the impending memory probe. We used an inverted encoding model (IEM) to track the time course of feature reconstructions and tested for attentional modulations at multiple delay-periods. The biased-competition model predicts that the neural representation of the retrospectively cued features would be strengthened regardless of binding (Desimone & Duncan, 1995). However, the effects of object- and feature-based attention would influence the unselected feature reconstructions

differently. In the bound condition, object-based attention automatically prioritizes features bound features via spread of activation that would result in elevated states of feature reconstructions (Driver, 2001; Duncan, 1984; Vecera et al., 2000). In the unbound condition however, features representations from the same feature spaces would compete for attentional resources which would eventually be reflected in the biased-competition effects at the level of individual feature.

# Methods

## Participants

Ten neurologically healthy students from the University of Wisconsin-Madison (3 females, 18-30 years, $M = 22$, all right-handed) participated in three 2-hour scanning sessions. One participant was excluded from the analyses due to excessive head movement. One other participant was author of this study. All participants had normal or corrected-to-normal vision and reported having normal color vision. The research complied to the guidelines of local ethics committee and all participants gave written informed consent.

**Design**

The experiment was comprised of an initial training and a subsequent testing phase. The rationale of conducting this study in two different phases was twofold. First, the task in the training phase would serve as an independent localizer for selecting voxels of interest that would be used to analyze attentional modulations induced by different cues in the testing phase. Second, fMRI data acquired in an independent training phase would generate a sufficient amount of data to model robust feature tuning curves of the training phase. These models would then be used to reconstruct the attended and unattended feature representations in the testing phase.

In the training phase, participants performed a one-item delayed estimation task in which a sample stimulus, either a patch of colored static dots (i.e. color condition) or a patch of coherently moving uncolored dots (i.e. direction condition), was presented for precise reproduction after a delay period. The direction and color conditions were pseudorandomly intermixed with a probability of 50% each. In the testing phase, participants performed the multiple-step retrocueing paradigm in which two multidimensional objects (two patches of dots conjointly being colored and coherently drifting in one direction) were successively presented as memoranda, and followed by several delay periods in which retrospective cues guided the FOA for selecting relevant stimulus dimensions for the impending probes. An initial "general cue" (denoted as Prelim Cue) indicated the Boundedness, namely whether one of the two objects (i.e. bound condition) or one of the two feature dimensions (i.e., unbound condition; the two color patches or the two directions of motion) would be relevant for that trial; then, each of two serially occurring "feature cues" (Retro cue 1 and Retro cue 2) indicated which feature would be tested by the impending memory probe. Half the trials were bound condition and the other half unbound condition trials. The type of Retro cues that were subsequently employed to shift the FOA among the maintained features was dependent on

whether a trial belonged to the bound or unbound condition. In the bound condition, Retro cue 1 (direction or color) and Retro cue 2 (direction or color) selecting either one of the features from the maintained multi-feature object were orthogonally manipulated. This manipulation also entails that Retro cue 2 could be selecting the same (i.e. stay trials) or different feature (switch trials). In the unbound condition, the prelim cue that had either selected the direction or color feature across two objects, were further selected among by means of Retro cue 1 (select 1st or 2nd feature) and Retro cue 2 (select 1st or 2nd feature) that were also orthogonally manipulated. This manipulation also entails that Retro cue 2 would either select the same (stay trials) of different feature (switch trials) as Retro cue 1.

**Task Procedures**

Participants performed 360 trials of the one-item delayed estimation task divided into 24 scan runs of 15 trials in the training phase. They also performed a multiple-step retrocueing task that was comprised of 160 trials in total, divided up into 16 scanning runs of 10 trials in the testing phase (fig 1). These two phases of the experiment were run across three scanning sessions of about two hours each. The first scanning session was composed of 16 runs of the training phase. The second scanning session was comprised of the 8 remaining runs of training phase and 6 runs of the testing phase. The third scanning session was comprised of the remaining 10 runs of the testing phase. Before the start of each scanning session, participants were given the specific task instructions of the particular task of each scanning session. The tasks were practiced both in- and outside the scanner to make sure participants fully understood the task. The task procedures in each phase will now be described in detail.

**Training Phase.** Each trial started with a white fixation cross (2˚ width and height) that was presented for two seconds, followed by the sample stimulus at the center of the screen for one second. The sample stimulus was either a patch of coherently moving dots (i.e.

direction trials) or static dots presented in a certain color (i.e. color trials). The patch contained 400 dots (0.1° visual angle, VA each) that were displayed within an invisible circular aperture (5° VA radius). In the motion trials, motion was 100% coherent (constant speed of 3 °/s) and randomly sampled (without replacement) from a list of 180 direction values that spanned over the full direction space of 360˚ in increments of 2˚ (i.e. 1˚ through 359˚). The moving dots were presented in a neutral color (i.e. grey) other than in the color trials. In the color trials, the dots were stationary and presented in colors that were randomly sampled (without replacement) from of a list of 180 color values that spanned over the full color space of 360˚ in increments of 2˚ (i.e. 1˚ through 359˚). This color list was generated from an evenly distributed circle on the CIE L*a*b colour space, centred at $L = 80$ with radius 60. All colours had an equal luminance and brightness and only varied in hue. All stimuli were presented at a black background.

The delay period started next in which a white fixation cross was presented for nine seconds. After delay, the probe display appeared in which participants were to indicate their response within the response deadline of three seconds. In direction trials, the probe was a white circle (10° VA diameter) presented at the center of the screen with a white line from the center to the edge of the circle. Participants used a trackball to adjusted the orientation of the line within the circle that was positioned at a random orientation to exclude response preparation on each trial. In the color trials, the probe was a colour bar (12˚ x 3.5˚) presented at the centre of the screen with all 360 colours from the color space horizontally arranged along the bar. A vertical white line (3.5˚) was presented at a random position within the bar that was to be adjusted as in the direction trials. Once participants responded, the white lines became thicker (4˚ VA) and remained at the responded positions until the response deadline. The probe was then followed by feedback that was determined by the degree with which the response orientations deviated from probed orientations. The feedback types ‘great’, ‘good’

and ‘poor’ (3˚ VA) were presented at the center of the screen along with the respective deviations smaller than 15˚, between 15˚ and 30˚ and larger than 30˚. Finally, a blank display with a grey fixation cross was presented for 8 seconds in the intertrial interval (ITI). The total duration of a trial was 24 seconds and participants performed 180 motion trials and 180 color trials. The Feature Dimensions were pseudorandomly intermixed across 24 blocks of 15 trials.

**Testing Phase.** All stimulus parameters were the same as in the training phase unless specified otherwise. Each trial started with a white fixation cross that was presented for two seconds, and followed by the first sample stimulus at the center of the screen for one second. In contrast to the training phase where one feature dimension was presented at a time (motion direction OR color), the sample stimuli in the testing phase were presented as conjunctions of two feature dimensions forming a multifeature object (motion direction AND color). Both motion direction and color were randomly drawn from their respective features spaces. Note that the features in the testing phase were randomly drawn (with replacement) from the full color and direction spaces (360˚) in increments of 1˚ contrary to list of 180 feature values in the training phase. After the interstimulus interval of one second, a second multifeature object was presented. The feature orientations were also randomly drawn from their respective feature spaces, however, with a minimum angular separation of 40˚ relative to the feature orientations of the first multifeature object. After the offset of the second multifeature objects, a delay period of seven seconds started in which both multifeature objects were to be maintained. The Prelim cue was then presented for one second instructing the participants to either attend to one of the two multifeature objects (i.e. bound condition) or to attend to one of the features across the two objects (i.e. unbound condition). The probability of bound and unbound conditions was 50% of which the trials were pseudorandomly presented in an intermixed fashion.

In the bound condition, the cues "< first >" or "<second>" were centrally presented and instructed the participants to maintain precise representations of the two features that were bound into the cued object throughout the trial. This Prelim cue also entails that the uncued multifeature object can be dropped as this would be of no relevance for the remainder of the trial. After another delay period of seven seconds, the first retro-cue was presented (Retro cue 1) for one second. Retro cue 1 instructed the participants to attend to either "direction" or "color" dimension of the multifeature object that was being maintained as instructed by the Prelim cue. The retrospectively cued feature was further maintained in another delay period for seven seconds after which the probe display appeared (Probe 1). This probe display was the same as in the training phase with either the direction wheel or the color bar being presented depending on the cued dimension. Participants used the trackball to indicate their responses within the response deadline of four seconds. Half of the trials were direction and the other half were color trials that were pseudorandomly presented. Immediately after the offset of probe display, Retro cue 2 was presented for one second. In half of the trials, Retro cue 2 was the same as Retro cue 1 instructing participants remain their attention on the same feature (i.e. stay trials). In other half of the trials, Retro cue 2 was the different from Retro cue 1 instructing participants to shift their attention to the other feature dimension of the same object (i.e. switch trials). A final delay period followed next for seven seconds after which the probe display appeared (Probe 2). Participants indicated their responses again by moving the trackball within the response deadline of four seconds.

In the unbound condition, the cues "< direction>" or "<color>" were centrally presented and instructed the participants to maintain precise representations of the cued feature dimension that was unbound across two objects throughout the trial. This Prelim cue also entails that the uncued feature dimension can be dropped as this would be of no relevance for the remainder of the trial. After another delay period of seven seconds, the first

retro-cue was presented (Retro cue 1) for one second. Retro cue 1 instructed the participants to attend to either “first” or “second” object of the maintained feature dimension as instructed by the Prelim cue. The retrospectively cued feature was maintained in another delay period of seven seconds after which the probe display appeared (Probe 1). This probe display was the same as in the training phase with either the direction wheel or the color bar being presented depending on the cued dimension. Participants used the trackball to indicate their responses within the response deadline of four seconds. The first object was cued on half of the trials and the second object on the remainder of the trials. Immediately after the offset of probe display, Retro cue 2 was presented for one second. In half of the trials, Retro cue 2 was the same as Retro cue 1 instructing participants remain their attention on the same feature (i.e. stay trials). In other half of the trials, Retro cue 2 was the different from Retro cue 1 instructing participants to shift their attention to the other feature of the same dimension (i.e. switch trials). A final delay period followed next for seven seconds after which the probe display appeared (Probe 2). Participants indicated their responses again by moving the trackball within the response deadline of four seconds.

Finally, Probe 2 was followed by feedback on both the first and second probe response that was determined by the degree with which the response orientations deviated from probed orientations as in the training phase. Feedback for the first and second response probes was respectively presented for one second above (e.g. “First: good”) and below (e.g. “Second: great”) the fixation cross, subtending 3˚ (VA). Finally, a blank display with a grey fixation cross was presented for 7 seconds in the ITI. The total duration of a trial was 52 seconds and participants performed 160 pseudorandomized trials across 16 blocks of 10 trials.

**Behavioral Data Analysis**

The behavioral analysis was performed by means of the methods used by Bays et al. (2009; http://www.paulbays.com/). Continuous measures of error were obtained on each trial as the angular distance between the reported feature orientations (first and second probe response) and the probed feature orientations (first and second target orientations). A precision measure was then calculated as the reciprocal of the standard deviation of the error for the first and second probe independently. Because both color and direction feature dimensions were drawn from circular spaces, Fischer's definition of standard deviation for circular data was used and corrected for expected guessing if participants had responded at random on each trial. The precision measure for the first probe was averaged across all participants and subjected to a repeated-measures analysis of variance (ANOVA) with Boundedness (Bound or UnBound) and Feature Dimension (Direction or Color) as within-subjects factors. The precision measure for the second probe was also averaged across all participants and subjected a repeated-measures ANOVA with Boundedness (Bound or UnBound), Feature Dimension (Direction or Color) and Trial Type (Stay or Switch) as within-subjects factors. Trials on which no responses were given, were excluded from the analyses. An alpha level of .05 was applied and Bonferroni correction was used on multiple tests to control for false-positives in post-hoc testing.

Furthermore, a probabilistic model of performance that was previously proposed by Bays and Husain (2008) decomposed the sources of errors into a mixture of Gaussian variability around the target colour ($P_T$), binding errors ($P_{NT}$) and a fixed probability of random guessing ($P_U$). This model can be described as follow:

$$p(\hat{\theta}) = (1 - \gamma - \beta)\phi_\kappa(\hat{\theta} - \theta) + \gamma\frac{1}{2\pi} + \beta\,\phi_\kappa(\hat{\theta} - \theta^*),$$

where $\theta$ is the target feature (probed), $\hat{\theta}$ the colour reported by the participants, and $\phi_\kappa$ the von Mises distribution (circular analog of the Gaussian) with mean zero and concentration

parameter $\kappa$. The probability of misremembering the target feature is $\beta$ with the orientation value of the non-target feature. The probability of random guessing is captured by $\gamma$. Maximum likelihood estimates of the parameters $\kappa$, $\beta$, $\gamma$ were separately obtained for each subject only in the unbound condition as this model only captures misbinding errors within a feature dimension. The mixture model components related to the first probe (probability target responses, binding errors and guesses) were independently subjected to a univariate repeated-measures ANOVA with Feature Dimension as a within-subjects factor. The mixture model components related to the second probe were independently subjected to the same analysis with Feature Dimension (Direction or Color) and Trial Type (Stay or Switch) as within-subjects factors. A graphical representation of these models components are given in figure 2B. An alpha level of .05 was applied and Bonferroni correction was used on multiple tests to control for false-positives.

**Data Acquisition and Preprocessing**

Whole brain images were acquired with the 3 T MRI scanner (Discovery MR750; GE Healthcare) at the Lane Neuroimaging Laboratory at the University of Wisconsin-Madison. High- resolution T1-weighted images were acquired for all subjects with an FSPGR sequence (8.132 ms time repetition (TR), 3.18 ms time echo (TE), 12° flip angle, 156 axial slices, 256 × 256 in- plane, 1.0 mm isotropic). Blood oxygen level-dependent (BOLD)- sensitive data were acquired using a gradient-echo, echoplanar sequence (2 s TR, 25 ms TE) within a 64 × 64 matrix (39 sagittal slices, 3.5mm isotropic).

*fMRI data analysis*

fMRI data analysis was performed using Analysis of Functional NeuroImages (AFNI) software package (http://afni.nimh.nih.gov; Cox, 1996). All volumes were spatially realigned to the final volume of the final functional run using rigid-body realignment. The processing

pipeline included slice time correction, detrending, conversion to percent signal change, and spatial smoothing with a 4-mm FWHM Gaussian kernel.

*Generation of ROIs*

In order to generate appropriate regions of interest (ROIs) for the MVPA, we first generated anatomical masks which broadly covered three separate regions of cortex: visual cortex (covering parts of both occipital and temporal cortex), parietal and frontal. We then selected a subset of voxels from each anatomical mask which showed the greatest sensitivity to our behavioral stimuli, as indexed by a functional GLM.

*Anatomical ROI Generation*

Anatomical ROIs were generated using the Talraich anatomical atlas (TTatlas; https://sscc.nimh.nih.gov/afni/doc/misc/afni_ttatlas/). Briefly, coordinates for relevant gyri in the TTatlas were used to generate masks for each gyrus, which were then warped into an individual's original space, and aggregated based on broad anatomical location to create a regional mask. For instance, to create the "frontal" anatomical mask, TTatlas masks were generated of the inferior frontal gyrus, middle frontal gyrus, superior frontal gyrus, medial frontal gyrus and precentral gyrus, which were then warped into subject space and aggregated.

*Functional ROI Generation*

We performed a general linear model (GLM) analysis for each subject on the data from the one-item change-detection task in order to identify brain ROIs for the MVPA. A single regressor was included for stimulus onset and another for the delay period, separately for color and direction, along with covariates to control for motion and block-specific effects.

Stimulus-onset was modeled as a boxcar of 3 seconds covering the full time both stimuli were present on the screen. The delay-period was modeled as an 9 second boxcar. All were convolved with a canonical hemodynamic response function. Each of these independent regressors was entered into a modified GLM for analysis using AFNI. In order to capture voxels which were likely to be useful for subsequent MVPA, for each subject, for each area (occipital/temporal, parietal and frontal) we extracted the top 400 voxels with the highest positive t-statistic associated with the color vs motion-direction comparison, separately for delay regressors and sample regressors. A general version of these two masks (delay sensitive and sample sensitive) is shown in figure 1). Once we had these two t-maps, for each broadly defined anatomical region (Occipital, Parietal, and Frontal) we extracted the top 400 voxels sensitive to the sample on color trials, and the same thing for direction trials to generate 12 total masks (3 anatomical areas x 2 timings x 2 features). A similar approach to ROI generation has been used in prior studies in our laboratory (Riggall and Postle 2012; Emrich et al. 2013) and has the advantage of accounting for individual differences in task-relevant neural activity.

**Forward (Inverted) Encoding Model**

In order to assess the representational fidelity of mnemonic representations, we built a forward encoding model from the motion-direction selective and color selective neural responses, separately for each ROI. Briefly, these models assume that the response in each voxel can be expressed as the linear sum of underlying neural activity, and that for at least some voxels, this activity varies as a function of motion direction or color, respectively (Brouwer and Heeger, 2009).

More specifically, following the steps laid out in Ester et al (2015), for each feature dimension (color and motion-direction) we modeled each voxel's response as a linear sum of

9 information channels which, taken together, spanned the full stimulus space for each feature. This relationship can be expressed in the form of the equation:

$$(Equation\ 1) \qquad b = \sum_{n=1}^{9} w_n * c_n$$

where $b$ is a single voxel's response, and $c_1$-$c_9$ are the expected responses of the nine information channels to the stimulus conditions that generated $b$. Under this formulation, $w_1$-$w_9$ thus represent the relative strength (or weight) of the contribution each channel makes to the voxel's overall response. This set of weights is sometimes referred to as a population receptive field or a voxel tuning function. Extending this framework to the full dataset yields the following general linear model:

$$(Equation\ 2) \qquad B_{v,t} = W_{v,n} * C_{n,t}$$

Here, $B_{v,t}$ is the matrix of responses for $v$ voxels on each of $t$ trials, $c_{n,t}$ is the analogous matrix of expected responses for $n$ channels on each of $t$ trials. Therefore, the weight matrix $w_{v,n}$ constitutes a mapping between "channel space" and "voxel space". The first step in implementing the model (the "training" phase) is to construct a basis set in stimulus space, upon which a portion of the voxel data can be regressed (equation 3), generating the weight matrix.

$$(Equation\ 3) \qquad W_{v,n} = B_{v,t} * C_{t,n}^T * \left(C_{n,t} * C_{t,n}^T\right)^{-1}$$

In the second step, this weight matrix can then be inverted and applied to the remaining voxel data in order to generate an expected output in channel space, like so:

$$(Equation\ 4) \qquad C_{n,t} = \left(W_{n,v}^T * W_{v,n}\right)^{-1} * W_{n,v}^T * B_{v,n}$$

For the motion direction model, our basis set consisted of nine von Mises functions (equation 5), each centered (by varying the $\mu$ parameter) around a direction 40 degrees apart (at 20˚, 60˚, 100˚, etc), thus covering the full 360 degree feature space. In equation 5, x is a vector of channel responses. $\mu$, k, and $\beta$ control the center (i.e., mean), concentration (i.e., inverse of

width) and baseline (i.e., vertical offset) of the function, while $\alpha$ corresponds to the amplitude of the function (i.e., vertical stretching/scaling; signal above a noisy baseline).

$$(Equation\ 5) \qquad f(x) = \alpha\left(e^{k(\mu-x)-1}\right) + \beta$$

We set the $\alpha, k$, and $\beta$ parameters to 1,7, and 0 respectively to best approximate tuning properties of MT neurons (Duijnhouwer et al, 2013). To model color, we extracted circular portion of *L*A*B* color space using the procedure outlined in Brouwer and Heeger (2009), and then used the same basis set outlined above to cover the full 360 degrees of extracted space (and importantly, we constrained our presented stimuli to this same portion of space).

As an initial validation step, we implemented a leave-one-run out cross-validation procedure where each of the runs from the one-item memory task (the Training Phase) was set aside and each time-point from the remaining runs was used to generate a weight matrix for each ROI for each feature dimension (color and motion direction) as described above. We then inverted the weight matrix and applied it to trials from the left-out run to generate reconstructions (separately for each time point, ROI and feature dimension). Reconstructions from each iteration of the leave-one-run-out procedure were then averaged together to generate reconstructions for the Training Phase. These results are shown in figure 3.

For most of the analyses, however, for each ROI, Voxel BOLD percent signal change values from the final delay period time-point in our one-item training task were regressed on each of our two basis sets separately, such that data from motion direction trials were regressed on the motion direction basis set for motion-direction ROIs, and data from color trials were regressed on the color basis set for color-sensitive ROIs. This late-delay TR was chosen to maximize the likelihood that subsequent reconstructions would reflect mnemonic representations, as the sample evoked BOLD response would have already run its course.

Once the weight matrix was generated from training phase data, data from EACH time-point in the Attention retrocuing task (Testing Phase) was multiplied by the inverted

weight matrix as described in equation 4 to generate a reconstruction timecourse for both motion direction and color. Each of these feature-specific reconstruction time-courses were then circularly shifted to a common center (0˚) and averaged with other trials of identical circumstance. Thus, to generate the "Attended" reconstructions in Figure 4 (the "Bound" condition), channel outputs from trials where either the first or second item was Prelim cued and "Direction" was the first retrocue were aligned along the either the first or second item's direction (respectively) and averaged together. To generate the smooth, 360-point functions shown in Figures 4,5 and 6, we repeated the encoding model analysis a total of 39 times and shifted the centers of the motion direction or color channels by 1˚ on each iteration.

Reconstructions were quantified using bootstrapping in a process similar to that employed by Ester et al (2015). In each ROI, for each feature dimension, each time-point, and each condition (e.g. BoundCued), reconstructions from all 9 subjects were randomly sampled with replacement 9 times to generate a 9x360 dimension resampled reconstruction matrix. This resampled matrix was averaged across the first dimension (subjects), yielding an averaged reconstruction that was then fit with a von Mises (equation 5). The $\mu$ parameter was fixed to 0 (reflecting our alignment across trials to a 0˚ centered channel) and the remaining parameters were fit using a combination of a gridsearch procedure and a General Linear Model. This procedure was repeated 10,000 times in total, yielding 10,000 bootstrapped estimates of amplitude, baseline, and concentration. A p-value for the reconstruction robustness was then calculated by assessing the percentage of bootstrapped iterations whose amplitude estimates were negative (in other words, $p<0.05$ implies that >95% of resampled reconstructions have a positive amplitude).

