UNIVERSITY OF CALIFORNIA SAN DIEGO

Mechanistic challenges in achieving precise top-down attentional modulations in the visual system

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Experimental Psychology

by

Sunyoung Park

Committee in charge:

Professor John Serences, Chair
Professor Anastasia Kiyonaga
Professor Eran Mukamel
Professor Viola Störmer

2024

The Dissertation of Sunyoung Park is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have come to fruition without the guidance and support from the amazing and intelligent people I have met throughout my journey in academia. My advisor, John, has been a great mentor throughout the last six years, including the two years of the global pandemic, which were filled with uncertainty. I could always rely on John to be supportive of my ideas and decisions, while also offering new perspectives and broadening my views, on both scientific and non-scientific matters. I have also gained so much from my labmates, past and present, who have become my found family in this foreign country. Thank you to Rosanne, Nuttida, Chaipat, Sirawaj, Vy, Steph, Maggie, Tim, Anna, Angus, Kirsten, Janna, Holly, Leah, Stella, Isa, and Luis. Through my time in John's lab, I have learned to become a better scientist and a better human being. I am deeply grateful to my committee members, Viola, Anastasia, and Eran for the invaluable feedback, that helped shape this dissertation.

I would also like to thank my friends in my cohort and the department, Minju, Mingi, Danbi, Alex, Aubrey, Lauren, Holly, Lim, and other grads that helped make San Diego feel like home. Thank you to the terrific undergraduate research assistants who I was fortunate to mentor: Daniella, Jenna, and Jennifer. Thank you to my friends outside of campus, especially those back in Korea, who ensured I had the social support I needed with endless jokes and laughter: Yoojin, Hyunsoo, Yeji, and Som. Thank you to Tony for sticking with me through the hardest part of the journey and taking care of me when I wasn't able to.

Lastly, thank you to my mom, who taught me the very first joys of learning. Thank you for believing in me even more than I believed in myself and inspiring me to take on challenges, however impossible they might seem.

Chapter 1, in full, is a reprint of the material as it appears in Journal of Neurophysiology, 127(2), 504-518, 2022, Park, Sunyoung; Serences, John T. The dissertation author was the primary investigator and author of this paper.

Chapter 2 and Chapter 3, in part, is currently being prepared for submission for publication of the material. Park, Sunyoung; Serences, John. T. The dissertation author was the primary researcher and author of this material.

VITA

2014 Bachelor of Arts in Psychology, Yonsei University

2016 Master of Science in Cognitive Sciences, Yonsei University

2024 Doctor of Philosophy in Experimental Psychology, University of California San Diego

ABSTRACT OF THE DISSERTATION

Mechanistic challenges in achieving precise top-down attentional modulations in the visual system

by

Sunyoung Park

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2024

Professor John Serences, Chair

Voluntary selective attention is a fundamental process in visual perception, enabling the prioritization of behaviorally relevant information from a complex sensory environment. While prefrontal cortex (PFC) has been implicated as the source of this top-down modulatory signal, how this is achieved is unclear given that PFC neurons have coarse and high-dimensional selectivity. Thus, this dissertation investigates how prefrontal areas with coarse and high-dimensional selectivity for spatial locations and visual features modulate sensory processing with

high precision. Chapter 1 explores the limits in the precision of top-down spatial attention, revealing that feedback signals from frontal and parietal regions are less spatially precise compared to the sensory responses in early visual areas, such as V1. Chapter 2 uses spiking neural network models to examine how feedback signals, despite their coarse and high-dimensional nature, can lead to precise modulations of early visual processing. These simulations highlight the importance of random projections between control and sensory networks in producing targeted attentional modulations and controlling disruptive spurious signals. Chapter 3 further investigates the impact of structured versus random feedback connections, showing that introducing slight structure to the network connections enables targeted attentional enhancement and the spread of feature-based attention. Together, these findings expand our understanding of how top-down attentional signals from prefrontal areas modulate sensory processing, emphasizing the interaction between high-dimensional representations in the prefrontal cortex and structured representations in early sensory areas to produce precise neural modulations required to achieve behavioral goals.

INTRODUCTION

A crucial aspect of the visual system is to process incoming information in a way that supports effective interactions with our surroundings to achieve behavioral goals. A key component serving this function is voluntary, or top-down, selective attention, which supports the ability to distinguish a behaviorally relevant visual target within a complex visual environment. As such, higher-order cognitive functions rely on being able to select and prioritize more relevant information over other less relevant information, which cannot be accomplished through passive processing of bottom-up sensory input.

Decades of research on top-down attention has shown that it leads to improvements in the sensory processing of the relevant visual inputs (1–4). For example, behavioral studies have revealed that top-down attention promotes efficient perception by enhancing the discriminability of basic visual features such as contrast (5, 6) accelerating the speed of processing (7), enhancing spatial resolution (8), and suppressing nearby distractors (9, 10). In addition, electrophysiological recording studies and functional magnetic resonance imaging (fMRI) studies have reported changes in the neural signals and representations underlying these behavioral enhancements. For example, directing attention to a spatial location or a feature can result in signal enhancement, noise reduction, and shifts in neural tuning (11–26; reviewed in 2, 3). All of these studies have demonstrated that our visual system has the ability to selectively target a specific visual input that matches current behavioral goals to prioritize important stimuli for processing.

As to how these behavioral and neural changes occur, studies suggest a key role for the prefrontal cortex in sending feedback signals that target relevant sensory neurons in early visual areas and modulate their activities. For example, the frontal eye field (FEF) has been found to be

the source area for spatial attention (27–31), as microstimulation to monkey FEF led to attention-like increase in V4 activity in a spatially selective way (28). Lateral prefrontal cortex (LPFC) has long been implicated in cognitive control, as disrupting LPFC activity leads to impaired top-down control over bottom-up stimulus drive (32–34). In addition, inferior frontal junction (IFJ) has been suggested as the source of attentional modulation for feature and object in humans (35–39). While there are some functional distinctions with respect to different types of top-down modulations, these brain regions are all higher-order areas that sit at the top of the feedforward hierarchy of sensory processing.

These prefrontal areas make good candidates as the source area for top-down attention signals, as they receive converging inputs from sensory cortex, motor cortex, the basal ganglia, and the medial temporal lobe (40). This convergence allows these areas to integrate different information - sensory, motor, memory, affect, and others - and compute which information is to be prioritized. Moreover, these prefrontal areas can conversely send control signals to modulate the activity of the different brain areas via widespread feedback projections (40, 41). However, as a consequence of converging inputs, the prefrontal cortex neurons end up having a coarse and high-dimensional selectivity to visual features. For example, studies have shown that FEF, a key area implicated in voluntary spatial attention, have much larger and variable spatial receptive fields (RFs) compared to early visual areas (42–44). Moreover, the prefrontal cortex has neurons that can encode multiple task parameters at the same time, with their responses varying to reflect different combinations of stimulus features and/or task demands (45, 46). This leads to the high dimensionality of representations in the prefrontal cortex, which cannot be easily mapped to neural codes in early sensory areas where RFs encode basic visual features in a relatively stereotyped and inflexible manner. In this dissertation I aim to address the question: Given the

coarse and high-dimensional selectivity in the prefrontal "control" areas, how do top-down attentional signals relayed from these areas modulate signals in early sensory cortex with the high precision required to achieve specific behavioral goals?

In Chapter 1, I investigate limits in the spatial precision of top-down modulations in the early visual areas in the absence of an external sensory signal. While the early visual areas that are the target of spatial attention modulations have highly selective spatial receptive fields, the frontal and parietal areas that are suggested to be the source for these signals have relatively coarse spatial receptive fields. Using fMRI, I demonstrate that while mid-level visual areas, such as V4, showed comparable top-down and bottom-up precision in their spatial representation, V1 showed a significant discrepancy, in which the top-down precision was substantially lower than the bottom-up precision. This finding provides a foundational insight that the nature of the top-down feedback signals is diffuse in early visual areas when there is no spatially specific bottom-up signal, reflecting the coarse spatial selectivity in the source areas for spatial attention.

In Chapter 2, I examine the consequences of relaying feedback signals through neurons with high-dimensional selectivity using a spiking neural network model. While prefrontal cortex neurons are found to have selectivity profiles that are complex and high-dimensional (e.g. maximally responding to a conjunction of two features, and less when the features are presented individually), it is unclear how top-down signals propagated through these neurons can result in a targeted modulation of the sensory response to the relevant feature. Through simulations in a neural model consisting of a sensory network and a control network connected through random and reciprocal projections, I demonstrate that feedback signals indeed propagate to unstimulated sensory neurons that are irrelevant to the target stimulus. However, this incidental spread of top-down modulations does not resemble sensory signals and thus does not lead to "illusory"

percepts. This finding highlights the importance of random connections between more structured early sensory areas and the later attentional control areas, which can cancel out spurious signals and selectively enhance the relevant neural activity.

In Chapter 3, I further test the importance of randomness in connectivity by parametrically manipulating the randomness of the projections in the spiking neural network model. While maximum randomness in between-network connections seems ideal in regulating spurious signals that can arise from feedback modulations, the simulations in Chapter 2 do not capture the global spread of feature-based attention (47, 48). By parameterizing the connectivity scheme on a continuum of randomness and structuredness, I demonstrate that more structure in the connections can give rise to spurious activity in unstimulated neurons that resemble coherent activity patterns for sensory stimuli. Moreover, I show that a balanced point can be achieved in which unstimulated neurons show elevated activity in a feature-specific way as reported in previous empirical studies, while the signals for the relevant stimulus receive the strongest top-down enhancement. This recapitulates the importance of a randomly connected network to show attentional enhancement in a targeted way while suggesting that minimal structure in the network may be needed to account for the global spread of feature-based attention.