# Results

## Behavioral

The analysis on the precision measure for first probe only revealed a main effect of Feature Dimension [$F(1,8) = 10.76$, $p<.05$, $\eta_p^2=.57$] with an overall higher precision for direction ($M = 2.99$, $SE = .61$) than for color ($M = 1.44$, $SE = .19$). Other effects did not reach significance ($Fs < 2.5$). Similarly for the second probe, there was only a significant main effect of Feature Dimension [$F(1,8) = 7.62$, $p<.05$, $\eta_p^2=.49$] with an overall higher precision for direction ($M = 2.46$, $SE = .49$) than for color ($M = 1.53$, $SE = .25$). None of the other effects reached significance. (fig 2)

## Encoding Model Reconstructions

*Reconstruction of one-item Training Phase*

The primary aim of the study was to utilize a forward encoding approach (see methods) in order to quantify the effects of top-down attention on multi-feature working memory representations. Before applying it to the data from the more complicated multi-retrocue paradigm, however, we first sought to validate the method by using it to characterize the far simpler one-item memory task using a leave-one-run-out r-fold cross-validation procedure (described in the methods). The first thing to note (fig. 3) is that while the general pattern of results is similar for both color and motion-direction models, the motion-direction model produced far more robust reconstructions than the color model. Because participants also performed worse on the color trials behaviorally (displaying larger errors), the increased noisiness of the color reconstructions likely reflects underlying noisiness of the neural activity itself and not a fundamental failure of the approach in modeling color space. Because of this result, the rest of this section that deals with reconstructions from Testing Phase data will focus on motion-direction reconstructions to maximize sensitivity. Additionally, these results provide evidence for a dynamic mnemonic code, in that models trained and tested on

the same time-point (the "diagonal") outperform models trained and tested on different time-points ("off-diagonal"). This result helped motivate the selection of the final delay-period TR (~10-12 seconds after trial onset) as a training time-point for subsequent analyses, in that underlying activity at that time-point was most likely to reflect mnemonic maintenance, given that it occurred past the sample-evoked bold time-course but before the probe onset.

*Effects of feature-based attention*

In order to assess the effects of feature-based selective attention on mnemonic representations, we next applied our model built upon data from the late delay period of the one-item Training Phase to our Testing Phase: a multiple-retrocue delayed estimation task. Figure 4 presents data from "Bound" Trials, which were defined as trials where the Prelim cue was either ">first<" or ">second<", instructing participants that both the color and motion-direction of the cued stimulus could possibly be tested on either of the two upcoming probes. Importantly, because this Prelim cue was 100% valid, participants could subsequently drop from memory the color and direction associated with the uncued stimulus. Reconstructions from these "Bound" Trials were then sorted into one of two groups, based upon whether Retro cue 1 instructed participants that the associated color or motion-direction of the Prelim-cued stimulus would be probed first. Reconstructing motion-direction during trials where motion-direction was cued by Retro cue 1 thus constituted "Attended" trials, in that motion-direction was the cued feature dimension, necessary for the upcoming probe. In contrast, reconstructing motion-direction on trials where *color* was cued by Retro-cue 1 constituted "Unattended" trials, in that motion-direction would be attentionally deprioritized by the cue. Unlike the Prelim cue however, participants couldn't fully drop the unattended feature-dimension, as it may still be needed for Probe 2. Once reconstructions from the

"Bound" trials were sorted this way, they were aligned along the direction of the Prelim cued item and averaged to generate the time-courses shown in figure 4. For Frontal (Fig. 4A,B) and Parietal (Fig. 4C,D) ROIs, reconstructions were robust only transiently at time-points surrounding Retro cue 1 onset, and only for trials where motion-direction was the attended feature dimension. This is consistent with data presented in Chapter 2, where only attended stimulus category could be decoded from parietal and frontal ROIs. For the Occipital ROI, however (Fig. 4 E,F), robust reconstructions were found for the majority of time-points surrounding Retro cue 1 for both Attended and Unattended motion-directions. More specifically, robust reconstructions for "Attended" and "Unattended" trials extend well into time-points 16 and 15 respectively (~10 seconds after Retro cue 1), by which time the probe has appeared on the screen. By TR 15, "Attended" Reconstructions have become significantly more robust (defined as possessing a larger amplitude fit, quantified using bootstrapping), indicating that a consequence of attentional selection is a suppression of the robustness of the unattended feature dimension. The Informational Connectivity Analysis of Chapter 2 argues that top-down signals from frontal and/or parietal areas directly modulate representational fidelity in visual cortex. Hints of a mechanism by which these areas accomplish this can be seen in Figure 5, which plots the reconstructions of individual time points for each ROI. For time-points just after the Retro cue 1 onset (11-13), motion-direction can still be robustly reconstructed regardless of whether the cued dimension was motion-direction or color. But, by time-point 15, the reconstructions on motion-direction cued trials have become more robust, a result which coincides with a baseline-shift in the Reconstructions in frontal and Parietal ROIs.

*Effects of Object-based attention*

In order to assess the consequences of object-based attention for representations in working memory, as before, we considered separately the effects of Retrocue 1 and the Prelim cue. Again, Retrocue 1 has the effect of prioritizing one of two remaining features in working memory for a subsequent behavior probe whereas the Prelim cue allows participants to drop two of the four initially encoded features. Figure 6 shows the results of "UnBound" trials, where the Prelim cue was "<direction>". The "Attended" time-course was generated by aligning reconstructions from these trials along the item's direction that was cued by Retrocue 1 and the "Unattended" time-course was generated by aligning along the uncued item's direction. In the Occipital ROI, in contrast to the results of feature-based attention, "Unattended" reconstructions quickly lose robustness following the onset of Retrocue 1, while "Attended" Reconstructions persist late into the delay. The single time-point plots of Figure 7 demonstrate another important distinction. Neither parietal nor frontal areas show robust reconstruction nor baseline shifts. Additionally, at time-point 15, "Unattended" reconstructions in the Occipital ROI appear to "invert" themselves, displaying minimal channel output along the aligned direction and maximal output for channels representing the opposite direction (180˚ away). While this effect reaches only trending significance here (quantified using the same bootstrap procedure as before but assessing for negative amplitude estimates), it does reach significance when considering the effects of an object-directed Prelim cue. Figure 8 shows individual time-point reconstructions immediately surrounding the Prelim cue. Cueing the first item causes reconstructions aligned along the uncued items direction to significantly invert in the Occipital ROI, with no corresponding effects in either Parietal or Frontal cortex.

## Discussion

Using a forward encoding model approach that allows for the quantification of specific parameters in representational feature space, we find specific neural population-level signatures of feature- and object-based attention directed towards internal representations held in visual working memory.

*Signatures of object-based attention*

When participants are cued that one of two objects will be relevant for an upcoming task, and that the other can be dropped from memory, the reconstruction in Occipital cortex of the representation of the cued item becomes sharper, while the reconstruction of the representation of the uncued item actually inverts its tuning profile (figure 8). A recent study linked the sharpening of the reconstruction of the attended item to increased behavioral performance (Sprague et al, 2014), though did not report a commensurate inversion of the uncued reconstruction. Under one interpretation, this inverted reconstruction is consistent with structured suppression of a particular set of channels in feature space. Taken this way, the combination of the sharpening of a cued representation and the suppression of the uncued representation is a fundamental prediction of biased competition. A similar pattern is evident following the first retrocue in "UnBound" trials, though the effects are much less pronounced. This is perhaps due to a task-specific difference between the cues, in that the Prelim cue informs participants that an item can be dropped, whereas the first retrocue is "softer" in its implications. Participants cannot fully drop the uncued item in response to the first retrocue, because it may still be relevant for the second probe. Data from the final delay period following the second retrocue are currently being analyzed and a prediction of the results so far is that the pattern will be similar to that

observed in response to the retrocue. It is important to note that a 180 degree shift of all channels in the feature space cannot account for these results, as it would also change the alignment of the cued dimension.

Somewhat puzzlingly, we were unable to reconstruct direction specific information in frontal or parietal areas. From the results of the experiment in Chapter 2, we predicted that reconstructions of the attended direction would be found in parietal or frontal areas. Ester and colleagues (2015) were able to reconstruct orientation in frontal and parietal areas, though there are a couple important methodological differences between their study and ours. The first is that the overall load in our study is higher, so it is possible that the increased noise has a more significant effect on frontal and parietal areas than in occipital. The second is that their stimuli were presented simultaneously in separate visual hemifields similar to the study presented in Chapter 2, where we were able to decode external representations in frontal and parietal areas. The third is that their retrocue immediately followed the sample presentation, a consideration whose relevance will be re-examined in the discussion of Chapter 4. Again, once the data from retrocue 2 are examined, it is possible that reconstructions will be obtained in these areas.

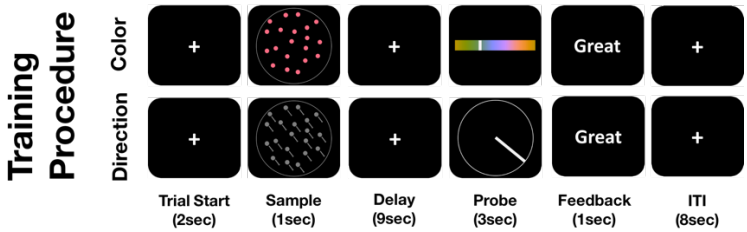
*Signatures of feature-based attention*

In the current experimental design, cues to a specific feature dimension were necessarily also cues that prioritized one half of a bound pair. Thus, a prediction made in the introduction is that different mechanisms will be needed to implement the prioritization (because prioritization of one feature will necessarily spread to the other dimension, potentially negating a prioritization benefit). What we observe is that in response to the first retrocue, unlike for the unbound condition, reconstructions of the unattended information remain robust late into the delay. Only

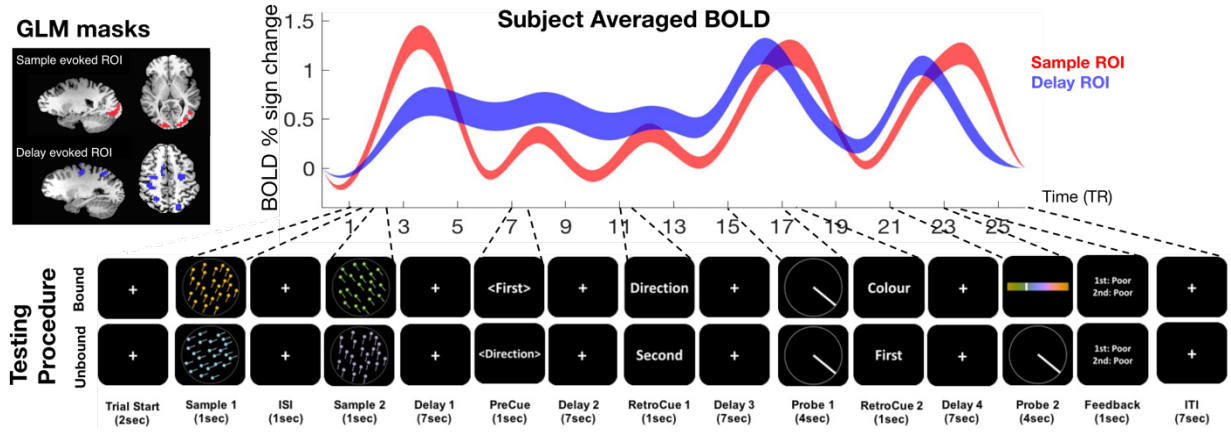
after a shift in the baseline parameter estimates of reconstructions in parietal and frontal does the reconstruction of the uncued feature dimension become non-robust. A shift in the baseline parameter can be thought of as a category level representation in that all channels of a particular feature space are elevated relative to another feature space. Thus, as expected, the allocation of feature-based attention is associated with an exemplar non-specific enhancement of the attended feature channels in both parietal and frontal areas. This pattern precedes the exemplar-specific enhancement of the attended feature in occipital cortex, which is necessary for output on the subsequent behavior probe. This is also a convenient way of handling the binding problem and getting around biased competition, in that by non-specifically elevating the feature space of the attended feature at the expense of the the unattended feature space, the exemplar specific information is not suppressed, allowing it to be brought back to the fore if the second retrocue demands. A prediction of the second retrocue will be that once the trial is coming to an end, this "work-around" will be unnecessary and perhaps a feature-specific suppression will be observed in the Occipital ROI. On the other-hand, if this base-line shift is indicative of feature-based attention in general, it will still be observed during the second retrocue.

# Figures

**A**




**B**

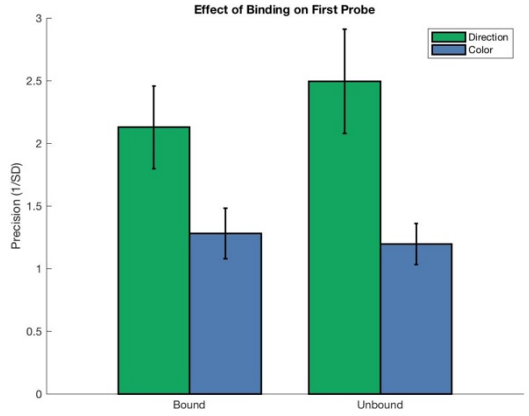



## *Figure 1 – Task Diagrams and ROIs*

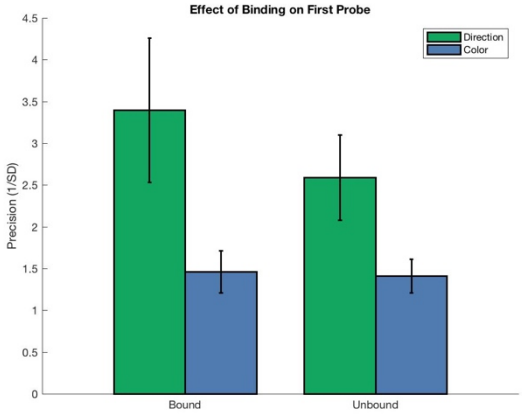
(A) *Training Task Diagram.* Participants completed 180 trials each of a one-item delayed recall test for Color and a one-item delayed recall test for motion-direction. Each trial constituted one of 180 evenly spaced exemplars in the feature space (see methods), randomized across all runs.

(B) *Testing Task Diagram and BOLD signal plot.* In addition to the training task, participants completed 160 trials of the Testing Task, which consisted of multiple retrocues to shift attention between features and exemplars held in working memory. "Bound" trials were defined as trials where the Prelim retrocue (TR 7; ~12 seconds in) informed participants that only one of two sequential patches would be relevant for the upcoming behavior probes. "Bound" refers to the fact that after this cue, participants only have to remember a single color and motion direction, which are "bound" into a single patch of moving dots. For "unbound" trials, the Prelim retrocue tells participants that only one feature dimension will be relevant, thus after the cue, the only information that is required to complete the task is either two "unbound" motion directions or colors. The BOLD plot shows BOLD percent signal change of voxels whose response is best explained by either the sample (red trace) and delay period (blue trace) regressors (see methods). The "GLM masks" inset shows where these voxels are located spatially. Masks presented in subsequent figures are the subset of delay period voxels in Frontal areas ("Frontal" mask), the subset of delay period voxels in parietal ("Parietal" mask), and the subset of sample voxels in Occipital Cortex ("Occipital Mask").

A




B




## *Figure 2 – Behavior*

(A) Data from separate behavior-only experiment run outside the scanner. Precision is plotted as 1/StandardDeviation(trial errors), with each bar representing the average precision across participants for each cue type. Because the experimental design allowed for angles of separation between exemplars of a single feature to be less than 40°, in order to compare these results with data collected inside the scanner (where the minimum angle of separation was set to 40°) we considered only those trials where the angle of separation was larger than 40°. Other than a main effect of feature type (color responses have larger errors overall), there is no effect of trial type. A 3-factor mixture-model (http://www.paulbays.com/code/JV10/index.php) also failed to show an effect of cue type (Bound vs UnBound) on the estimated concentration parameter, proportion of non-target responses, or proportion of guess responses.

(B) Data from trials participants completed during scanning. Consistent with the behavioral experiment outside the scanner, participants showed greater errors overall during Color cued trials, but no binding effects of cue type were observed. A mixture-model analysis was unable to be performed on these data, due to lack of sufficient trial number

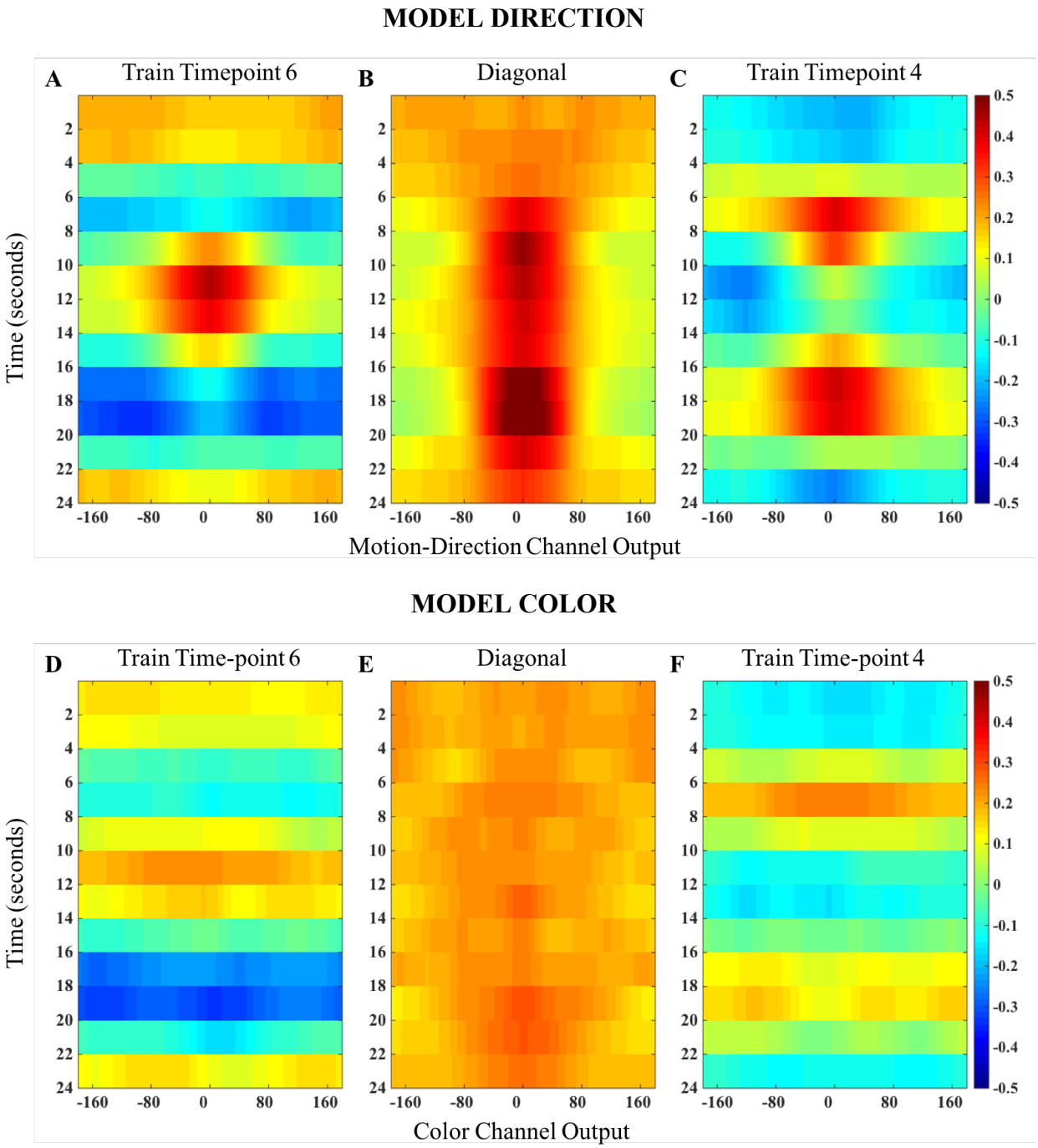
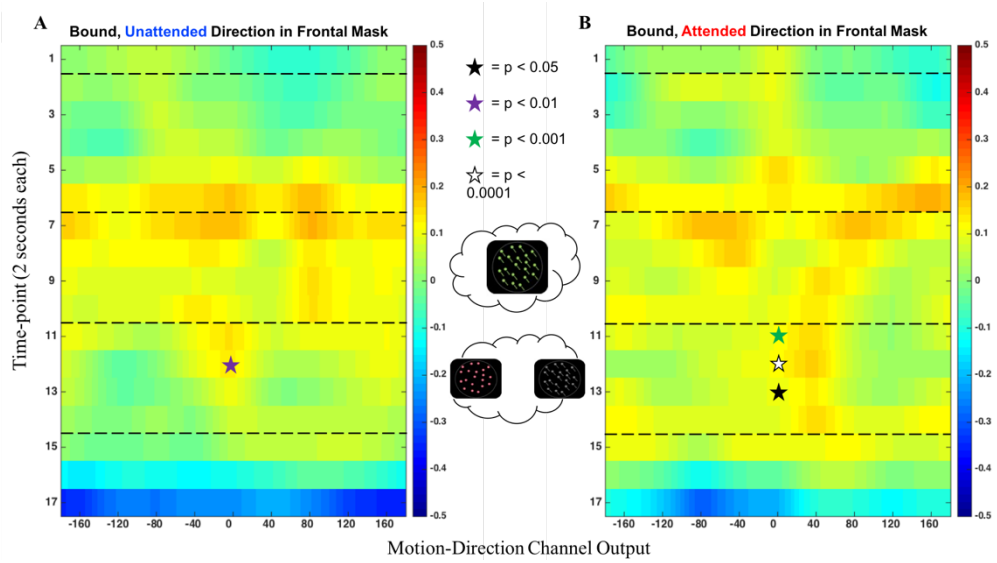
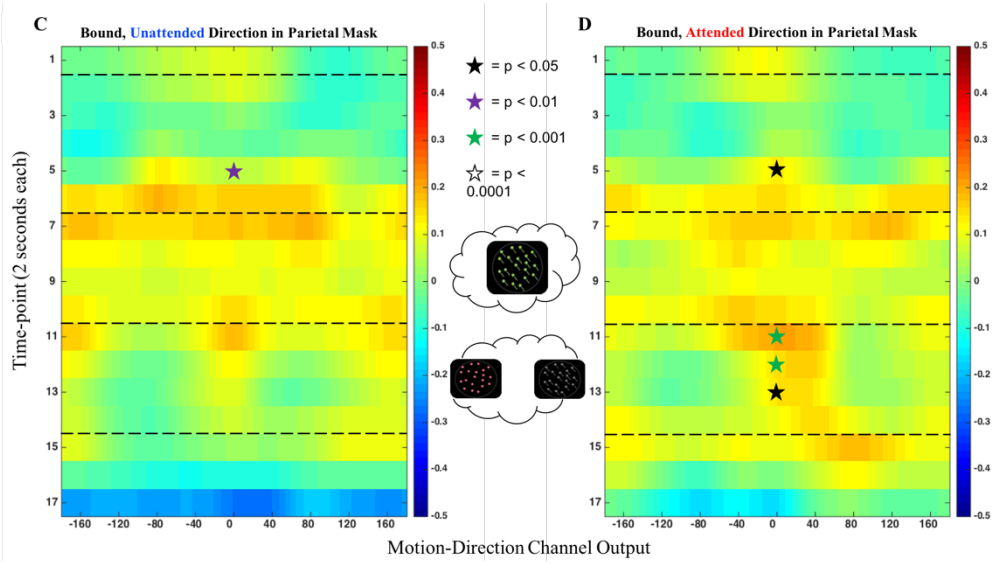
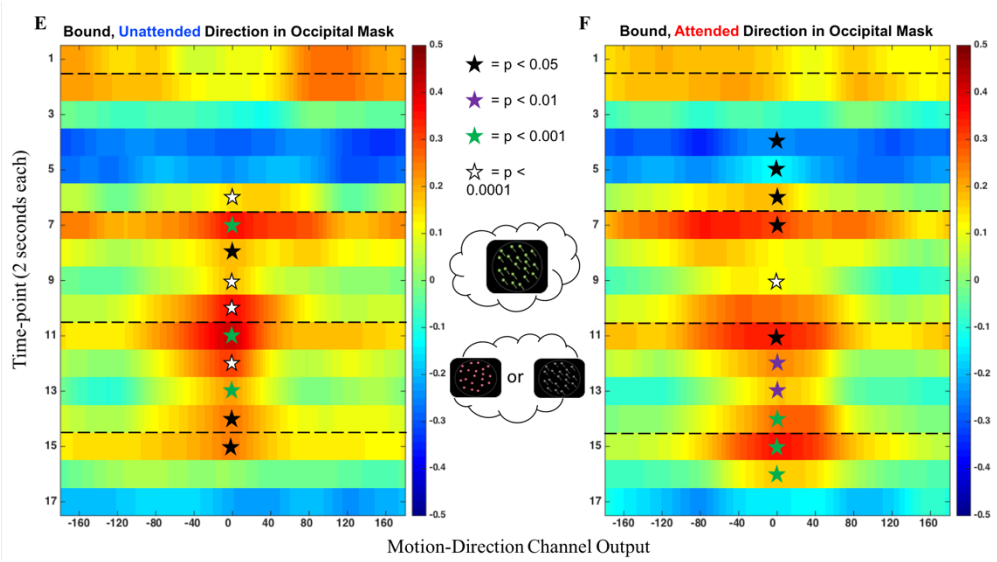