In conclusion, this dissertation examines the nature of the voluntary attentional signals originating from prefrontal cortex and how they can provide precise modulatory feedback to highly selective neurons in early sensory cortex. Based on the known properties of the prefrontal cortex, I argue that the limited spatial precision of top-down modulations is the consequence of propagating feedback signals through neurons with high-dimensional selectivity, and that random connections between higher-order and early sensory neurons cancel out unwanted and spurious feedback signals. These findings provide new insights into the neural computations

underlying voluntary selective attention and highlight the interplay between the less structured, high-dimensional prefrontal regions and the more structured, low-dimensional sensory areas. Together, this work contributes to a deeper understanding of the mechanistic aspect of voluntary attention and outlines possible requirements in a biologically plausible model of our brain.

References

1. **Carrasco M**. Visual attention: the past 25 years. *Vision Res* 51: 1484–1525, 2011. doi: 10.1016/j.visres.2011.04.012.

2. **Desimone R**, **Duncan J**. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18: 193–222, 1995. doi: 10.1146/annurev.ne.18.030195.001205.

3. **Reynolds J**, **Chelazzi L**. Attentional Modulation of Visual Processing. *Annu Rev Neurosci* 27: 611–47, 2004. doi: 10.1146/annurev.neuro.26.041002.131039.

4. **Serences JT**, **Kastner S**. A Multi-level Account of Selective Attention. .

5. **Ling S**, **Carrasco M**. Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision Res* 46: 1210–1220, 2006.

6. **Liu T**, **Abrams J**, **Carrasco M**. Voluntary attention enhances contrast appearance. *Psychol Sci* 20: 354–362, 2009.

7. **Giordano AM**, **McElree B**, **Carrasco M**. On the automaticity and flexibility of covert attention: A speed-accuracy trade-off analysis. *J Vis* 9: 30–30, 2009.

8. **Barbot A**, **Carrasco M**. Attention Modifies Spatial Resolution According to Task Demands. .

9. **Awh E**, **Matsukura M**, **Serences JT**. Top-down control over biased competition during covert spatial orienting. *J Exp Psychol Hum Percept Perform* 29: 52, 2003.

10. **Serences JT**, **Yantis S**, **Culberson A**, **Awh E**. Preparatory activity in visual cortex indexes distractor suppression during covert spatial orienting. *J Neurophysiol* 92: 3538–3545, 2004.

11. **Brefczynski JA**, **DeYoe EA**. A physiological correlate of the'spotlight'of visual attention. *Nat Neurosci* 2: 370–374, 1999.

12. **Cohen MR**, **Maunsell JHR**. Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* 12: 1594–1600, 2009. doi: 10.1038/nn.2439.

13. **Colby CL**. The neuroanatomy and neurophysiology of attention. *J Child Neurol* 6: S90–S118, 1991.

14. **Kastner S**, **Ungerleider L**. Mechanisms of Visual Attention in the Human Cortex. *Annu Rev Neurosci* 23: 315–341, 2000. doi: 10.1146/annurev.neuro.23.1.315.

15. **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* 77: 24–42, 1997. doi: 10.1152/jn.1997.77.1.24.

16. **Mitchell JF**, **Sundberg KA**, **Reynolds JH**. Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron* 55: 131–141, 2007. doi: 10.1016/j.neuron.2007.06.018.

17. **Moran J**, **Desimone R**. Selective attention gates visual processing in the extrastriate cortex. *Science* 229: 782–784, 1985.

18. **Reynolds JH**, **Pasternak T**, **Desimone R**. Attention Increases Sensitivity of V4 Neurons. *Neuron* 26: 703–714, 2000. doi: 10.1016/S0896-6273(00)81206-4.

19. **Sundberg KA**, **Mitchell JF**, **Reynolds JH**. Spatial Attention Modulates Center-Surround Interactions in Macaque Visual Area V4. *Neuron* 61: 952–963, 2009. doi: 10.1016/j.neuron.2009.02.023.

20. **Vo VA**, **Sprague TC**, **Serences JT**. Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. *J Neurosci* 37: 3386–3401, 2017. doi: 10.1523/JNEUROSCI.3484-16.2017.

21. **Womelsdorf T**, **Anton-Erxleben K**, **Treue S**. Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *J Neurosci Off J Soc Neurosci* 28: 8934–8944, 2008. doi: 10.1523/JNEUROSCI.4030-07.2008.

22. **Hayden BY**, **Gallant JL**. Combined effects of spatial and feature-based attention on responses of V4 neurons. *Vision Res* 49: 1182–1187, 2009. doi: 10.1016/j.visres.2008.06.011.