***Figure 3 – Leave-one-run-out Reconstruction Validation procedure for Experiment One***

Reconstruction time-courses for one-item memory task of Experiment 1 in the Occipital ROI. The color of each point indicates the output, at each 2-second long time-point volume (TR) of a motion direction encoding model (A-C) and a color encoding model (D-F). Outputs from each trial are aligned at 0° and averaged as detailed in the methods. Direction outputs are significantly more robust than Color outputs, though both display evidence of a dynamic representational code.


A
Bound, Unattended Direction in Frontal Mask
B
Bound, Attended Direction in Frontal Mask
= p < 0.05
= p < 0.01
= p < 0.001
= p < 0.0001
Time-point (2 seconds each)
Motion-Direction Channel Output


C
Bound, Unattended Direction in Parietal Mask
D
Bound, Attended Direction in Parietal Mask
= p < 0.05
= p < 0.01
= p < 0.001
= p < 0.0001
Time-point (2 seconds each)
Motion-Direction Channel Output


E
Bound, Unattended Direction in Occipital Mask
F
Bound, Attended Direction in Occipital Mask
= p < 0.05
= p < 0.01
= p < 0.001
= p < 0.0001
or
Time-point (2 seconds each)
Motion-Direction Channel Output

### *Figure 4 - Experiment 2 "Bound" Trial Time-course Reconstructions*

Time-courses for an encoding model trained on one-item direction trials from Experiment 1 and tested on "Bound" trials from Experiment 2. Time-courses display time-point by time-point reconstructions the first 17 time-points, up through the first probe (see figure 1). Significance of robustness of reconstruction indicated by star color, using a bootstrap procedure outlined in the methods. Black dotted lines indicate trial events, including sample, Prelim cue, Retrocue 1, and probe onsets. Cartoon in center indicates what general feature information is cued for each section of the trial. (A) and (B) show reconstruction time-courses for the Frontal ROI, (C) and (D) for the Parietal ROI and (E) and (F) for the Occipital ROI. The left column averages together trials where Retrocue 1 was "color", and the right column shows trials where Retrocue 1 was "direction".

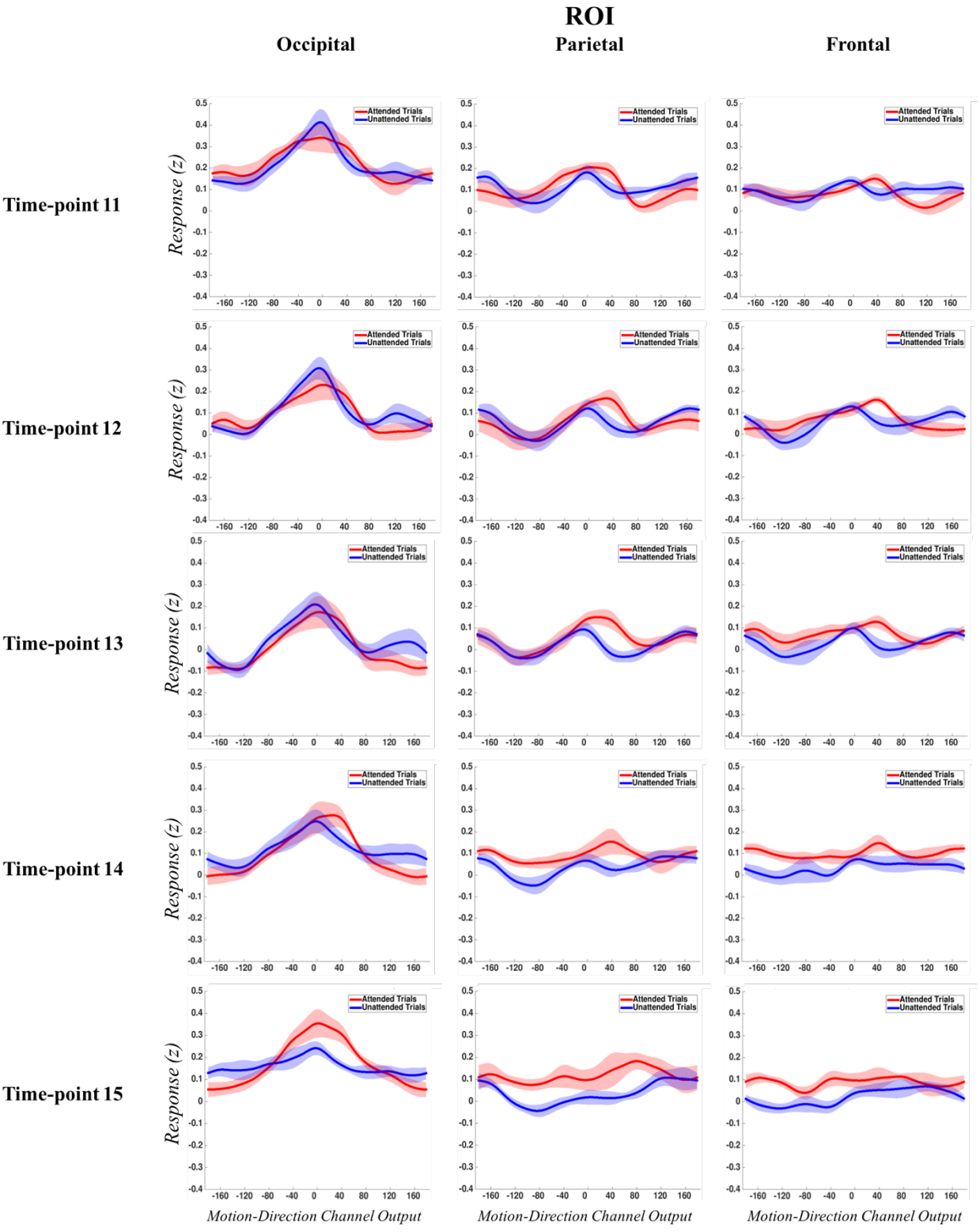

ROI
Occipital
Parietal
Frontal
Time-point 11
Time-point 12
Time-point 13
Time-point 14
Time-point 15
Response (z)
Attended Trials
Unattended Trials
Motion-Direction Channel Output

***Figure 5 – Individual Time-point Plots for "Bound" Trials***

Channel response is plotted on the Y-axis for individual time-points, and the specific Motion-direction channel is plotted on the x-axis, centered around 0°. Each column is one of the 3 ROIs: Occipital, Parietal and Frontal respectively. Each row is a set of reconstructions from a single time-point starting with the Retrocue 1 onset (Timepoint 11) and ending with the probe onset (Timepoint 15). Red traces represent reconstructions of directions on trials when direction was cued ("Attended" reconstructions) and blue traces represent reconstructions of direction on trials when color was cued ("Unattended" reconstructions).

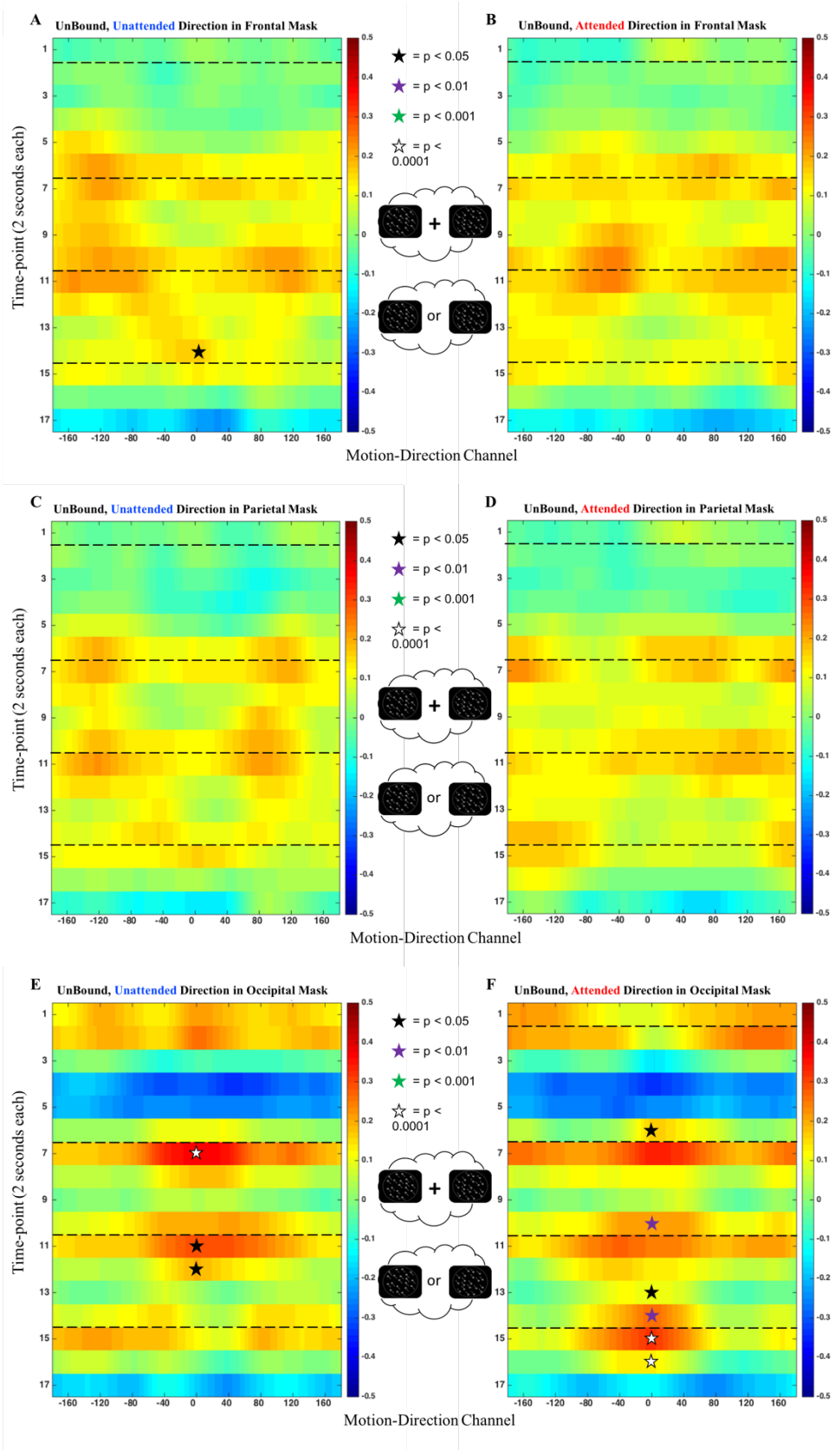

A
UnBound, Unattended Direction in Frontal Mask
B
UnBound, Attended Direction in Frontal Mask
C
UnBound, Unattended Direction in Parietal Mask
D
UnBound, Attended Direction in Parietal Mask
E
UnBound, Unattended Direction in Occipital Mask
F
UnBound, Attended Direction in Occipital Mask
= p < 0.05
= p < 0.01
= p < 0.001
= p < 0.0001
+
or
Time-point (2 seconds each)
Motion-Direction Channel

***Figure 6 - Experiment 2 "UnBound" Trial Time-course Reconstructions***

Conventions are the same as in Figure 4, except that "UnBound" trials are considered. These are trials where the Prelim Cue was "<Direction>".

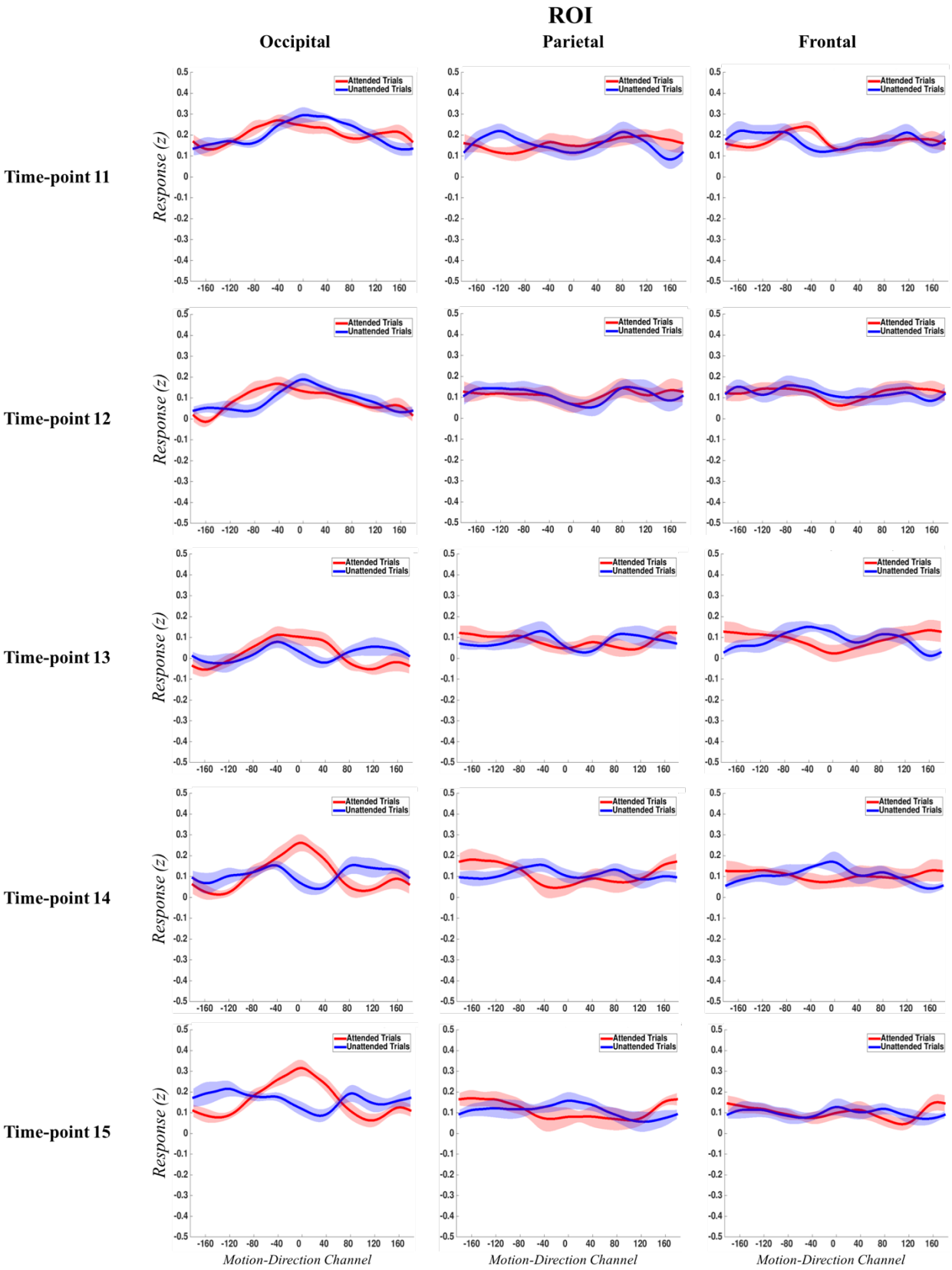

ROI
Occipital
Parietal
Frontal
Time-point 11
Time-point 12
Time-point 13
Time-point 14
Time-point 15
Response (z)
Attended Trials
Unattended Trials
Motion-Direction Channel

***Figure 7 – Individual Time-point Plots for "UnBound" Trials***
Same Conventions as figure 5.

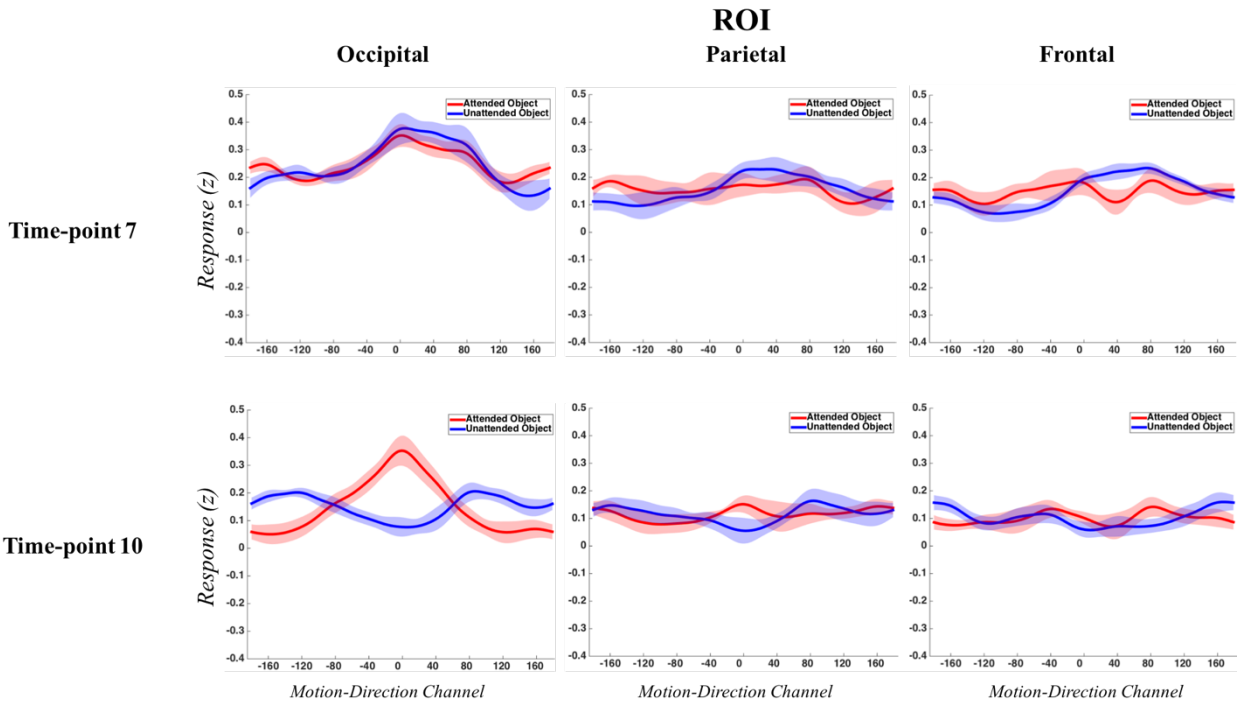



***Figure 8 – Individual Time-point Plots for object-based Prelim Cue Trials***

Similar Conventions as figure 5 and 7, however, what is being plotted here is the output of Motion-direction channels aligned to either the direction of the item which is Prelim-cued (red trace) or aligned to the direction of the item which is not Prelim-cued (blue trace). Data from two time-points are presented: Timepoint 7 is the cue onset, and thus reflects the neural state before the cue, and Timepoint 10, which represents a maximal cue evoked response.

# Chapter 4: The Interaction of Task-relevance and Attention in Visual Working Memory

## Introduction

How does task-relevance affect working memory representations? The concepts of task relevance and selective attention are inextricably linked, in that selective attention is defined as the prioritization of information relevant to the task at hand. Selective attention is clearly important at encoding (Gazzaley et al 2011; Schmidt et al, 2002). Knowing ahead of time that only a subset of presented information will be tested by an upcoming memory probe enables only those stimuli to be encoded into memory, which is critical given limited working memory capacity (Zanto and Gazzaley 2009). However, task-relevance is not always known ahead of time. Experimentally, retrocues can be used to indicate changes in task-relevance after encoding, allowing the experimenter to examine the effects of selective attention in prioritizing information that has already been encoded into working memory. Prioritization of representations already in working memory has been shown to lead to performance benefits though the mechanisms underlying this benefit remain less clear. Most current models invoke a benefit of sustained attention applied to internal representations (Kiyonaga and Wegner, 2013; Ma et al, 2014) often via mechanisms proposed to operate on external representations, such as the principles of biased competition (Desimone, 1998). Recently, however, Myers, Stokes and Nobre (2017) have argued for a theory of internally directed selective attention in working memory that integrates in a more involved way the concepts of attentional prioritization and task-relevance. Namely, they argue

that the prioritization of WM contents is both an act of selection, coupled with a reformatting of the selected information to best guide the next behavioral action.

Behaviorally, measuring the format that information in the brain takes is nearly impossible. However, recent advances in multivariate imaging techniques have allowed for the assessment of representational fidelity, allowing for the direct testing of this hypothesis.

To examine the interaction between task-relevance and internal selection of representations in working memory, we had participants perform a series of behavioral tasks that varied the task-relevance of direction information, and used a forward encoding model to reconstruct direction information from neural data both before and after a retrocue that prioritized one of the two items encoded into working memory.

# Methods

## Behavioral Study

### *Participants*

Participants included 20 healthy, right-handed adult volunteers recruited from the University of Wisconsin-Madison. All of the participants reported being medically and neurologically healthy with either normal or corrected to normal vision. Participants were compensated for their time.

### *Behavioral Data Acquisition*

Participants were shown two sequential patches of dots, with each patch having a unique combination of color, direction of movement, and speed of dot movement. The parameters associated with these dot patches as well as fixation cross, probe and screen resolution were identical to those outlined in Chapter 3, with the exception that dot speed was not fixed. Instead, this value was randomly selected from a linear space of 360 possible values between 1 degrees of visual angle per second and 10 degrees of visual angle per second, with a minimum separation between stimuli of 40 steps in this linear space. While the stimuli used for each trial varied based on trial type, timing and general object type was consistent throughout the study. Each trial consisted of an inter-trial interval (ITI) (3 s), first stimulus period (S1) (1 s), inter-stimulus interval (ISI) (0.25 s), second stimulus period (S2) (1 s), first delay period (D1) (1.5 s), cue period (1 s), second delay period (D2) (1.5 s), probe period (3.5 s) and feedback period (Fb) (0.25 s) (Figure 1b, but with the timings listed above).

Participants completed a total of five trial types. In all trials, unless otherwise noted, participants used speed of dot movement as the cue and reported color on a color bar (again, see chapter 3 for more details). In the first trial type ("Absent"), direction was completely absent from the stimuli (i.e. dots were moving with zero coherence). For the second trial type ("Irrelevant"), 100% coherent motion in a single direction was present in each of the sample stimuli but not in the cue (which was still 0% coherent motion), making it irrelevant to the task. In the third trial type (Cue), participants used direction as a cue instead of the speed at which the dots were moving. In the fourth trial type, the speed of the sample dot patches was used to cue, but the cue patch, rather than being 0% coherent motion, instead took on either the direction of the item whose speed was being cued ("Stroop-Congruent"; 50% of trials) or the direction of the uncued item ("Stroop-Incongruent"; 50% of trials). In the fifth trial type ("Remember"), participants were instructed to remember and report the dots' direction of movement instead of color. The probe for these trials changed accordingly to the direction probe used in Chapter 3.

Participants completed a total of 200 trials, 50 of each trial type. Block order was one of two configurations, split evenly and counterbalanced across participants (Absent, Irrelevant, Cue, Stroop, Remember or Absent, Irrelevant, Stroop, Cue, Remember). The order was designed so that the salience of direction would be roughly increasing throughout the study.

Feedback, written and numerical, was given after each trial. Numerical feedback was the number of degrees between the target color/direction and the reported color/direction. Written feedback consisted of one of three words "Great" (within 15° of target), "Okay" (between 16 and 40°), or "Poor" (greater than 40°). Precision (degrees from target) and reaction times (RTs) were recorded.

## fMRI Study

### *Participants*

Participants included 10 healthy, right-handed adult volunteers recruited from the University of Wisconsin-Madison. All of the participants were screened for neurological and psychiatric disorders and to determine their ability to undergo magnetic resonance imaging. Participants were compensated for their time.

### *fMRI Data Acquisition*

The fMRI data for each participant were collected over two study sessions, both were performed while being scanned in a 3-T MRI scanner. The purpose of the first and second scan was to collect the data necessary to test and train the inverted encoding model, respectively. The first, testing, session was a modified version of the behavioral study described above (Figure 1b). Modifications were made to the timing of the trials as to allow for detection of BOLD changes in spite of the inherent lag in the hemodynamic response. The length of the ITI, D1, and D2 were increased to seven seconds each. The second, training, session consisted of participants viewing and reporting the direction of movement of a patch of moving, grey dots (Figure 1B). The directions used in the training scan were the same as those in the testing scan. We had participants complete the training scan after the testing scan explicitly not to tip them off that direction was the feature of interest in the study.

### *Behavioral Data Analysis*

To analyze the behavioral data we used a three-factor mixture model (Bays, Catalao, & Husain, 2009). The model was used to discern between trials in which participants were

responding to the target, responding to the non-target (binding error), or simply guessing. These values are reported as proportions. Again, see description in Chapter 3.

*fMRI Data Analysis*

Functional magnetic resonance imaging data was processed and analyzed just as detailed in Chapter 3, except no masks were generated for color data, and only direction forward encoded models were constructed. The testing scan of each participant was aligned to the training scan. Additionally, the methods for Chapter 3 detail that we trained the forward encoding model using the latest possible delay-period time-point (TR) from the training task (corresponding to a time in which the sample-evoked bold signal had returned to baseline). This was done to assess the representational fidelity of activity associated with the late maintenance period, instead of direct sample stimulus-processing per se. Because we wanted to assess different possible representational formats, we trained an additional forward model for motion direction, which was identical to that of the "late-delay" model in all aspects, save that it was constructed from data from the time-point in the training task of maximal bold excursion in response to the sample stimulus (TR 4). This model will hereafter be referred to as the "sample-evoked" model.

*Family-wise error rate correction*

We used the same method of quantifying reconstructions from individual time-points indicated in the methods of chapter 3, however, we also corrected for family-wise error rates in the following way. Data from "absent" trials represent an effective negative control for quantifying the model's expected type I error, because no direction information was present on those trials. To quantify the false discovery rate we built the forward models as normal, applied

them to each timepoint from “Absent” trials, and then aligned them along a randomly generated list of directions (because there was no “true” list of directions to align them along). Data from one of these iterations are shown in figures 3 and 4. Once aligned, the reconstructions were fit as normal and quantified using the subject-level bootstrap procedure outlined in Chapter 3. A reconstruction was considered robust for a positive amplitude if < 5% of bootstrap iterations produced negative amplitude estimates. Similarly, a reconstruction was considered robust for a negative amplitude if < 5% of bootstrapped iterations produced positive amplitude estimates. The number of significant reconstructions (positive or negative) was noted, along with the largest number of consecutive “robust” time-points. This process was repeated 2500 times, generating a percentage of robust reconstructions and a maximum cluster-size. A cluster size of 3 appeared on < 1% of iterations of this procedure, which we selected as an appropriate cluster threshold.

# Results

## Behavioral Study

A one-way ANOVA revealed that there were no differences in proportion of target responses between conditions, $F(4,95) = 2.43$, $p = 0.053$ (Figure 2). Participants correctly responded to the target similarly when direction was absent ($M = 0.863$, $SD = 0.117$), irrelevant ($M = 0.823$, $SD = 0.114$), used as the cue ($M = 0.885$, $SD = 0.105$), "stroop-like" ($M = 0.8719$, $SD = 0.0796$) or to be remembered ($M = 0.797$, $SD = 0.110$).

An additional one-way ANOVA showed that there were no differences in proportion of binding errors between conditions, $F(4,95) = 2.33$, $p = 0.061$ (Figure 2). Participants responded to the non-target equally as frequently when direction was absent ($M = 0.106$, $SD = 0.107$), irrelevant ($M = 0.124$, $SD = 0.090$), used as the cue ($M = 0.062$, $SD = 0.064$), "stroop-like" ($M = 0.0828$, $SD = 0.058$), or to be remembered ($M = 0.115$, $SD = 0.075$).

A third one-way ANOVA showed that there were no differences in proportion of guesses between conditions, $F(4,95) = 1.83$, $p = 0.129$ (Figure 2). Participants responded with a guess in like proportions when direction was absent ($M = 0.031$, $SD = 0.032$), irrelevant ($M = 0.053$, $SD = 0.087$), used as the cue ($M = 0.053$, $SD = 0.078$), "stroop-like" ($M = 0.0453$, $SD = 0.0451$), or to be remembered ($M = 0.088$, $SD = 0.088$).

A final one-way ANOVA of Precision showed that while there was a significant difference between conditions, that this difference was driven entirely by the difference in precision between reporting color and reporting direction. When the remember condition was removed no significant differences were found, $F(3,76) = 0.871$, $p = 0.46$ (Figure 2). Participants evinced similar precision when direction was absent ($M = 8.6187$, $SD = 2.99$), irrelevant ($M = 8.7913$, $SD = 4.63$), used as the cue ($M = 7.37$, $SD = 2.98$), or "stroop-like" ($M = 7.67$, $SD = 2.37$).

While we did expect to see differences for the stroop condition relative to the other trials, it is perhaps not surprising that we found no difference between trial types given the poor performance on color trials in general. Across all three tasks reported in this work, participants showed poor color performance. This is likely due to a poor choice of region of color space to use.

**fMRI study**

*Encoding Model Results – "Absent" Trials*

The first condition of the experiment (the "Absent" trials) served as a negative control, because no unique direction information was present on the trial, in that dots moved with 0% coherent motion. Therefore, the ability to reconstruct direction (i.e. find a robust amplitude parameter, either positive or negative), on these trials can only be due to model type I error. To generate "cued" and "uncued" plots, reconstructions were aligned along two randomly generated directions that followed the same rules used to generate the actual directions used for future trials (namely that the two directions had to be at least 40 degrees apart). Reconstruction time-courses for each ROI were generated for a model trained on a time-point (TR) from the "Training" task reflecting the maximal sample-evoked BOLD response (TR = 4; fig 4,A-E) and for a model trained on a time-point corresponding to mid/late delay activity (TR 6 of the training phase; fig 3,A-E). Black stars indicate reconstructions robust for positive amplitudes, while white stars indicate reconstructions robust for negative amplitudes (i.e. inverted tuning). Insets for both figures (G-L) show individual time-point reconstructions of the cue onset and probe onset (which, due to hemodynamic lag, thus reflect activity prior to and as a consequence of, the cue, respectively). While there aren't many significant reconstructions, there are more than would be

expected by chance (~17% instead of 5% expected rate of false discovery). To account for this, we implemented a cluster correction procedure (see methods). Results are interpreted here in light of a cluster threshold of < 3 consecutive time-points (TRs). Level 1 significance is indicated on plots for descriptive purposes. Additionally, it can be noted that most of the false positive results occur during the first 2 and last 2 TRs of each trial, which are time-points whose reconstructions reflect activity during the inter-trial interval (ITI). This is similar to the findings of the Study detailed in chapter 2, where decoding was noisiest during the ITI. Because these time-points were not of theoretical interest a priori, their reconstructions will not be discussed further, however, in the interest of transparency, data are still included in the figures.

*Encoding Model Results – "Remember" Trials*

As a positive control, (and a condition most analogous to the study discussed in chapter 3), we attempted to reconstruct motion direction on trials where participants were required to maintain and subsequently recall motion direction. Figures 5 and 6 show reconstructions for models trained, respectively, on the "delay-period" (TR 6) and "sample-evoked" (TR 4) time-points from the training task. Similar to findings in Chapter 3, reconstructions trained on the "delay-period" TR are quite robust in the Occipital ROI (figs 5E and 5F). Additionally, both the cued and uncued directions can be robustly reconstructed prior to cue onset (3K), but after the cue, only the cued direction remains reliable (5L). And, similar to the findings of Chapter 3, these "delay-period"-trained models fail to reconstruct in parietal (fig 5 C,D) or frontal (fig 5 A,B) regions. For models trained with the "sample-evoked" TR, the Occipital ROI shows only transient reconstructions immediately following the sample presentation (time-points 5 and 6, figure 6 E,F) and probe, though these reconstructions do not survive cluster correction.

Reconstructions again fail in the Frontal ROI (fig 6 A,B) though a significant cluster is observed in parietal cortex immediately following the sample onset, though only for the uncued item, and indicative of an inverted-tuning profile. Because this cluster occurs only for the uncued item at time-points before cue onset (and the distinction between UnCued and Cued becomes evident to participants), it is most likely the case that the profile is present for the Cued direction is well, or for neither. Post-hoc t-tests conducted to assess whether reconstructions for these timepoints were significantly different between the Cued and UnCued directions found no significant difference. Overall, reconstructions of remembered directions are reliable almost exclusively in the Occipital ROI, and models constructed from a "late-delay" period time-point in the training task fare better than those constructed from a "sample-evoked" time-point.

*Encoding Model Results – "Irrelevant" Trials*

For these trials, participants were informed prior to the start of the block that they would be performing a task similar to the "Absent" block except that for each of the two dot patch stimuli they viewed, all the dots would be moving in a single direction. They were explicitly told that they could ignore this direction however, as it was irrelevant to completing the task (reporting the color of the patch whose speed was cued mid-way through the delay period). Because of the result from Chapter 3 that features bound into the same item could still be reconstructed late into the delay after a retrocue informed participants that only one of the features would be relevant for an upcoming recall probe (along with the results of many behavioral studies discussed in Chapter One), we hypothesized that despite being told that they could ignore direction, participants would still encode motion direction, though perhaps to a lesser extent. Instead, we were completely unable to reconstruct motion direction from any area

during the delay period, regardless of whether we trained on the "sample-evoked" (fig 8) or "late-delay" period (fig 7) TRs from the Training Task, implying that participants were able to completely filter out irrelevant information.

*Encoding Model Results – "Cue" Trials*

These trials were nearly identical to "irrelevant" trials, except that direction, instead of speed, was used to cue participants which color to report. We hypothesized that because participants did not have to recall the precise direction, but only had to use the information to distinguish between two color memoranda in response to the retrocue, that the information would be held in a less precise state. In response to the retrocue, as expected, for a model trained on "late-delay", the cued item's direction can be robustly reconstructed from Occipital (~ time-point 9, fig 9E), but not the uncued (fig 9f). This pattern also holds true leading up to the retrocue (when the distinction between "cued" and "uncued" hasn't yet been made known to the participant, so the inability to resolve the "uncued" direction here is likely just noise. Indeed, a post-hoc bayes factor calculation at this timepoint between amplitude estimates for the two reconstructions reveals a factor of 4.1:1 in favor of the null hypothesis that they are not different. Additionally, it should be noted that the cue, onsetting at TR 7 in these plots, itself contains direction information (because direction is the cue). Therefore, if this underlying activity reflected only processing of the cued stimulus, we would expect robust reconstructions for the model trained on the "sample-evoked" time-point from the training task in the Occipital ROI. Instead, only transient robust reconstructions were found (TR 9; fig 10E), with none surviving cluster correction. This indicates that the sustained reconstruction robustness post-cue in the "late-delay" trained model reflects the underlying mnemonic representation-related activity.

Comparing to "Remember" direction trials (fig 5E), it becomes apparent that while the cued direction can be robustly reconstructed post-retrocue both when that direction is the thing to be reported and when that direction is used to cue a color to be reported, the time-courses are different. Reconstructions in "cue" trials are briefer, becoming non-robust at TR 11 (probe-onset), whereas reconstructions for "remember" trials last well into the probe-evoked response time-points (TR 13). This is expected given the differences in task-relevance between the conditions: direction is needed to respond to the probe in the "remember" condition, but needed only to select the correct color memoranda in response to the retrocue in the "cue" condition. Additionally, in order to test the hypothesis that reconstructions of direction on "remember" trials would be more precise (or at least more robust), than on "cue" trials, we conducted a post-hoc two-tailed t-test for both the precision and amplitude estimates of the reconstructions at TRs 7,9 and 11. To our surprise, significant differences were found only for TR 11 for the amplitude estimate ($p=0.023$), though it is likely that we were not powered to find a difference given the noisiness of the estimates. Finally, though in the Occipital ROI we failed to robustly reconstruct direction in the "cue" condition for the "sample-evoked" time-point trained model, the Parietal ROI showed a significant reconstruction cluster for the cued direction post-retrocue (fig. 10C). Because the reconstructions observed in the Parietal ROI for the "sample-evoked" model are not robust before the retrocue, nor are they robust for at any point for a "late-delay" trained model (fig. 9C), this implies that this activity reflects cue stimulus processing (see discussion).

*Encoding Model Reconstructions – "Stroop" Trials*

In the "irrelevant" trials, participants were instructed that they could ignore the direction information present in the stimulus, but there was no direct task-related disincentive to encoding

it. We therefore had participants complete blocks of trials where the relationship between the stimuli and the cue varied in a "Stroop"-like fashion, in that while the speed of the cue was 100% informative and congruent with the stimulus whose color was to be tested, the direction of the cue was incongruent (ie was the direction associated with the *other stimulus*) on 50% of trials. This meant that if participants encoded the direction of the stimuli, this information could interfere with cue-processing mediated selection of the correct stimulus for report on the 50% of trials where the direction was incongruent, thus giving participants an incentive for suppressing direction-related information. Trials where the cue was congruent were analyzed separately from trials where the cue was incongruent, to best observe differences in cue processing. Though treated separately, during the course of the experiment participants experienced these trials intermixed within blocks.

For "stroop-congruent" trials, in the Occipital ROI, across both models, no robust reconstructions were observed that survived cluster correction (figs 11E,11F,12E,12F). This is most similar to the pattern observed for "irrelevant" trials (figs. 7E,7F,8E,8F) and reaffirms the interpretation that the reconstructions observed post-retrocue in the Occipital ROI for "cue" trials reflect active maintenance of direction information (fig 9E). Where these trials differ from the "irrelevant" trials is in the reconstruction time-courses for Parietal and Frontal ROIs. As stated previously, we hypothesized that in order to correctly navigate the "Stroop"-like condition, participants would need to actively suppress direction information present in both the stimuli and the cue, increasing the likelihood that fronto-parietal networks would be recruited to accomplish this. Consistent with this hypothesis, we observed significant inverted clusters (negative amplitudes) for reconstructions of the cued direction in both Frontal and Parietal ROIs (fig. 11A,12A,12C). For "late-delay" trained models, this cluster was observed following probe onset

(fig 11A) and for "sample-evoked" models, clusters were observed post-retrocue leading into the probe onset (fig 12A,12C). The implications of this result are addressed in the Discussion below.

For "stroop-incongruent" trials, a similar pattern is observed (figs 13,14). In the Parietal ROI, a significant cluster of robust inverted reconstructions is present for time-points surrounding probe onset when aligning along the cued direction for a "sample-evoked" model (fig 14C). When the same data are aligned along the UnCued direction (which, for these trials is the direction information presented in the cue), a significant cluster of robust *positive*-amplitude reconstructions can be seen. For a comparison of one time-point in the center of these two clusters see figure 14J. The results from this same model in the Frontal ROI show the same general pattern, though the robust reconstructions never quite aggregate into a significant cluster (fig 14A,B).

## Discussion

Building off the results of the experiment in Chapter 3, we used a forward encoding model approach to reconstruct the neural representation of motion-direction across a number of different task conditions for which the behavioral relevance of motion-direction information was varied. The concept of task-relevance is central to the definition of selective-attention, in that what information is selected for prioritized processing is that which is most task-relevant. Traditionally, single-unit investigations into selective attention during sustained attention tasks have operationalized task-relevance, and thus attention, as a binary (Martinez Trujillo and Treue 2004; Egly et al, 1994; Chelazzi et al 1998). While it is widely-acknowledged that for complex tasks such as working memory, the task-relevance and thus attentional prioritization of different

kinds of information can vary in a more graded fashion (reviewed in LaRocque et al, 2014), it is less clear whether the same mechanisms of prioritization engaged at encoding subserve task-dependent prioritization during maintenance (Myers, Stokes and Nobre, 2017). In the previous study detailed in Chapter 3, encoding was held constant, in that participants were required to encode two objects (dot patches), each of which possessed a unique direction of motion and color. When a series of retrocues instructed participants after the fact which of the four features (2 colors and 2 directions) would be relevant for a behavioral report, features that were bound together showed a kind of "stickiness". Specifically, when two features of a single object were still being held in memory, and one of those two features was cued, neural representations of the uncued feature could be reconstructed later into the subsequent delay-period than when one of two features from *different* objects was cued. Here, using nearly identical stimuli (but adding a potential third dimension of speed) we manipulated which features were task-relevant at encoding by blocking together trials with different task instructions. When participants knew that direction information was going to be tested, reconstructions of direction were extremely robust in the Occipital ROI, analogous to when the feature dimension of direction was cued in Chapter 3. Reconstructions were qualitatively more robust here than in Chapter 3, both before and after a retrocue instructing participants which of the two directions held in memory was relevant for an upcoming behavior probe, implying a cost to having encoded both direction AND color in Chapter 3. A direct test of this using the same participants would confirm a prediction of distributed resource models; specifically that the addition of features, even those bound into the same object, costs resources and thus lowers the fidelity of the representations (Bays, 2013). In a deviation from the results of chapter 3 however, when direction information was irrelevant to completing the task at hand (participants knew they were to report color at encoding), direction

information was unable to be reconstructed at any time point from any region. This argues that "feature-binding" is not automatic to all features present in the same spatial location, but is instead under executive control (Jaswal, 2012).