23. **McAdams CJ**, **Maunsell JH**. Attention to both space and feature modulates neuronal responses in macaque area V4. *J Neurophysiol* 83: 1751–1755, 2000. doi: 10.1152/jn.2000.83.3.1751.

24. **Martinez-Trujillo JC**, **Treue S**. Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. *Curr Biol* 14: 744–751, 2004. doi: 10.1016/j.cub.2004.04.028.

25. **Motter BC**. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J Neurosci* 14: 2178–2189, 1994. doi: 10.1523/JNEUROSCI.14-04-02178.1994.

26. **David SV**, **Hayden BY**, **Mazer JA**, **Gallant JL**. Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. *Neuron* 59: 509–521, 2008. doi: 10.1016/j.neuron.2008.07.001.

27. **Moore T**, **Fallah M**. Microstimulation of the frontal eye field and its effects on covert spatial attention. *J Neurophysiol* 91: 152–162, 2004. doi: 10.1152/jn.00741.2002.

28. **Moore T**, **Armstrong KM**. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373, 2003. doi: 10.1038/nature01341.

29. **Gregoriou GG**, **Gotts SJ**, **Desimone R**. Cell-type-specific synchronization of neural activity in FEF with V4 during attention. *Neuron* 73: 581–594, 2012. doi: 10.1016/j.neuron.2011.12.019.

30. **Veniero D**, **Gross J**, **Morand S**, **Duecker F**, **Sack AT**, **Thut G**. Top-down control of visual cortex by the frontal eye fields through oscillatory realignment. *Nat Commun* 12: 1757, 2021. doi: 10.1038/s41467-021-21979-7.

31. **Ekstrom LB**, **Roelfsema PR**, **Arsenault JT**, **Bonmassar G**, **Vanduffel W**. Bottom-Up Dependent Gating of Frontal Signals in Early Visual Cortex. *Science* 321: 414–417, 2008. doi: 10.1126/science.1153276.

32. **Feredoes E**, **Heinen K**, **Weiskopf N**, **Ruff C**, **Driver J**. Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. *Proc Natl Acad Sci* 108: 17510–17515, 2011. doi: 10.1073/pnas.1106439108.

33. **Nee DE**, **D'Esposito M**. Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *eLife* 6: e28040, 2017. doi: 10.7554/eLife.28040.

34. **Mendoza-Halliday D**, **Xu H**, **Azevedo FAC**, **Desimone R**. Dissociable neuronal substrates of visual feature attention and working memory. *Neuron* 112: 850-863.e6, 2024. doi: 10.1016/j.neuron.2023.12.007.

35. **Zanto TP**, **Rubens MT**, **Bollinger J**, **Gazzaley A**. Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage* 53: 736–745, 2010. doi: 10.1016/j.neuroimage.2010.06.012.

36. **Baldauf D**, **Desimone R**. Neural Mechanisms of Object-Based Attention. *Science* 344: 424–427, 2014. doi: 10.1126/science.1247003.

37. **Gong M**, **Liu T**. Biased Neural Representation of Feature-Based Attention in the Human Frontoparietal Network. *J Neurosci* 40: 8386–8395, 2020. doi: 10.1523/JNEUROSCI.0690-20.2020.

38. **Bedini M**, **Baldauf D**. Structure, function and connectivity fingerprints of the frontal eye field versus the inferior frontal junction: A comprehensive comparison. *Eur J Neurosci* 54: 5462–5506, 2021. doi: 10.1111/ejn.15393.

39. **Meyyappan S**, **Rajan A**, **Mangun GR**, **Ding M**. Role of Inferior Frontal Junction (IFJ) in the Control of Feature versus Spatial Attention. *J Neurosci* 41: 8065–8074, 2021. doi: 10.1523/JNEUROSCI.2883-20.2021.

40. **Miller EK**, **Cohen JD**. An Integrative Theory of Prefrontal Cortex Function. *Annu Rev Neurosci* 24: 167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167.

41. **Pandya DN**, **Yeterian EH**. Prefrontal cortex in relation to other cortical areas in rhesus monkey: architecture and connections. *Prog Brain Res* 85: 63–94, 1990. doi: 10.1016/s0079-6123(08)62676-x.

42. **Blatt GJ**, **Andersen RA**, **Stoner GR**. Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. *J Comp Neurol* 299: 421–445, 1990. doi: 10.1002/cne.902990404.

43. **Mayo JP**, **DiTomasso AR**, **Sommer MA**, **Smith MA**. Dynamics of visual receptive fields in the macaque frontal eye field. *J Neurophysiol* 114: 3201–3210, 2015. doi: 10.1152/jn.00746.2015.

44. **Mohler CW**, **Goldberg ME**, **Wurtz RH**. Visual receptive fields of frontal eye field neurons. .