We were unable to reconstruct stimulus specific information for remembered information in Parietal or Frontal areas, consistent with some prior studies (Emrich, Riggall and Postle 2013; Harrison and Tong, 2009) though not others (Bettencourt and Xu, 2016; Ester et al 2015). One potential explanation for the discrepancy in the literature is a difference in task design (ie the presence or absence of distractors). Here, when direction information was used to cue which color was to be reported, a robust reconstruction of direction was observed in Parietal in response to the cue. This implies that parietal cortex represents information necessary to select a representation from memory (cue information) but not the to-be-selected memoranda (in that reconstruction fails post retrocue in "remember" trials). Additionally, when direction information is used as a distractor (see "stroop-like" condition methods), robust reconstructions of the distracting information are observed in both Parietal and Frontal ROIs. Specifically, both areas display an inverted tuning output following the cue when aligned along the direction of the cued object *regardless of what direction is currently displayed on the screen* (compare 13A to 14A and 13C to 14C). One interpretation of inverted tuning reconstructions is that they reflect structured suppression of a particular portion of feature space (Martinez-Trujillo and Treue, 2004). Under this interpretation, inverted reconstructions are observed in frontal and parietal cortex because the direction information in the cue is being actively suppressed. Thus, the role of the information in the task plays a clear role in how that information is represented. It is possible that the representations decoded by Bettencourt and Xu are the temporary instantiation of a representation in parietal cortex to compare with current visual stimulus processing of a cue or

distractor. When we train our model on the "sample-evoked" time-point of the one-item task and reconstruct the cue, we observe a temporary instantiation of that information in Parietal cortex (10C). A second interpretation is that the inverted tuning curves are a product of tuning shifts in feature space (David et al 2008).

The present results demonstrate that regardless of whether information will be used to cue a specific item to report, or itself be reported, the format of the representation is relatively similar in Occipital cortex (figure 5E, figure 9E). However, the functional role that the information serves in a task seems to determine its representation in frontal or parietal cortex. We find no evidence that Parietal or Frontal ROIs represent information that is to be reported, but do find evidence of representations for information used to cue which stimulus to report (figure 10C). This is especially curious given that these reconstructions were obtained from a model that was trained on a one-item task without the need for attentional selection.

**Figures**

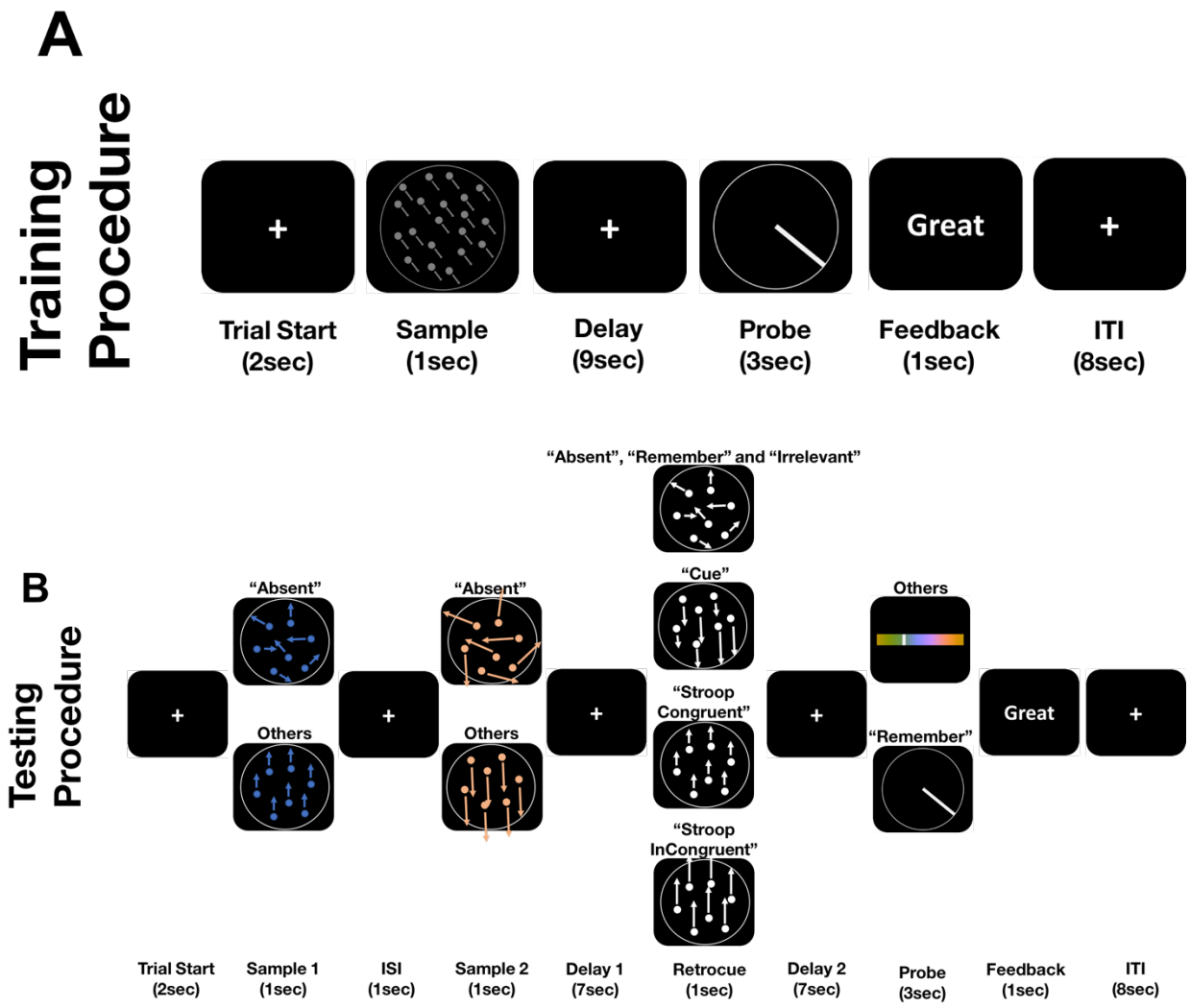



Figure 1: Experimental tasks (A) The training task. Participants completed a one-item memory task for direction of motion aligning a radial probe along the remembered direction (B) The fMRI testing task. Arrow size and direction reflect speed and motion direction respectively. Time proceeds from left to right and where multiple insets exist for a segment, the label indicates what was presented on trials of that condition. 6 unique conditions were presented: "Absent", "Irrelevant", "Cue", "Stroop-Congruent", "Stroop-Incongruent", and "Remember". Each condition was its own block, with the two "Stroop" conditions blocked together. See methods for more details

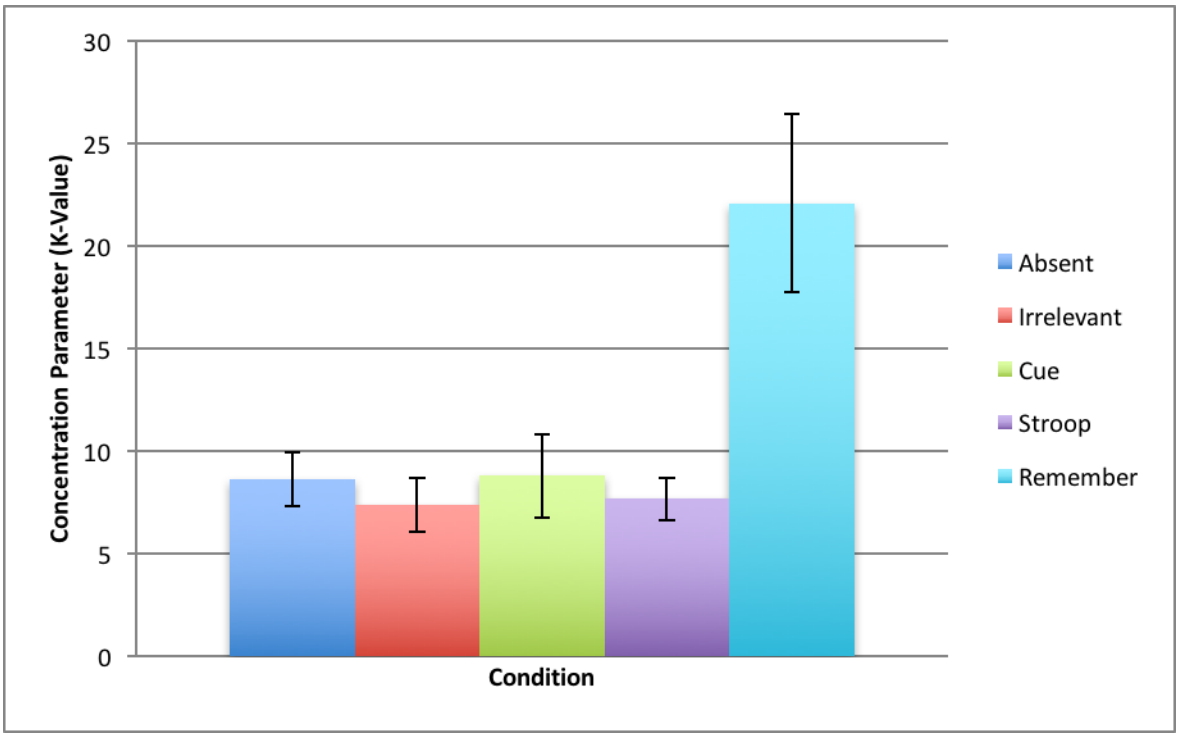
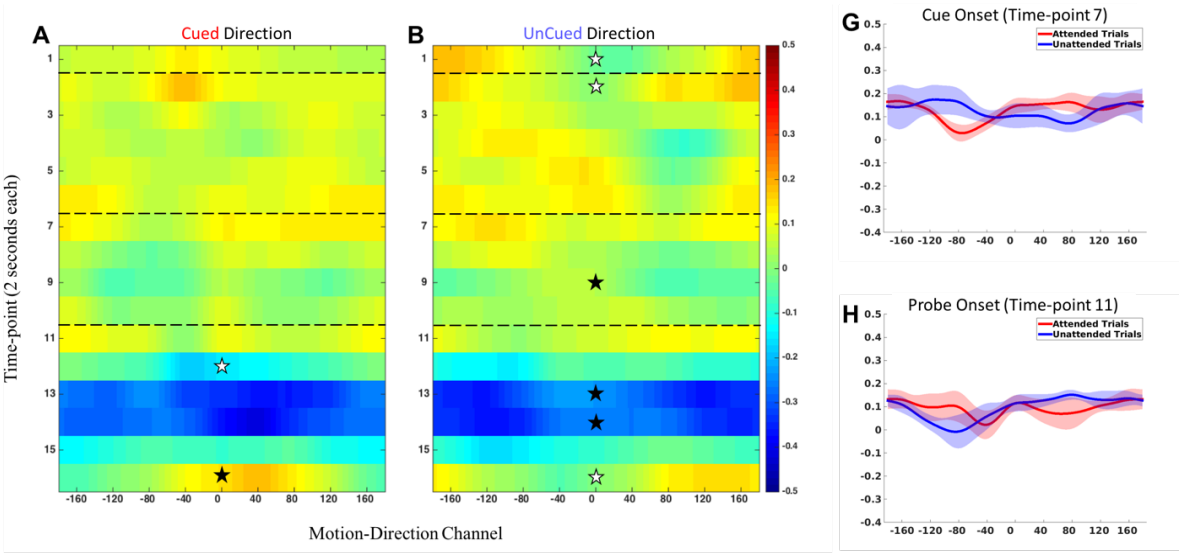
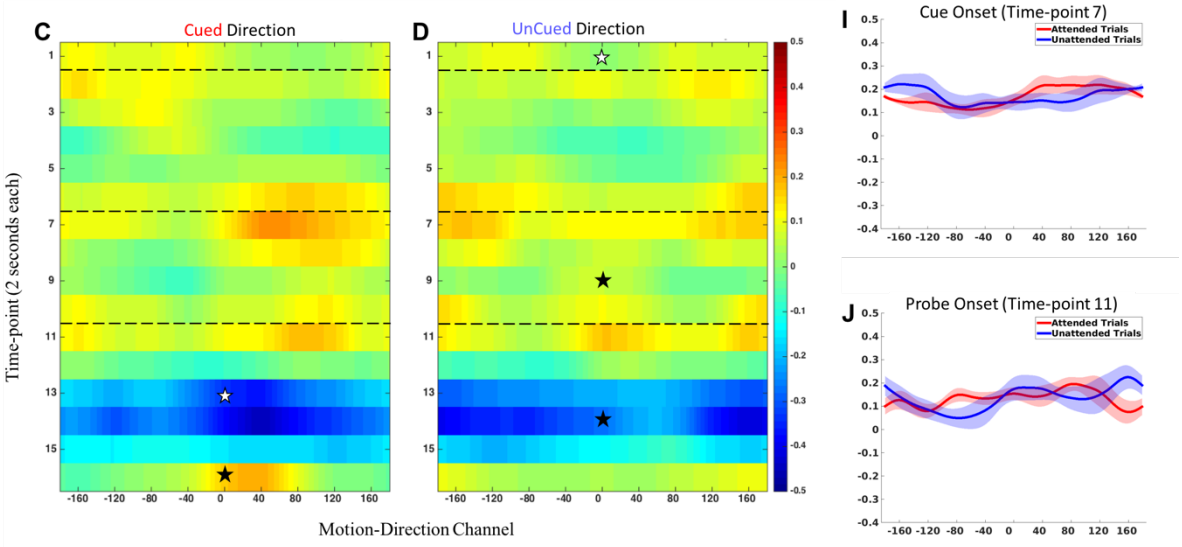



Figure 2: Results from a 3-factor mixture model applied to data from the behavioral study. Participants showed greater precision for direction trials and roughly equal precision across all other conditions. Errors bars represent +/- S.D. Participants responded with equal proportions of target responses, non-target responses and guesses across conditions (not shown).

Frontal ROI




Parietal ROI




Occipital ROI

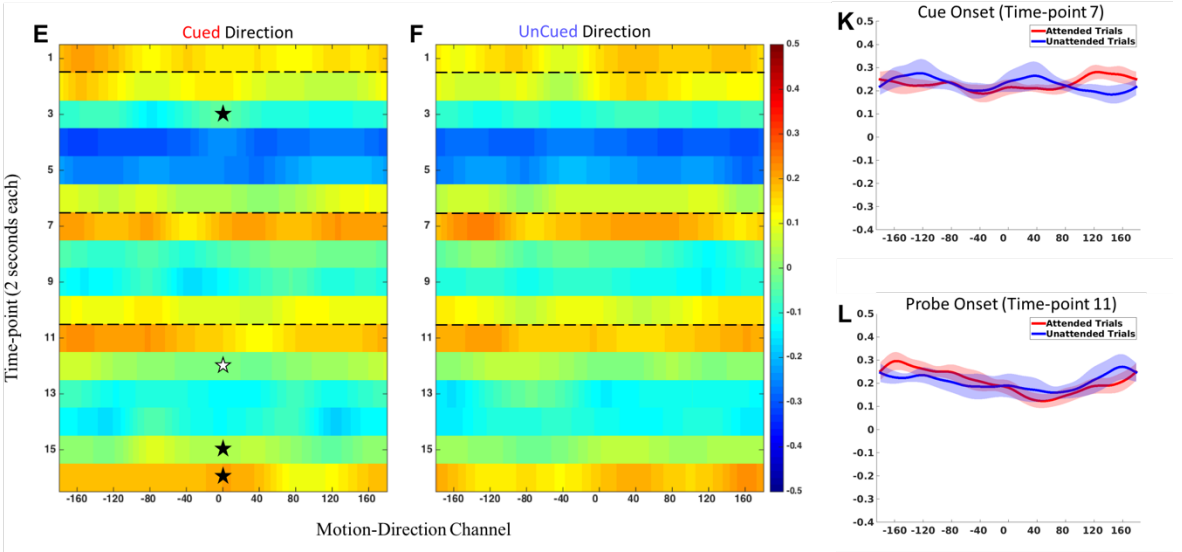


**Figure 3 – "Absent" Trial Reconstructions, "late-delay" trained model.**
Frontal (A-B), Parietal (C,D) and Occipital ROI (E,F) reconstruction time-courses. Y-axis represents time-points (TRs) of a single trial. X-axis represents the motion-direction channels, centered at 0. Color indicates the output of the channel in arbitrary units. Black stars indicate robust positive-amplitude reconstructions, white stars indicate robust negative-amplitude reconstructions (for a description of "robust" see methods). Insets (G-L) plot single TR reconstructions for timepoints 7 and 11. Y-axis is channel output, and x-axis is motion direction channel.

Frontal ROI

Parietal ROI

Occipital ROI

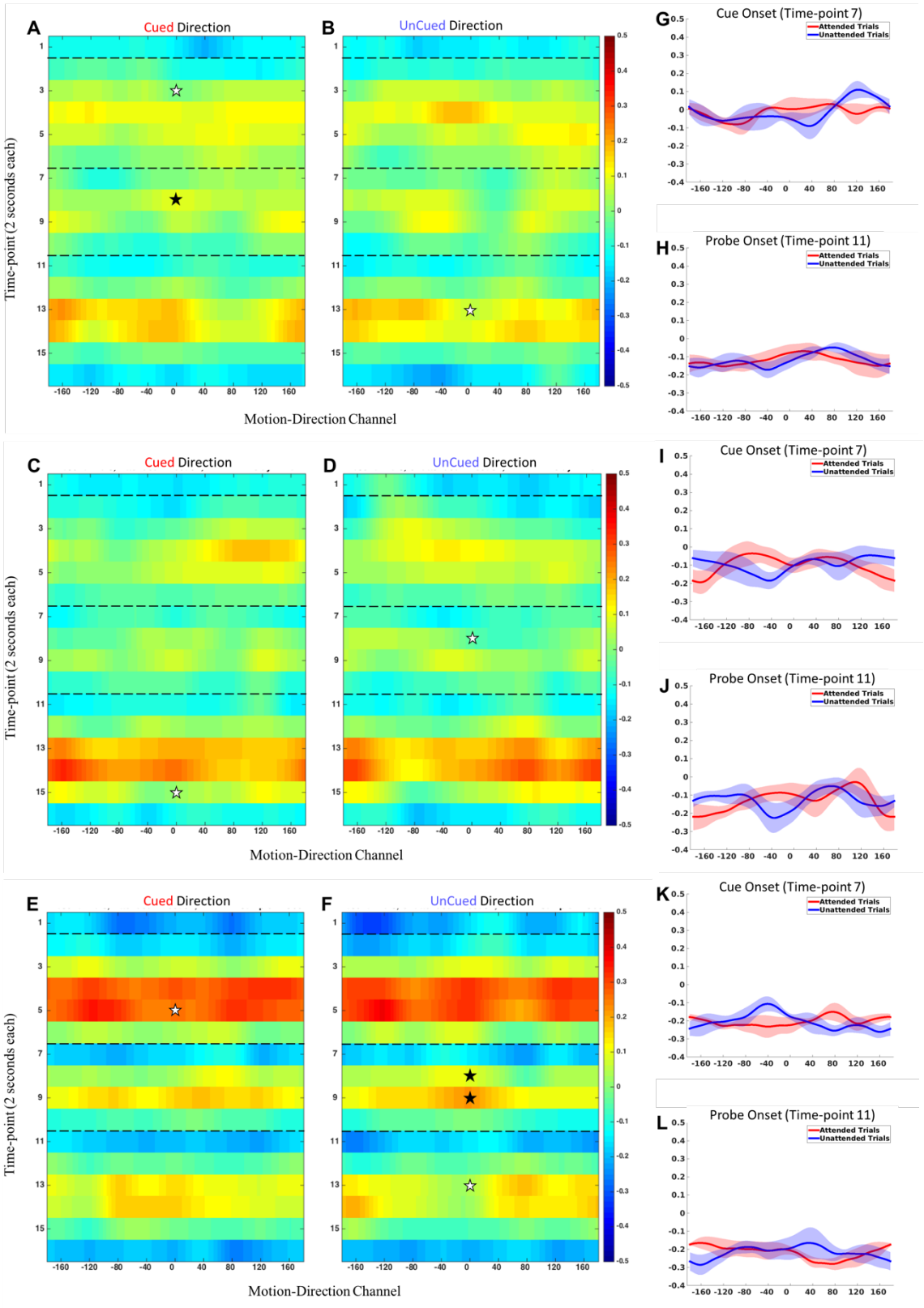



**Figure 4 - "Absent" Trial Reconstructions, "sample-evoked" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

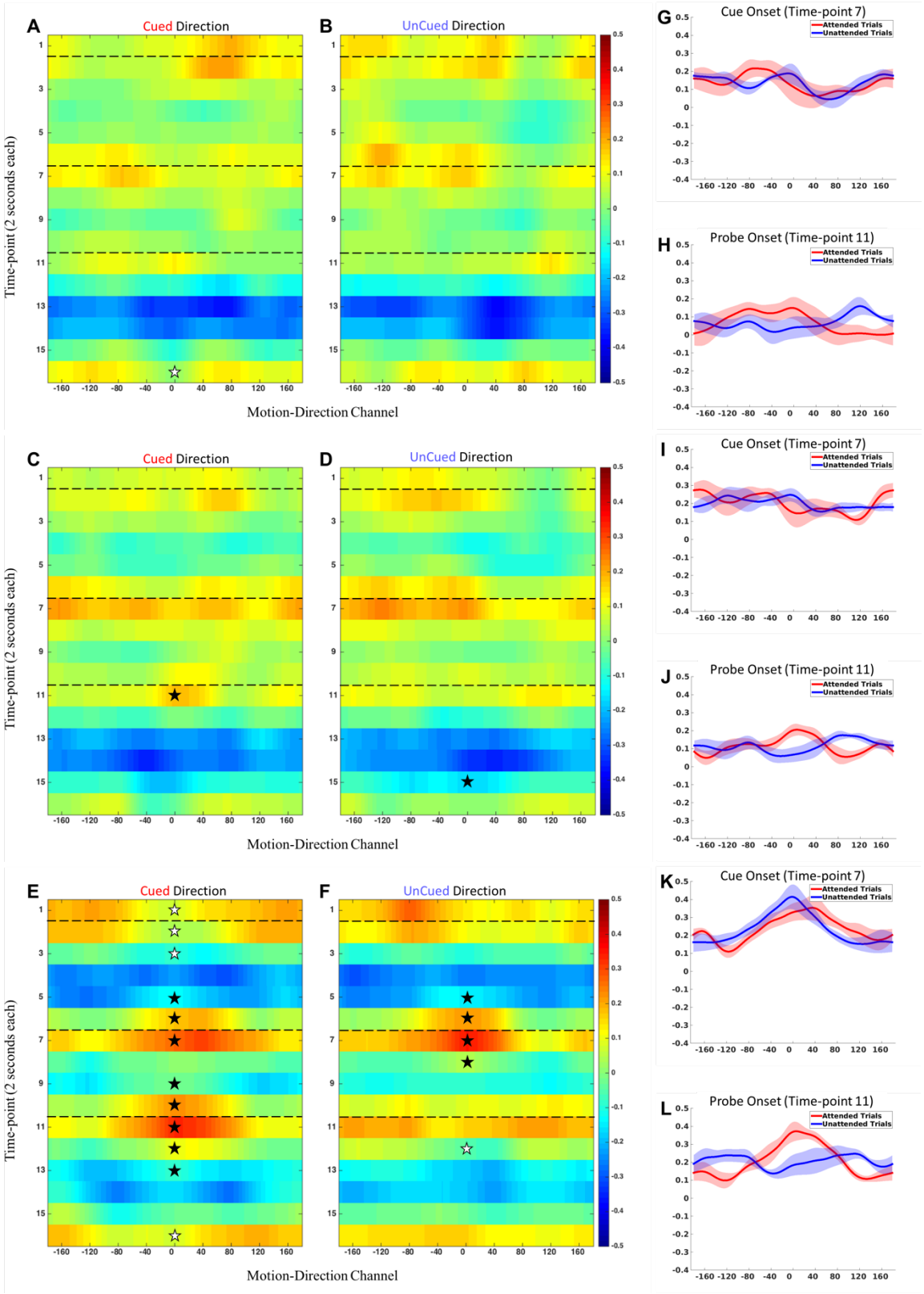



**Figure 5 - "Remember" Trial Reconstructions, "late-delay" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

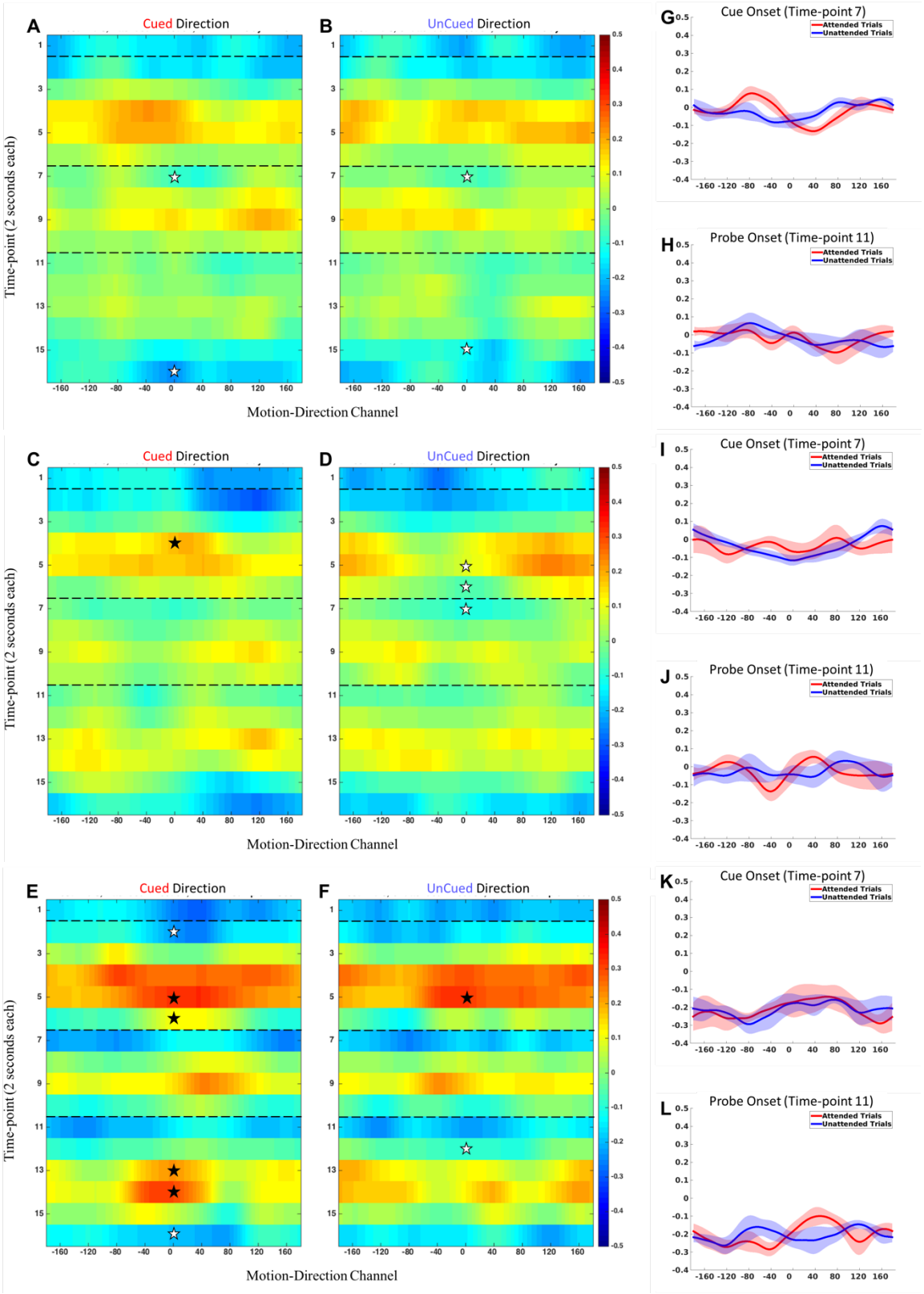



**Figure 6 – "Remember" Trial Reconstructions, "sample-evoked" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

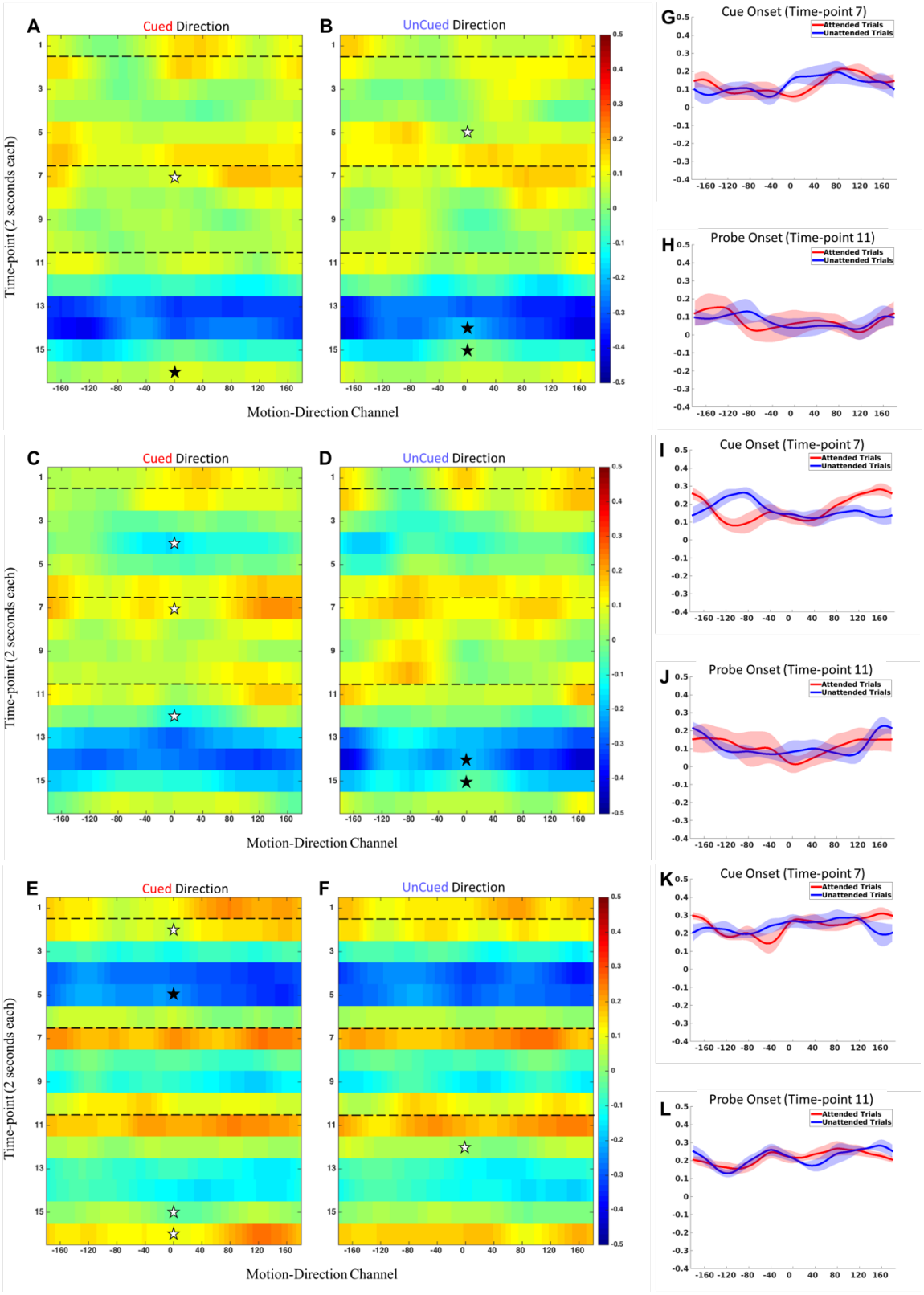



**Fig 7 – "Irrelevant" Trial Reconstructions, "late-delay" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

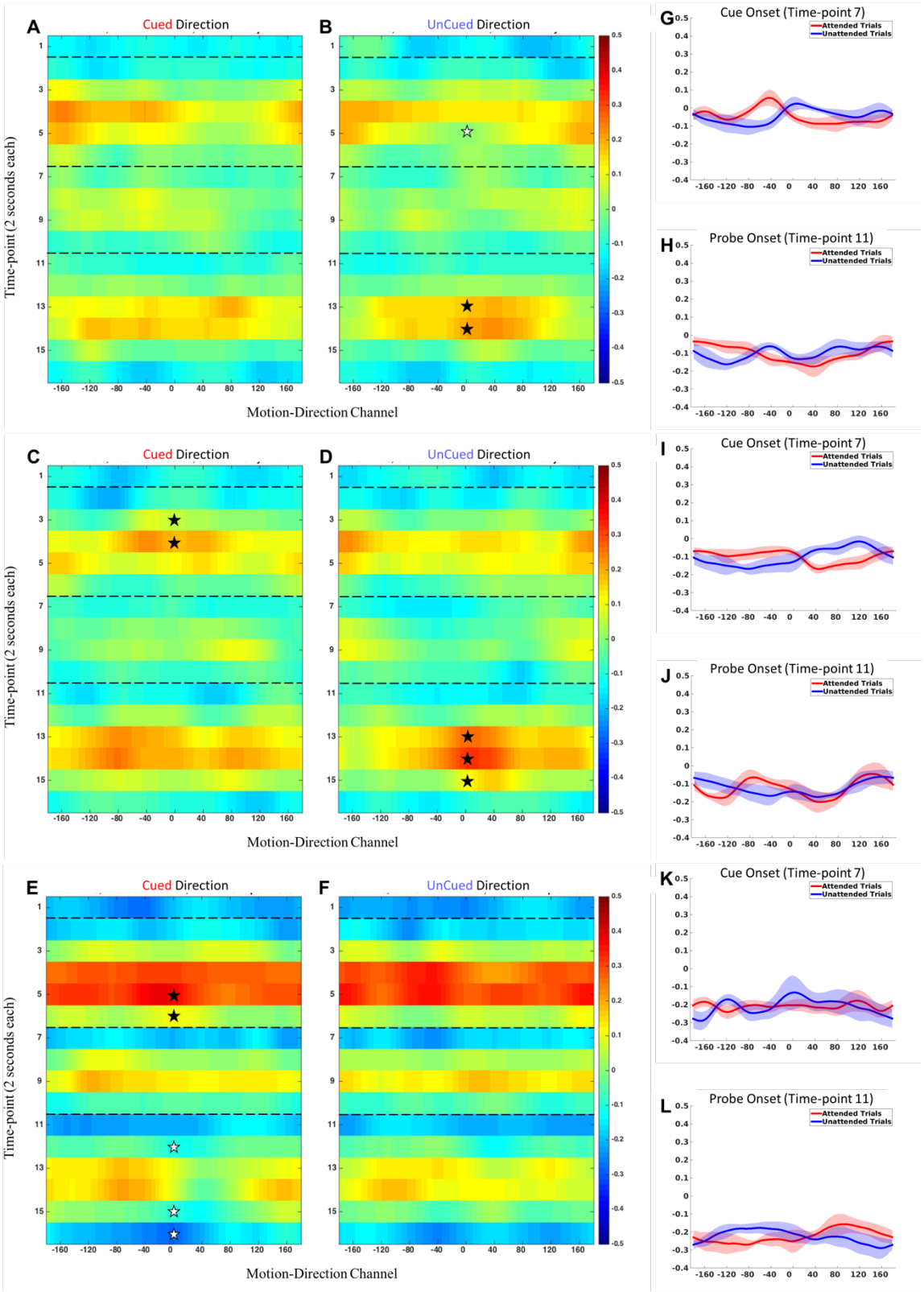



**Fig 8 – "Irrelevant" Trial Reconstructions, "sample-evoked" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

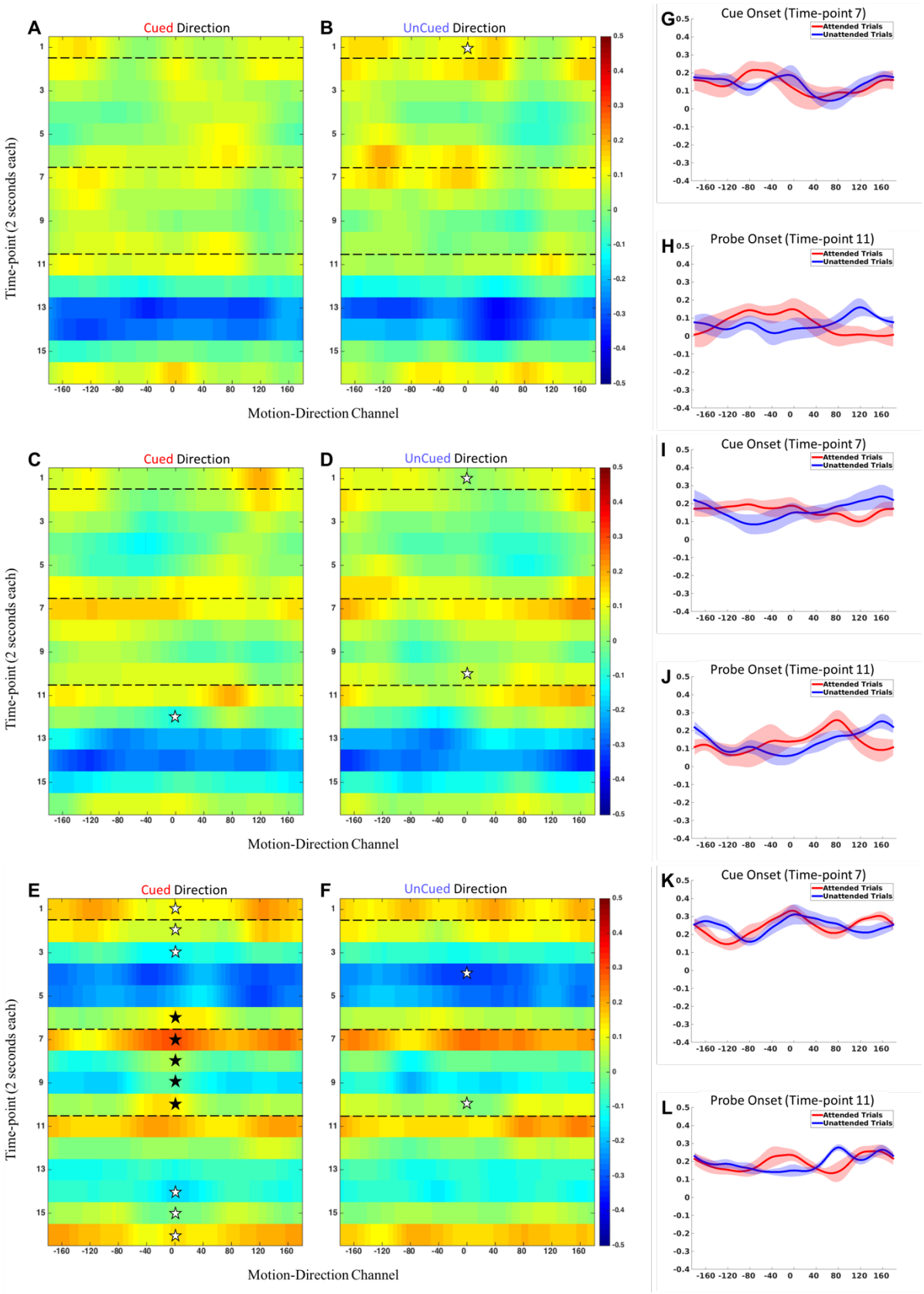



**Figure 9 - "Cue" Trial Reconstructions, "late-delay" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

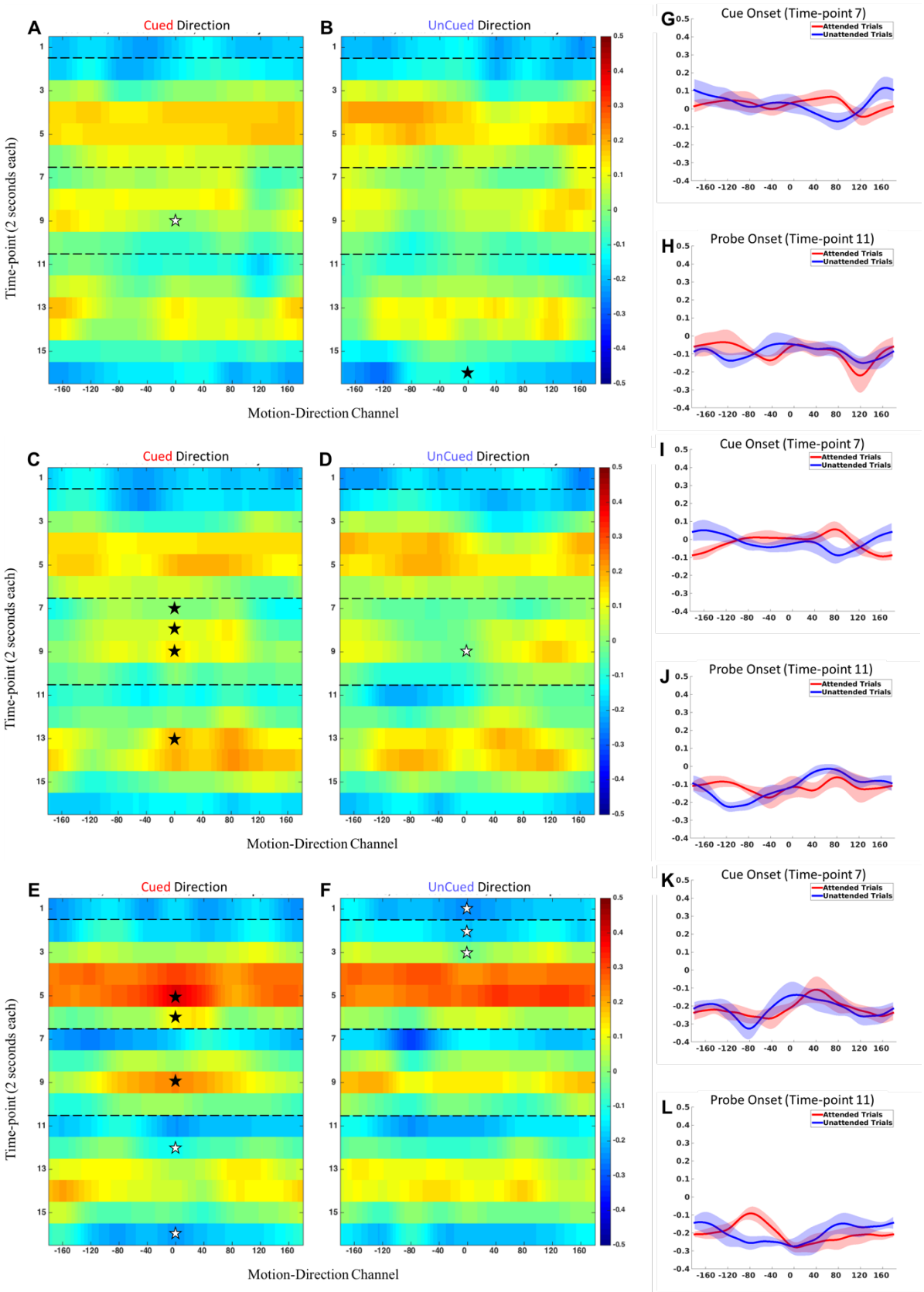



**Figure 10 - "Cue" Trial Reconstructions, "sample-evoked" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

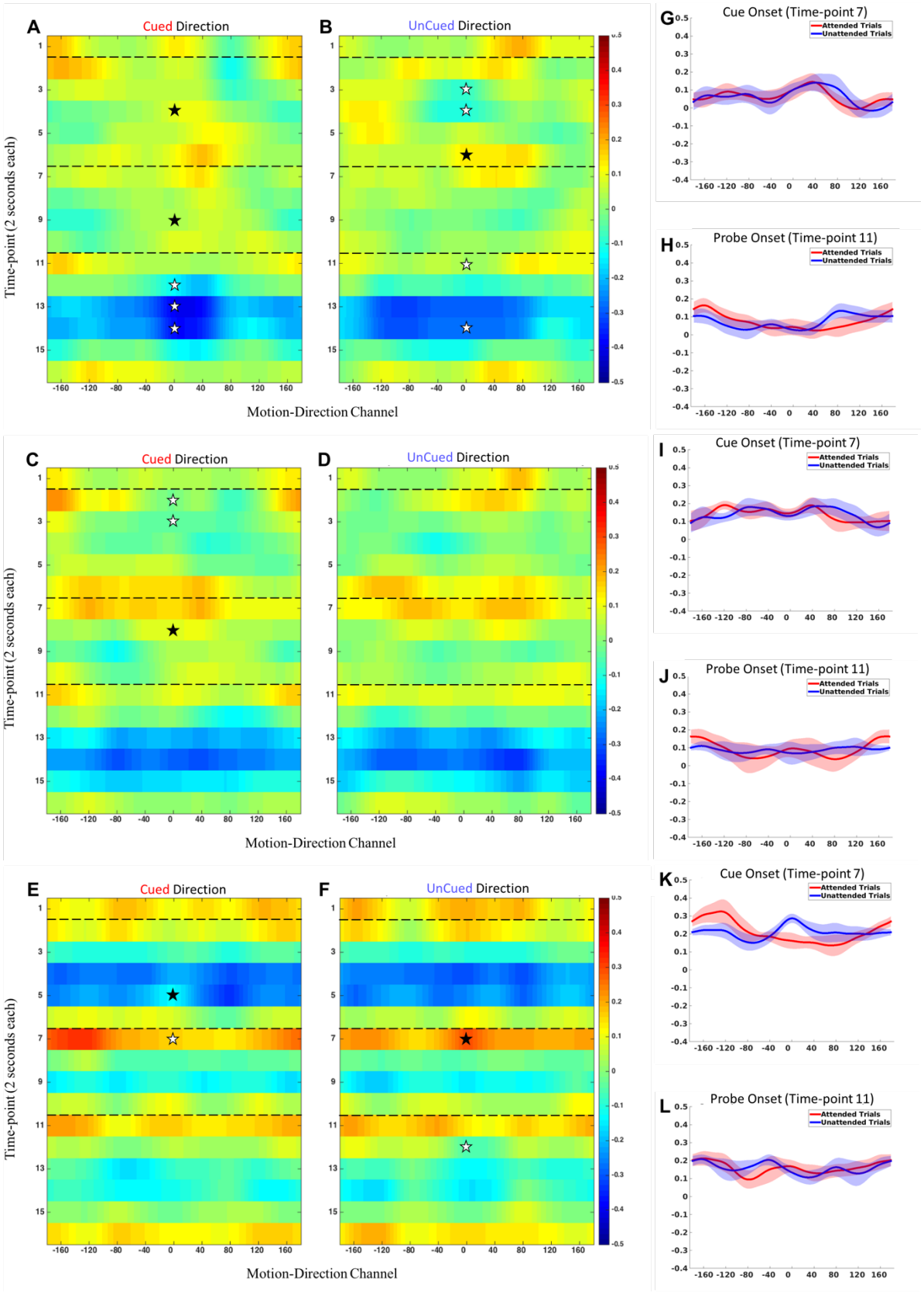



**Figure 11 - "Stroop Congruent" Trial Reconstructions, "late-delay" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

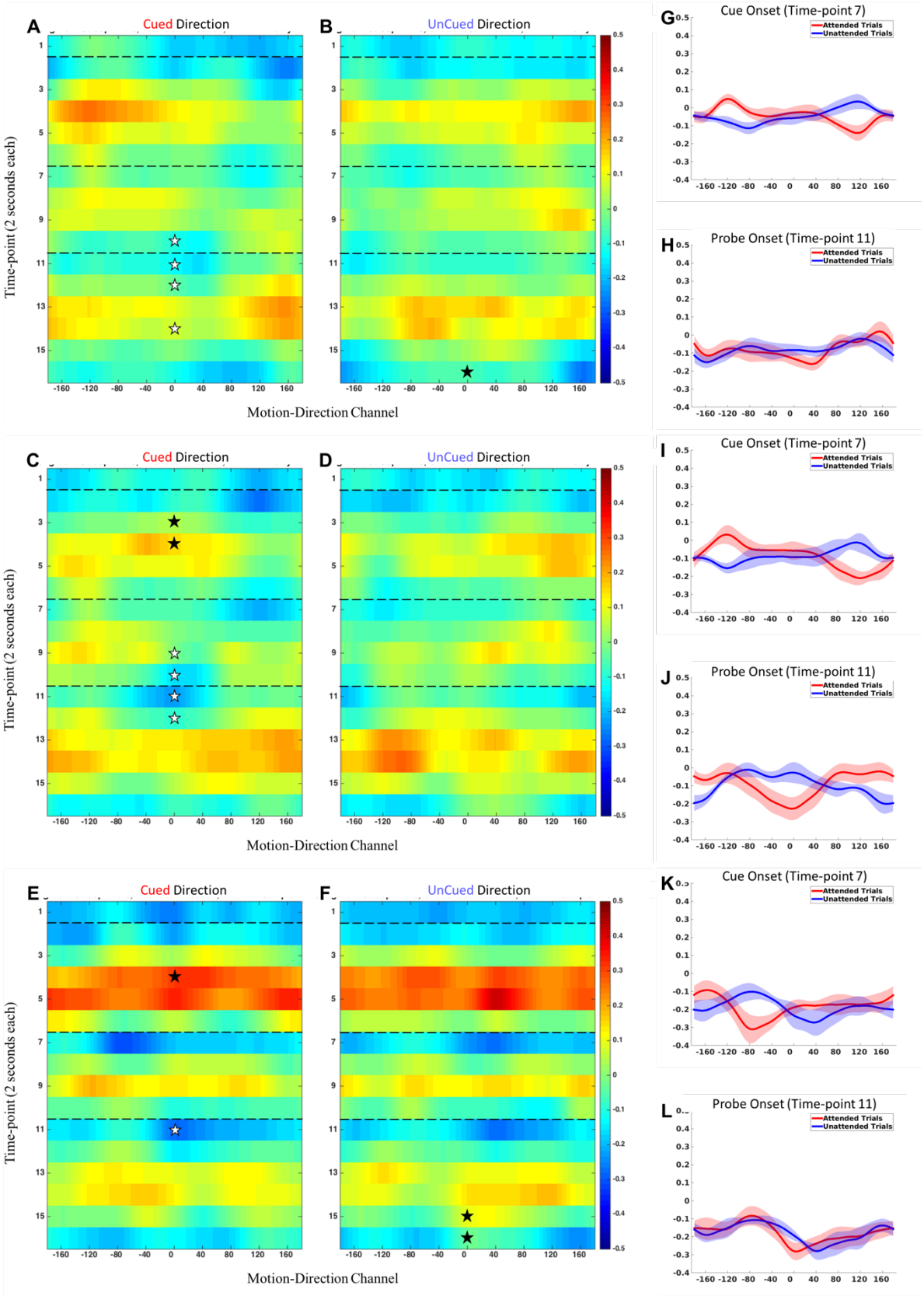



**Figure 12 - "Stroop Congruent" Trial Reconstructions, "sample-evoked" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

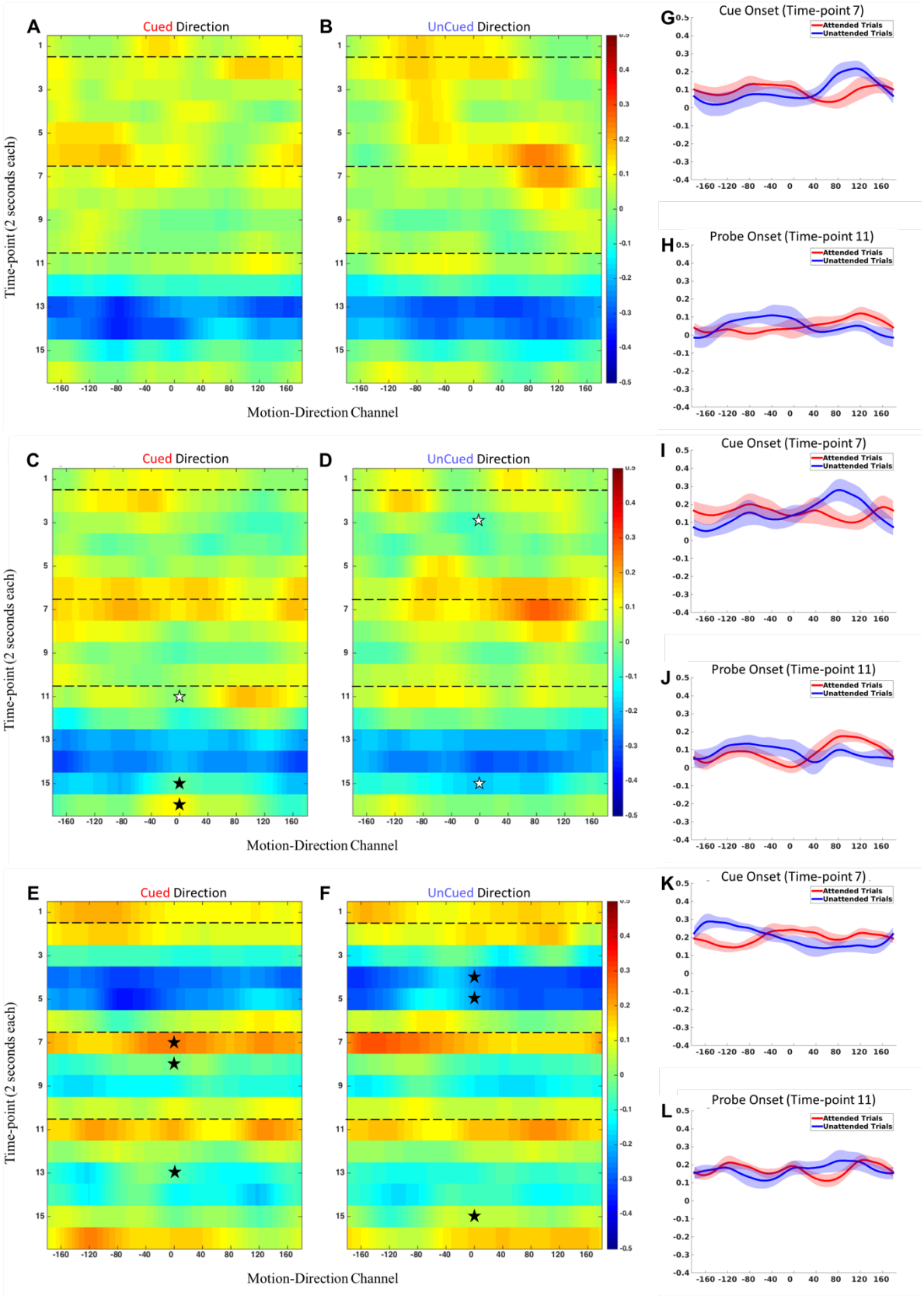



**Figure 13 - "Stroop Incongruent" Trial Reconstructions, "late-delay" trained model.**

Frontal ROI

Parietal ROI

Occipital ROI

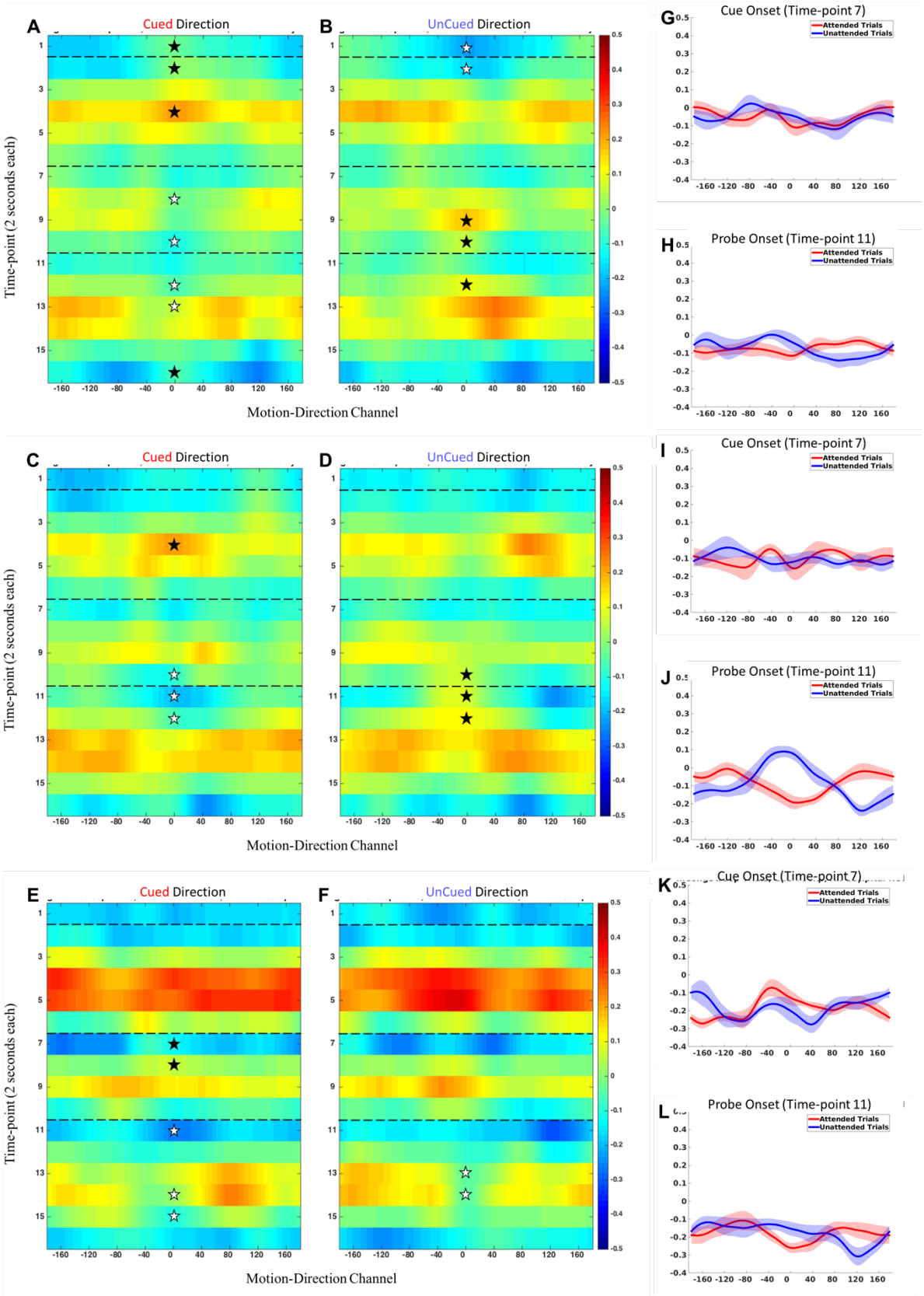



**Figure 14 - "Stroop Incongruent" Trial Reconstructions, "sample-evoked" trained model.**

# Concluding Remarks

The current work utilizes multivariate methods to explore common mechanisms underlying the prioritization of external representations (selective attention) and internal representations. In the first study, principles of biased competition were observed to apply at the population level, enhancing attended representations and suppressing unattended information while both are still present in the environment. The effects of attentional action were qualitatively different in visual cortex when compared to frontal and parietal areas. A multivariate connectivity metric was used to demonstrate an active role for frontal regions in suppressing unattended information at the level of the population representation.

In the second study, attentional prioritization of internal representations in working memory was investigated. Specifically, a triple retrocue task demonstrated remarkable flexibility of representations in visual cortex, long after initial encoding had taken place. Participants were able to navigate the task with no noticeable difference in behavioral precision between navigating object-based or feature-based cues. Despite this, unique neural signatures were observed for the two types of attentional prioritization, with a characteristic inverted reconstruction observed in response to suppression of an item. When the information could not be fully dropped however (because it might be relevant later) or when attention had to prioritize one aspect of a bound feature over another, the uncued feature could still be reconstructed late into the delay. This implies that shifts in prioritization for internal representations does not necessarily need to be zero-sum.

In the third experiment, the relationship between task-relevance and attentional prioritization was examined. Specifically, the very act of "cueing" means that some information retained across a working memory delay is the information that needs to be selected, while some information is held to guide the selection itself. Using forward model reconstructions we demonstrate that frontal and parietal areas preferentially represent information of the latter sort. These representations of task-rule information can also be suppressed if necessary, with the same neural signature of suppression observed in experiment 2.

Future work is still needed to fully tease apart the distinction between boundedness and feature-similarity addressed in experiment 2. Additionally, the sensitivity of the reconstruction method could possibly be improved through some methodological changes. Regardless, multivariate techniques represent a powerful window into long-standing questions about representational architecture in the field of working memory, enabling the direct testing of hypotheses related to neural population-level codes.

# References:

Anderson, D. E., Serences, J. T., Vogel, E. K., and Awh, E. (2014). Induced alpha rhythms track the content and quality of visual working memory representations with high temporal precision. *The Journal of Neuroscience* 34: 7587-7599.