45. **Fusi S**, **Miller EK**, **Rigotti M**. Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol* 37: 66–74, 2016. doi: 10.1016/j.conb.2016.01.010.

46. **Rigotti M**, **Barak O**, **Warden MR**, **Wang X-J**, **Daw ND**, **Miller EK**, **Fusi S**. The importance of mixed selectivity in complex cognitive tasks. *Nature* 497: 585–590, 2013. doi: 10.1038/nature12160.

47. **Treue S**, **Trujillo JCM**. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399: 575–579, 1999. doi: 10.1038/21176.

48. **Saenz M**, **Buracas GT**, **Boynton GM**. Global effects of feature-based attention in human visual cortex. *Nat Neurosci* 5: 631–632, 2002. doi: 10.1038/nn876.

Chapter 1 Relative precision of top-down attentional modulations is lower in early visual cortex compared to mid- and high-level visual areas

Abstract

Top-down spatial attention enhances cortical representations of behaviorally relevant visual information and increases the precision of perceptual reports. However, little is known about the relative precision of top-down attentional modulations in different visual areas, especially compared to the highly precise stimulus-driven responses that are observed in early visual cortex. For example, the precision of attentional modulations in early visual areas may be limited by the relatively coarse spatial selectivity and the anatomical connectivity of the areas in prefrontal cortex that generate and relay the top-down signals. Here, we used fMRI and human participants to assess the precision of bottom-up spatial representations evoked by high contrast stimuli across the visual hierarchy. Then, we examined the relative precision of top-down attentional modulations in the absence of spatially-specific bottom-up drive. While V1 showed the largest relative difference between the precision of top-down attentional modulations and the precision of bottom-up modulations, mid-level areas such as V4 showed relatively smaller differences between the precision of top-down and bottom-up modulations. Overall, this interaction between visual areas (e.g. V1 vs V4) and the relative precision of top-down and bottom-up modulations suggests that the precision of top-down attentional modulations is limited by the representational fidelity of areas that generate and relay top-down feedback signals.

New & Noteworthy

When the relative precision of purely top-down and bottom-up signals were compared across visual areas, early visual areas like V1 showed higher bottom-up precision compared to top-down precision. In contrast, mid-level areas showed similar levels of top-down and bottom-up precision. This result suggests that the precision of top-down attentional modulations may be

limited by the relatively coarse spatial selectivity and the anatomical connectivity of the areas generating and relaying the signals.

Introduction

The ability to voluntarily attend to a part of the visual field to process relevant sensory information is key to navigating visually cluttered environments. Attentional control signals are thought to originate in subregions of prefrontal cortex (PFC) and parietal cortex and target neurons in early visual cortex to improve information processing. However, relatively large receptive field sizes and the anatomical organization of feedback pathways in PFC may impose structural constraints on the precision of these top-down attentional modulations in earlier visual areas.

Previously, behavioral studies have shown that voluntary, or top-down, selective attention leads to spatially precise improvements in sensory processing (1–3). For example, the selective deployment of top-down attention to a relevant spatial position promotes more efficient perception in a variety of visual tasks (4, 5) by improving discriminability of basic visual features such as contrast (6, 7), spatial frequency (8), and textures (9). Moreover, top-down spatial attention can accelerate the speed of processing (2), enhance spatial resolution (10), and suppress nearby distractors (11, 12).

In addition to the changes in behavior, neurophysiological recording studies in monkeys and fMRI studies in humans have found signal enhancement, noise reduction, and shifts in spatial tuning in visual areas when top-down attention is directed inside single-unit or population-level receptive fields (13–22; reviewed in: 23, 24). Indeed, top-down modulations in early and later visual areas have been reported even in the absence of any kind of external stimulation (25–34), and activation patterns in visual cortex can be used to decode attended spatial locations and non-spatial features (35–44). However, these studies primarily focused on

establishing the presence of different kinds of attention effects, so little is known about the precision of top-down attentional modulations across different areas of the visual hierarchy.

The high precision of bottom-up sensory representations in early visual areas like V1 makes these areas potentially ideal targets for supporting equally precise top-down attentional modulations. For example, receptive fields of early visual area neurons are quite small (smaller than 1° at fovea and around 1° at 10° periphery in V1; Hubel & Wiesel, 1974) and are highly tuned to specific features such as orientation, motion trajectory, spatial frequency, or color (46–48). Thus, neurons in early visual areas can precisely code low-level visual features compared to neurons in mid- and high-level visual areas that have larger and more complex receptive fields, such as V4 (49–55).