Alderson, R. M., Kasper, L. J., Patros, C. H., Hudec, K. L., Tarle, S. J., and Lea, S. E. (2015). Working memory deficits in boys with attention deficit/hyperactivity disorder (ADHD): An examination of orthographic coding and episodic buffer processes. *Child Neuropsychol.* 2015;21(4):509-30.

Armstrong KM, Fitzgerald JK, Moore T. Changes in visual receptive fields with microstimulation of frontal cortex. Neuron. 2006:50791–798

Arnell KM, Jolicoeur P. (1999). The attentional blink across stimulus modalities: evidence for central processing limitations. *J. Exp. Psychol.: Hum. Percept. Perform.* 25:630–48

Axmacher N., Mormann F., Fernández G., Elger C. E., and Fell J. (2006) Memory formation by neuronal synchronization. *Brain Res Rev* 52:170–182.

Awh, E., Jonides, J., & Reuter-Lorenz, P. A. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(3), 780–790. https://doi.org/10.1037/0096-1523.24.3.780
Baddeley, A. D. (1986). Working Memory. London, Oxford University Press.

Baddeley, A. D. and G. J. Hitch (1974). Working Memory. *The Psychology of Learning and Motivation.* G. H. Bower. New York, Academic Press. 8: 47-89.

Baluch, F., Itti, L. Mechanisms of top-down attention. Trends in Neurosciences, 34 (4) (2011), pp. 210–224

Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. *J Neurosci* 30(28): 9424-9430.

Barrouillet, P., Bernardin, S., and Camos, V. (March 2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology. General* 133 (1): 83–100.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–31. https://doi.org/10.1016/j.neuropsychologia.2010.12.023

Bays, P. M., Catalao, R. F. G., and Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *J Vision* 9:7.

Bays, P. M. and Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science* 321:851–854.

Bettencourt, K. C. and Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience* 19, 150-157. do: 10.1038/nn.4174

Bisley, J.W. Goldberg, M.E. (2010). Attention, intention, and priority in the parietal lobe Annu. Rev. Neurosci., 33, pp. 1–21

Bisiach, Edoardo, and Claudio Luzzatti. (1978). "Unilateral Neglect of Representational Space." Cortex 14(1):129--133

Born, R. and Bradley, D. (2005). Structure and function of visual area MT. *Annu Rev Neurosci* 28: 157–89.

Brady TF, Konkle T, Alvarez GA, Oliva A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci U S A*. 2008 Sep 23;105(38):14325-9. doi: 10.1073/pnas.0803390105. Epub 2008 Sep 11.

Briggs, F., Mangun, G. R., & Usrey, W. M. (2013). Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature, 499(7459), 476-480

Broadbent, D (1958). *Perception and Communication*. London: Pergamon Press.

Brouwer, G. J. and Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 29(44), 13992–14003. http://doi.org/10.1523/JNEUROSCI.3577-09.2009

Buffalo EA, Fries P, Landman R, Liang H, Desimone R. A backward progression of attentional effects in the ventral stream. Proceedings of the National Academy of Sciences. 2010;107: 361–365

Buschman TJ, Miller EK. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science. 2007:3151860

Carrasco M. Visual attention: The past 25 years. Vision Res. 2011;51: 1484–1525.

Chelazzi, L. John Duncan, Earl K. Miller, Robert Desimone. Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search. Journal of Neurophysiology Dec 1998, 80 (6) 2918-2940

Christophel, T. B., M. N. Hebart, and J. D. Haynes (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *J Neuroscience* 32(38): 12983-12989

Chun, M.M. *et al.* A taxonomy of external and internal attention**.** Annu. Rev. Psychol., 62 (2011), pp. 73‑101

Clark, V. P. *et al*. Selective attention to face identity and color studied with fMRI. *Hum. Brain Mapp.* **5**, 293– 297 (1997).

Cohen MR, Maunsell JHR. Attention improves performance primarily by reducing interneuronal correlations. Nature Neuroscience. 2009:121594–1600

Corbetta, M. *et al*. Attentional modulation of neural processing of shape, color, and velocity in humans. *Science* 248, 1556 –1559 (1990)

Coull JT, Nobre AC. 1998. Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *J. Neurosci.* 18:7426–35

Coutanche, M. N. and Thompson-Schill, S. L. (2013). Using Informational Connectivity to Measure the Synchronous Emergence of fMRI Multi-voxel Information Across Time. *Journal of Visualized Experiments : JoVE*, (89), 51226. Advance online publication. http://doi.org/10.3791/51226

Cox, C. R., Seidenberg, M. S., and Rogers, T. R. (2014). Connecting functional brain imaging and Parallel Distributed Processing. *Language, Cognition and Neurosience* Volume 30, Issue 4 2015

Cox, D. D. and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*. Jun;19(2 Pt 1):261-70.