Contrarily, regions in parietal and frontal cortex that are thought to be sources of top-down attentional control signals, such as the lateral intraparietal area (LIP) or the frontal eye fields (FEF), have much larger spatial receptive fields and thus relatively coarse codes for spatial position (e.g., receptive fields are around 10-20° with a large variance at 10° periphery in FEF; 56–59). In addition, anatomical studies suggest that areas such as FEF primarily send feedback projections to mid-level visual areas such as V4, with only sparse connections to earlier areas such as V1 (60–63). Consistent with these anatomical observations, microstimulations to FEF neurons lead to a spatially selective increase in the firing rates in V4 neurons (64), and V4 shows larger and earlier attentional enhancements in firing rates compared to V1 and V2, suggesting that attentional feedback signals target mid-level visual areas first and then are back-propagated to earlier visual areas (65). Thus, the precision of top-down attention signals in early visual areas might be limited by the coarser selectivity of the higher level ‘control’ areas and the mid-level

areas that appear to serve as the primary relay stations for top-down feedback to the highly selective neurons in early visual cortex.

In this study, we tested whether the precision of top-down attentional modulations was relatively coarser than bottom-up modulations in early visual areas compared to later visual areas. First, to measure the precision of bottom-up spatial representations in each visual area, participants were shown a series of small checkerboard stimuli at different locations in the visual field (termed the *bottom-up spatial mapping* task). Then, we used a separate task in which participants were cued to attend to a small or a large part of the visual field to measure the precision of top-down modulations in each visual area (termed the *top-down spatial attention* task). Note that, by necessity, the bottom-up mapping task and the top-down attention task are different (e.g. one has a small and highly salient peripheral stimulus, while the other requires covert attention to a cued location). However, despite using different paradigms, we can still compare the *relative* precision of both bottom-up and top-down modulations in early vs. later areas (and critically, the same tasks were used to assess the respective precision in all visual areas). As such, the key comparison in our analysis was evaluating the interaction between task type (bottom-up vs. top-down) and visual area, which indicates that there is a change in the relative precision of bottom-up and top-down modulations that cannot be explained by differences between the two tasks.

We applied multivariate pattern analyses (MVPA) to single-trial voxel activation patterns to generate several measures of the precision of spatial representations. Bottom-up representational precision was measured as decoding accuracy for a single stimulus location in the bottom-up mapping task. Top-down attentional precision was measured as decoding accuracy for attended locations in the attention task. In addition to the decoding accuracy, we also

analyzed confusion matrices from the raw classifier outputs to provide a more nuanced assessment of the spatial precision in visual areas during each task. We found that earlier visual areas, such as V1, showed higher relative precision for bottom-up representations compared to top-down representations. However, mid-level and parietal areas such as V3, V4 and IPS showed comparable bottom-up and top-down precision, leading to an interaction between region of interest and the relative precision of different modulations. Overall, these findings are consistent with the hypothesis that the precision of top-down modulations in early visual cortex is limited by the precision of areas that generate and relay top-down feedback signals.

Methods

Participants

Five participants (3 females), who had normal or corrected-to-normal vision and were right-handed, were recruited from the University of California, San Diego (UCSD) community (mean age 26.0 ± 2.7 yr). All participants performed a one-hour behavioral training session followed by two fMRI scan sessions of the two main experimental tasks, each session lasting two hours. Participants were also scanned to obtain data for retinotopic mapping of occipital and parietal cortex in 1-2 separate fMRI sessions. In addition, extra runs of the bottom-up spatial mapping task were obtained for some participants who also participated in a separate study that used the same mapping task. Participants were compensated for their participation at a rate of $10/h for the behavioral training and $20/h for the scanning sessions. The local Institutional Review Board at UCSD approved the experiment, and all participants provided written informed consent.

Stimulus generation and presentation

All stimuli were projected on a 24 x 18 cm screen placed inside the bore of the scanner and viewed through a head-coil mounted mirror from 47 cm in a darkened room. Stimuli were generated and presented using MATLAB and the Psychophysics toolbox (66, 67). The luminance output from the projector was linearized in the stimulus presentation software and all stimuli were presented against a uniform grey background (286 cd/m^2).