Cowan, N. (1995). Attention and Memory: An Integrated Framework. New York: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24: 87–185.

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., and Saults, J. S. (2006). Scope of Attention, Control of Attention, and Intelligence in Children and Adults. *Memory & Cognition*, *34*(8), 1754–1768.

Curtis, C. E. and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci* 7:415–423.

Cutrell EB, Marrocco RT. Electrical microstimulation of primate posterior parietal cortex initiates orienting and alerting components of covert attention. Experimental Brain Research. 2002:144103–113

David SV, Hayden BY, Mazer JA, Gallant JL. Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. Neuron. 2008;59: 509–521.

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Desimone R. Visual attention mediated by biased competition in extrastriate visual cortex. Philos Trans R Soc Lond B Biol Sci. 1998;353: 1245–1255

Deutsch, J. A.; Deutsch, D. (1963). "Attention: Some Theoretical Considerations". *Psychological Review*. **70**: 80–90. doi:10.1037/h0039515

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, *92*(1), 53–78. https://doi.org/10.1348/000712601162103

Duijnhouwer, J., Noest, A. J., Lankheet, M. J. M., van den Berg, A. V., & van Wezel, R. J. A. (2013). Speed and direction response profiles of neurons in macaque MT and MST show modest constraint line tuning. *Frontiers in Behavioral Neuroscience*, *7*, 22. http://doi.org/10.3389/fnbeh.2013.00022

Duncan J (1980), "The locus of interference in the perception of simultaneous stimuli." *Psychol Rev* 87(3):272-30

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology. General*, *113*(4), 501–17. https://doi.org/10.1037/0096-3445.113.4.501

Duncan J, Martens S, Ward R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature* 387:808–1

Egly, R.; Driver, J.; Rafal, R. D. (1994). "Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects" (PDF). *Journal of Experimental Psychology: General*. **123** (2): 161–177. doi:10.1037/0096-3445.123.2.161

Emrich, S. M., Riggall, A.C., Larocque, J. J., and Postle, B.R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *J Neurosci* 33(15): 6516-6523.

Eriksen, B. A.; Eriksen, C. W. (1974). "Effects of noise letters upon identification of a target letter in a non- search task". *Perception and Psychophysics*. **16**: 143–149. doi:10.3758/bf03203267

Erickson, M. A., Maramara, L. A., and Lisman, J. E. (2010) A single brief burst induces GluR1-dependent associative short-term potentiation: A potential mechanism for short-term memory. *J Cognitive*
*Neurosci* 22:2530–2540.

Ester, E. F., Anderson, D. E., Serences, J. T., and Awh, E. (2013). A neural measure of precision in visual working memory. *J Cogn Neurosci* 25(5): 754-761.

Ester E. F., Sprague, T. C., and Serences, J. T. (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron*. 87, 1–13

Fries P. Neuronal gamma-band synchronization as a fundamental process in cortical computation. Annual Review of Neuroscience. 2009:32209–224

Fries P, Womelsdorf T, Oostenveld R, Desimone R. The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. Journal of Neuroscience. 2008:284823

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61: 331-349.

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *Journal of Neurophysiology* 63: 814-831.

Fuster, J. M. and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science* 173: 652-654.

Gazzaley, A. (2011) Influence of early attentional modulation on working memory. Neuropsychologia 49, 1410–1424

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–35. https://doi.org/10.1016/j.tics.2011.11.014

Giesbrecht B, Woldorff MG, Song AW, Mangun GR. 2003. Neural mechanisms of top-down control during spatial and feature attention. *Neuroimage* 19:496–512

Goldman-Rakic, P. S. (1987). Circuitry of the prefrontal cortex and the regulation of behavior by representational memory. Handbook of Neurobiology. V. B. Mountcastle, F. Plum and S. R. Geiger. Bethesda, American Physiological Society: 373-417.

Goldman-Rakic, P. S. (1992). "Working memory and the mind." Scientific American 267: 110-117.

Gottlieb, J. From thought to action: the parietal cortex as a bridge between perception, action, and cognition Neuron, 53 (2007), pp. 9–16

Gnadt, J. and Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Exp Brain Res* 70:216–220.

Gregoriou GG, Gotts SJ, Zhou H, Desimone R. High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention. Science. 2009:3241207–1210.

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–94. https://doi.org/10.1162/089892903322598139

Hamidi, M., Slagter, H. A., Tononi, G., and Postle, B. R. (2009). Repetitive transcranial magnetic stimulation affects behavior by biasing endogenous cortical oscillations. *Frontiers in Integrative Neuroscience*, 3(14)

Han, X., Berg, A. C., Oh, H., Samaras, D., and Leung, H. C. (2013). Multi-voxel pattern analysis of selective representation of visual working memory in ventral temporal and occipital regions. *Neuroimage* 73: 8-15.

Harrison, S. A. and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458: 632-635.

Haynes J. D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neuroscience* 7:523–534.

Hebb, D. O. (1949). The Organization of Behavior: A Neuropsychological Theory. John Wiley & Sons Hoboken, NJ.

Hou, Y., & Liu, T. (2012). Neural correlates of object-based attentional selection in human cortex. *Neuropsychologia*, *50*(12), 2916–2925. http://doi.org/10.1016/j.neuropsychologia.2012.08.022

Hubel, D. H. and Wiesel, T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. J *Physiology* Mar;195(1):215-43.

Huth AG, Lee T, Nishimoto S, Bilenko NY, Vu AT, Gallant JL. Decoding the Semantic Content of Natural Movies from Human Brain Activity. Front Syst Neurosci. 2016 Oct 7;10:81. eCollection 2016

Huk, A. C., Dougherty, R. F., and Heeger, D. J. (2002). Retinotopy and functional subdivision of human areas MT and MST. *J Neurosci* 22:7195–7205.

Ikkai, A. and Curtis, C. E. (2011). Common neural mechanisms supporting spatial working memory, attention and motor intention. *Neuropsychologia* 49:1428-1434

Itti L, Koch C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. 40:1489–506

James, W. The Principles of Psychology**,** Henry Holt, New York (1890)

Jaswal S. The importance of being relevant. *Front Psychol. 2012; 3():309*

Jensen, O. (2006). Maintenance of multiple working memory items by temporal segmentation. *Neuroscience* 139:237–249.

Johnson, J. S., Hamidi, M., and Postle, B. R. (2010) Using EEG to explore how rTMS produces its effects on behavior. *Brain Topography,* 22, 281-293

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychological Science.* Apr;18(4):312-3.

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–42. https://doi.org/10.3758/s13423-012-0359-y

Klein BP, Harvey BM, Dumoulin SO (2014) Attraction of position preference by spatial attention throughout human visual cortex. Neuron 84:227–237

Kriegeskorte, N., Goebel, R., and Bandettini, P. A. (2006) Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103:3863–3868.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A., Oberauer, K., and Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *Journal of Cognitive Neuroscience* 25: 127-142.

LaRocque, J. J., Riggall, A. C., Emrich, S. M., and Postle, B. R. (2013). Active representations of individual items in short-term memory: A matter of attention, not retention. *Annual Meeting of the Society for Neuroscience*: 507.506.

LaRocque, J. J., Lewis-Peacock, J. A., and Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience* 8, 5. doi:10.3389/fnhum.2014.00005

LaRocque, J. J., Eichenbaum, N. S., Starrett, M. J., Rose, N. S., Emrich, S. M., & Postle, B. R. (2015) The short- and long-term fate of memory items retained outside the focus of attention *[Special Issue on Working Memory] Memory & Cognition*, 43(3), 453-468

Lauritzen TZ, D'Esposito M, Heeger DJ, Silver MA. Top–down flow of visual spatial attention signals from parietal to occipital cortex. J Vis. 2009;9: 18

Lee, S. H., Kravitz, D. J., and Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nat Neurosci* 16(8): 997-999.

Lepsien, J., and Nobre, A. C. (2007). Attentional modulation of object representations in working memory. *Cereb. Cortex* 17, 2072–2083.

Lewis-Peacock, J. A. and Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *The Journal of Neuroscience* 28: 8765-8771.

Lewis-Peacock, J. A. and Postle, B. R. (2012). Decoding the internal focus of attention. *Neuropsychologia* 50: 470-478.

Lewis-Peacock, J. A., Drysdale, A. T., and Postle, B. R. (2015). Neural evidence for the flexible control of mental representations. *Cerebral Cortex*, 25 (10), 3303-3313 doi: 10.1093/cercor/bhu130

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412: 150-157

Lorente de Nó, R. (1933). Vestibulo-ocular reflex arc. *Arch Neuro Psychiatr* 30:245–291.

Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature* 390:279–281.

Luck SJ, Chelazzi L, Hillyard SA, Desimone R. Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex. J Neurophysiol. 1997;77: 24–42

Luria, R., & Vogel, E. K. (2011). Shape and color conjunction stimuli are represented as bound objects in visual working memory. *Neuropsychologia*, *49*(6), 1632–1639. https://doi.org/10.1016/j.neuropsychologia.2010.11.031

Ma, W.J. et al. (2014) Changing concepts of working memory. Nat. Neurosci. 17, 347–356

Martínez-Trujillo J, Treue S. Attentional modulation strength in cortical area MT depends on stimulus contrast. Neuron. 2002;35: 365–370

Martinez-Trujillo JC, Treue S. Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. Curr Biol. 2004;14: 744–751

Maunsell J. H. & Treue S. Feature-based attention in visual cortex. Trends in Neurosciences 29, 317–322 (2006).

McElree, B. (1998). Attended and non-attended states in working memory: accessing categorized structures. *J. Mem. Lang.* 38, 225–252.

McElree, B. (2006). Accessing recent events. *Psychol. Learn. Motiv.* 46, 155–200.

Mendoza-Halliday, D., Torres, S., and Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat. Neurosci*. 17, 1255–1262.

Meyers, E. M., Qi, X. L., and Constantinidis, C. (2012). Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *Proc Natl Acad Sci USA* 109(12): 4651-4656.

Miconi T, VanRullen R. A Feedback Model of Attention Explains the Diverse Effects of Attention on Neural Firing Rates and Receptive Field Structure. Siegel M, ed. *PLoS Computational Biology*. 2016;12(2):e1004770. doi:10.1371/journal.pcbi.1004770

Miller, E. K., Li, L., Desimone, R. (1993) Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J Neurosci* 13:1460–1478.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2): 81–97.

Milner, B. (1964). Some effects of frontal lobectomy in man. The Frontal Granular Cortex and Behavior. J. M. Warren and K. Akert. New York, McGraw-Hill: 313-334.

Mitchell JF, Sundberg KA, Reynolds JH. Differential attention-dependent response modulation across cell classes in macaque visual area V4. Neuron. 2007:55131–141

Mongillo, G., Barak, O., and Tsodyks, M. (2008) Synaptic theory of working memory. *Science* 319:1543–1496.

Moore T, Fallah M. Control of eye movements and spatial attention. Proceedings of the National Academy of Sciences. 2001:21549498

Myers, Nicholas E. , Mark G. Stokes, Anna C. Nobre, Prioritizing Information during Working Memory: Beyond Sustained Internal Attention, Trends in Cognitive Sciences, Available online 25 April 2017, ISSN 1364-6613, https://doi.org/10.1016/j.tics.2017.03.010

Nee, D. E. and Jonides, J. (2008). Neural correlates of access to short-term memory. *Proc. Natl. Acad. Sci. U S A* 105, 14228–14233.

Nelissen, N., Stokes, M. G., Nobre, A. C., and Rushworth, M. F. (2013). Frontal and Parietal Cortical Interactions with Distributed Visual Representations during Selective Attention and Action Selection. *J Neurosci* 33(42): 16443-16458.

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–73. https://doi.org/10.1162/089892904322926700

Norman, K. A., Polyn, S. M., Detre, G. J., Haxby, J. V. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10:424–430.

Noudoost B, Chang MH, Steinmetz NA, Moore T. Top-down control of visual attention. Curr Opin Neurobiol. 2010;20: 183–190

Oberauer, K., & Hein, L. (2012). Attention to Information in Working Memory. *Current Directions in Psychological Science*, *21*(3), 164–169. https://doi.org/10.1177/0963721412444727

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Vol 28(3), May 2002, 411-421. http://dx.doi.org/10.1037/0278-7393.28.3.411

O'Craven KM, Downing PE, Kanwisher N. 1999. fMRI evidence for objects as the units of attentional selection. *Nature* 401:584–87

Pashler, H. (1988). Familiarity and visual change detection. *Percept Psychophys* 44:369–378.

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–31. https://doi.org/10.1037/a0030947

Posner, M. I., Snyder, C. R. R., and Davidson, B. J. (1980). Attention and Detection of Signals." *Journal of Experimental Psychology: General* 1980, Vol. 109, No. 2, 160-174

Posner, M. I., Walker, J. A., Friedrich, F. J. & Rafal, R. D. Effects of parietal lobe injury on covert orienting of visual attention. *J. Neurosci.* **4**, 1863–1874 (1984).

Postle, B. R. (2007). "Activated long-term memory"? The bases of representation in working memory. *The Cognitive Neuroscience of Working Memory.* N. Osaka, R. H. Logie and M. D'Esposito. Oxford, U.K., Oxford University Press: 333-350.

Postle, B. R. (2015). Activation and information in working memory research. In A. Duarte, M. Shields, and D.R. Addis (Eds.) The Wiley-Blackwell Handbook on the Cognitive Neuroscience of Memory. Wiley-Blackwell (Oxford, U.K.), 21-43.

Postle, B. R. (2015) The cognitive neuroscience of visual short-term memory *Current Opinion in Behavioral Sciences*, 1, 40-46.

Postle, B. R. and Pasternak, T. (2009) Short term and working memory. In L.R. Squire (Ed.) Encyclopedia of Neuroscience. Elsevier (Oxford), Volume 8, pp. 793-799

Pribram, K. H., Ahumada, A., Hartog, J., and Roos, L. (1964). A progress report on the neurological processes disturbed by frontal lesions in primates. The Frontal Granular Cortex and Behavior. J. M. Warren and K. Akert. New York, McGraw-Hill Book Company: 28-55.

Ranganath, C. and Blumenfeld, R. S. (2005). Doubts about double dissociation between short- and long-term memory. *Trends in Cognitive Sciences* 9: 374-380.

Recanzone, GH, Wurtz, RH, Schwartz, U. Responses of MT and MST neurons to one and two moving objects in the receptive field. J. Neurophysiol. 78, 1997; 2904–15

Reynolds JH, Chelazzi L. Attentional modulation of visual processing. Annu Rev Neurosci. 2004;27: 611–647

Reynolds JH, Pasternak T, Desimone R. Attention Increases Sensitivity of V4 Neurons. Neuron. 2000;26: 703–714

Riggall, A.C. (2014). The Neural Underpinnings of Short-term Memory for Visual Motion. *Manuscript in Preparation*

Riggall, A. C. and B. R. Postle (2012). The relationship between working memory storage and elevated activity as measured with funtional magnetic resonance imaging. *The Journal of Neuroscience* 32: 12990-12998.

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature.* May 30;497(7451):585-90.

Ruff DA, Cohen MR. Attention Increases Spike Count Correlations between Visual Cortical Areas. *The Journal of Neuroscience*. 2016;36(28):7523-7534. doi:10.1523/JNEUROSCI.0610-16.2016

Saalmann YB, Pigarev IN, Vidyasagar TR. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. Science. 2007:3161612

Saalmann YB, Pinsk MA, Wang L, Li X, Kastner S. Pulvinar regulates information transmission between cortical areas based on attention demands. *Science (New York, NY)*. 2012;337(6095):753-756. doi:10.1126/science.1223082

Salinas E, Sejnowski TJ. Correlated neuronal activity and the flow of neural information. Nature Reviews Neuroscience. 2001:2539–550.

Samaha, J., Bauer, P., Cimaroli, S., and Postle, B.R. (2015). Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. *Proceedings of the National Academy of Sciences*, 112(27), 8439-8444 doi: 10.1073/pnas.1503686112

Sarazin, M., Stern, Y., Berr, C., Riba, A., Albert, M., Brandt, J., Dubois, B. (2005). Neuropsychological predictors of dependency in patients with Alzheimer disease. *Neurology.* Mar 22;64(6):1027-31.

Schmidt, B.K. et al. (2002) Voluntary and automatic attentional control of visual working memory. Percept. Psychophys. 64, 754–763

Serences, J. T., Saproo, S., Scolari, M., Ho, T., and Muftuler, T. (2009) Estimating the influence of attention on population response profiles. *NeuroImage*. Jan 1;44(1):223-31.

Serences, J. T., Ester, E. F., Vogel, E. K. and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science* 20: 207-214.

Snitz, B. E., MacDonald, A. W., and Carter, C. S. (2006). Cognitive Deficits in Unaffected First-Degree Relatives of Schizophrenia Patients: A Meta-analytic Review of Putative Endophenotypes. *Schizophrenia Bulletin*, *32*(1), 179–194.

Snyder, Larry H., Aaron P. Batista, and Richard A. Andersen. 1997. "Coding of Intention in the Posterior Parietal Cortex." Nature 386(6621):167--170. doi:10.1038/386167a0

Sprague, T. C., Ester, E. F., Serences, J. T. (2014) Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology.* Sep 22;24(18):2174-80

Sreenivasan, K., Vytlacil, J., and D'Esposito, M. (2014). Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *Journal of Cognitive Neuroscience* 26: 1141-1153.

Sreenivasan, K. K., Curtis, C. E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity in working memory. *Trends in Cognitive Sciences* 18: 82-89.

Stanton GB, Bruce CJ, Goldberg ME. Topography of projections to posterior cortical areas from the macaque frontal eye fields. J Comp Neurol. 1995:353291–305

Stokes, M. G. (2015), ‘Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn Sci*, 19, 394 - 405

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron* 78(2): 364-375.

Stroop, John Ridley (1935). "Studies of interference in serial verbal reactions". *Journal of Experimental Psychology*. **18** (6): 643–662. doi:10.1037/h0054651

Thut G, Nietzel A, Brandt SA, Pascual-Leone A. alpha-Band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. Journal of Neuroscience. 2006:269494

Tiesinga P, Fellous JM, Sejnowski TJ. Regulation of spike timing in visual cortical circuits. Nature Reviews Neuroscience. 2008:997–109

Tipper, S.P. (1985). The negative priming effect: Inhibitory priming by ignored objects. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 37A, 571–590

Torriente, I., Valdes-Sosa, M., Ramirez, D. & Bobes, M. A. Visual evoked potentials related to motion-onset are modulated by attention . *Vision Res.* **39**, 4122– 4139 (1999)

Treisman, A., 1964. Selective attention in man. *British Medical Bulletin*, 20, 12-16.

Treisman, A (1960). "Contextual cues in selective listening". *Quarterly Journal of Experimental Psychology*. **12**: 242–248. doi:10.1080/17470216008416732

Treisman, A and Gelade, G. (1980). "A feature-integration theory of attention." *Cognitive Psychology*, Vol. 12, No. 1, pp. 97–136.

Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolić, D., and Singer, W. (2009) Neural synchrony in cortical networks: History, concept and current status. *Front Integr Neurosci* 3:1–19.

Valdés-Sosa MJ, Iglesias-Fuster J, Torres R. Attentional selection of levels within hierarchically organized figures is mediated by object-files. *Frontiers in Integrative Neuroscience*. 2014;8:91. doi:10.3389/fnint.2014.00091

van Bergen RS, Ma WJ, Pratte MS, Jehee JF. Sensory uncertainty decoded from visual cortex predicts behavior. Nat Neurosci. 2015 Dec;18(12):1728-30. doi: 10.1038/nn.4150. Epub 2015 Oct 26

Vecera, S. P., Behrmann, M., & McGoldrick, J. (2000). Selective attention to the parts of an object. *Psychonomic Bulletin & Review*, *7*(2), 301–308. https://doi.org/10.3758/BF03212985

Vy A. Vo, Thomas C. Sprague and John T. Serences. Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. Journal of Neuroscience 27 February 2017, 37 (12) 3386-3401

Warren, J. M. and Akert, K., Eds. (1964). The Frontal Granular Cortex and Behavior. New York, McGraw-Hill Book Company.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64. https://doi.org/10.1037//0096-3445.131.1.48

Wilken, P. and Ma, W. J. (2004). A detection theory account of change detection. *J Vision* 4:1120–1135.

Womelsdorf T, Fries P, Mitra PP, Desimone R. Gamma-band synchronization in visual cortex predicts speed of change detection. Nature. 2006:439733–736

Womelsdorf T, Anton-Erxleben K, Pieper F, Treue S. Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. Nat Neurosci. 2006;9: 1156–1160

Womelsdorf T, Schoffelen JM, Oostenveld R, Singer W, Desimone R, Engel AK, Fries P. Modulation of Neuronal Interactions Through Neuronal Synchronization. Science. 2007:3161609

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review*, *15*(1), 223–229. https://doi.org/10.3758/PBR.15.1.223

Woolgar, A., Williams, M.A., Rich, A.N. (2015). Attention enhances multi-voxel representation of novel objects in frontal, parietal and visual cortices. *NeuroImage*, Volume 109, 1 April 2015, Pages 429-437, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2014.12.083.

Yantis, S., Serences, JT. Cortical mechanisms of space-based and object-based attentional control, Current Opinion in Neurobiology, Volume 13, Issue 2, April 2003, Pages 187-193

Zanto, T.P. and Gazzaley, A. Neural Suppression of Irrelevant Information Underlies Optimal Working Memory Performance. Journal of Neuroscience 11 March 2009, 29 (10) 3059-3066; DOI: https://doi.org/10.1523/JNEUROSCI.4621-08.2009

Zhang W. and Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature* 453:233–235.

Zokaei, N., Manohar, S., Husain, M., and Feredoes, E. (2014). Causal Evidence for a Privileged Working Memory State in Early Visual Cortex. *The Journal of Neuroscience*, *34*(1), 158–162. http://doi.org/10.1523/JNEUROSCI.2899-13.2014

Zokaei, N., Ning, S., Manohar, S., Feredoes, E., and Husain, M. (2014). Flexibility of representational states in working memory. *Frontiers in Human Neuroscience*, *8*, 853. http://doi.org/10.3389/fnhum.2014.00853