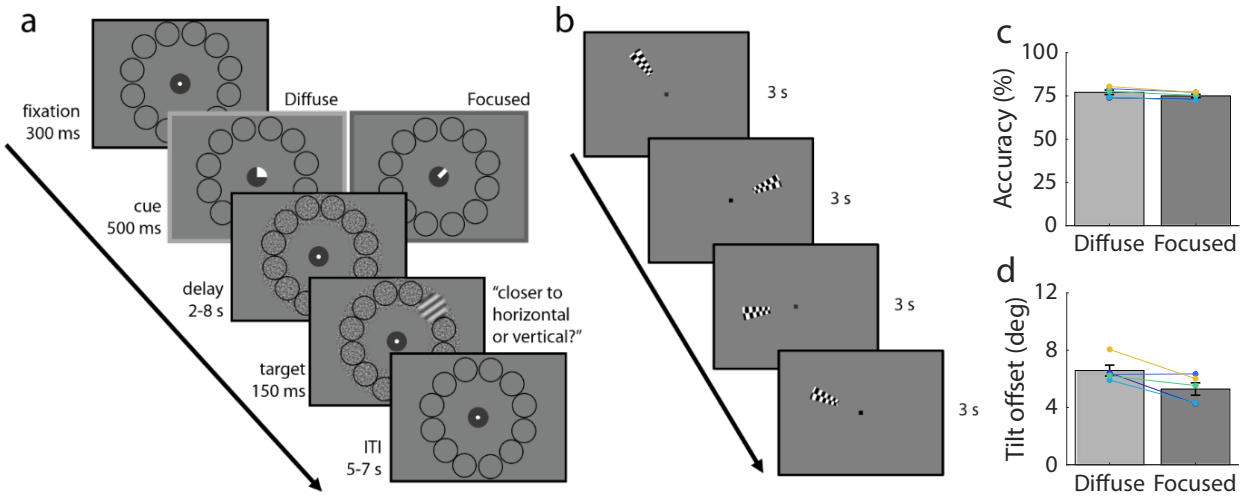


Figure 1.1 (a) Task procedure of top-down spatial attention task. Each trial started with a brief flicker of the white fixation dot at the center of the screen. After 300 ms, one of the two central cues indicating the quadrant (Diffuse) or the exact location (Focused) of the upcoming target was shown for 500 ms. The cues validly predicted the target location in all trials. A cue-to-target delay period followed, and the duration of the delay period varied between 2-8 s (2 s in catch trials and 6-8 s in non-catch trials; for detail, see section *Top-down spatial attention task* in Methods). A uniform flickering noise stimulus that was at the same contrast level as the target for the trial was present during the delay period. After the delay, the target grating was presented for 150 ms in one of the twelve possible locations, and the participants responded by button press whether the orientation of the target was closer to a horizontal or a vertical orientation. Placeholders (black circular lines) were present throughout the task, marking the possible target locations arranged on an imaginary circle. After the target offset, the subsequent trial started after an intertrial interval (ITI) of 5-7 s. Pre-scan training sessions used a slightly modified version of this task (see section *Pre-scan training session* in Methods for detail). (b) Task procedure for bottom-up spatial mapping task. In each of the 3-s trials, a wedge-shaped flickering checkerboard was presented in one of 24 locations, arranged on an imaginary circle. The 24 wedges altogether tiled the target location areas in the attention task, and in the decoding analyses, data from adjacent wedges were combined to match the 12 target locations for the cross-generalization analysis (see section *MVPA decoding* in Methods). In this task, participants responded by button press whenever the contrast of the fixation changed. (c, d) Behavioral performance for the top-down spatial attention task, data combined from the scanning sessions and an independent eye-tracking session (see Supplementary Methods). (c) Mean behavioral accuracy. Accuracy was higher in the diffuse compared to the focused condition by 2%. (d) Mean tilt offset. Tilt offsets were higher in the diffuse than in the focused condition. Colored dots represent data from individual participants. Error bars represent ±1 standard error of the mean (SEM).

Bottom-up spatial mapping task

To estimate the precision of bottom-up spatial representations, we presented one of 24 wedge-shaped flickering checkerboards at full contrast on each trial (Figure 1.1b). The wedges tiled an imaginary donut-shaped annulus that was 5 degrees of visual angle from the inner to the outer edge and spanned 4.5 - 9.5 degrees of visual angle from the center fixation point. Thus, each wedge covered 15° of arc (360°/24). The annulus covered the target locations in the top-down attention task. The checkerboard stimulus on each trial was presented on a uniform grey background and a 0.2° central fixation point was present throughout the task. Participants were instructed to indicate, using a button press with their right index finger, changes in the contrast of the fixation dot that occurred on 20 out of 96 (20.8%) trials in each run, evenly split between increment and decrement of the contrast.

Each trial consisted of a 3-s stimulus presentation period without any inter-trial intervals (ITIs). The location of the wedge was randomly selected on each trial with the constraint that wedges did not appear in the same quadrant consecutively. The contrast of the fixation dot changed at least 7s after the previous change. To ensure that participants performed below ceiling, we manually adjusted the task difficulty after every run by changing the percent contrast change of the contrast increment/decrement of the fixation. Each run consisted of 96 trials, a 12.8-s blank period at the beginning of the run, and a 12-s blank period at the end of the run, adding up to 312.8 s. Participants completed a total of 9-15 runs of the spatial mapping task across 2-4 separate sessions.

Top-down spatial attention task

Throughout the top-down spatial attention task, a white central fixation dot (0.2° diameter) on a dark grey central dot (105 cd/m^2, 2° diameter) was used as a central fixation point. There were also placeholders for target locations present throughout the task, which consisted of 12 black circular lines equidistant from the center of the screen, marking the 12 possible target locations. The placeholders tiled a donut-shaped annulus with a thickness of 5° visual angle, spanning 4.5 – 9.5° from the center of the screen, with three placeholders positioned in each of the four quadrants. Placeholders were used to facilitate attentional selection of the cued location(s). Each trial started with the white fixation dot flickering to indicate the trial onset and 300 ms later, followed by a cue presented for 500 ms (Figure 1.1a). One of two attention cues was presented on the dark grey central dot. The cue was either a diffuse cue or a focused cue; the diffuse cue was a filled, white arc covering a quadrant of the dark grey dot, indicating the quadrant that contains the upcoming target, and the focused cue was a 0.08° thick white line extending from the center to the outer edge of the dark grey dot, indicating the exact target location out of the 12 possible locations. The cues always validly predicted the target location, and the participants were explicitly instructed to covertly attend to the cued location while maintaining fixation at the white fixation dot.

There was a delay period between the cue and the target which, in 24 trials (non-catch) out of 27 trials in a single run, varied between 6-8 s in steps of 400 ms, and in the remaining 3 trials (catch), was fixed at 2 s. These different lengths of delay period were used to ensure that we had enough time points during the delay period to estimate neural response patterns before the onset of the target, and at the same time to force participants to direct attention right after the presentation of the cue in anticipation of the rare 2-s delays. The 2-s delay catch trials were

excluded from all analyses, behavioral or fMRI. During the delay period, a donut-shaped Fourier-filtered noise stimulus, which spanned 4.5 – 9.5° from the center of the screen, covering all possible target locations, counterphase flickered at 4 Hz. The noise stimulus consisted of white noise filtered to include only spatial frequencies between 1 and 4 c/°. During the first second of the delay period, the contrast of the noise ramped up linearly from 0% to the contrast of the target grating for that trial and continued flickering until target offset. The noise stimulus served as a dynamic continuous bottom-up input and also reduced transient signals caused by abrupt onset or offset of the target (68).

After the delay, one of the placeholders switched to the target grating for 150 ms. Participants were instructed to report whether the orientation of the target was closer to a vertical or a horizontal orientation with a button press with their right index or middle finger, within 2 s after the target onset. The target stimulus was a circular sinusoidal grating (2.7° radius) with a smoothed edge (filtered through a 2.7° radius, 0.9° sd Gaussian kernel), presented at the eccentricity of 7° from the center fixation. The grating was presented in a phase that was chosen at random on each trial, and at one of five possible contrast levels (0, 5, 10, 20, 50% Michelson contrast). The grating had a spatial frequency of 0.91 c/° and an orientation that was tilted clockwise or counter-clockwise from 45° or 135° by variable offsets. The tilt offsets for each of the cue and the contrast conditions (10 conditions in total) started at the orientation threshold values measured from the pre-scan behavioral training session (see section *Pre-Scan Training Session)* and were adjusted manually by the experimenter between every run to match the cumulative accuracies across all conditions around 75%. For example, if the cumulative accuracy for a given condition (i.e., 5% contrast, focused cue condition) was lower than 75%, the tilt offset

for that condition was adjusted to a higher value, making it easier to detect the tilted direction. The response mapping was counterbalanced between participants.

The ITI between the offset of the target and the onset of the next trial varied between 5-7 s in steps of 400 ms. Each run consisted of 27 trials, a 12.8-s and 8-s blank period at the beginning and the end of the run, respectively, adding up to 383.2 s. Two cue conditions (diffuse or focused), five levels of contrast, and 12 target locations were fully counterbalanced across 5 runs. The reference orientation (45° or 135°) and the direction of tilt from the reference (clockwise or counter-clockwise) were balanced for each of the cue, contrast, and target location combinations across the total of 20 runs carried out across two separate sessions. Due to technical difficulties, in one of the participants, behavioral and fMRI data for 20% contrast level in one session and behavioral responses for two runs were not collected.

Pre-scan training session

Before the scanning sessions, participants were trained on the top-down attention task in a separate pre-scanning session in the lab using the same task with the following exceptions. The cue-to-target delay period was shortened to 1 s, and there was no ITI between trials so that the next trial started as soon as the response for the current trial was made. Responses were made by pressing one of the two keys on a keyboard. The target grating was shown at lower contrast levels (2, 3, 5%), and tilt offsets were fixed at five levels (0, 5, 10, 15, 20°). There were 24 trials for each of the cue, contrast, and tilt offset combinations, and a total of 720 trials were split up into 12 blocks with short breaks in between.

Analysis of behavioral data

To assess the effect of the attention cue and the contrast of the target on behavior, we performed a two-way ANOVA on the behavioral accuracy and the tilt offset. Cue was considered as a categorical factor, and contrast was considered as a continuous factor. First, we tested the main effect of the two factors on accuracy. P-values were computed using randomization testing to avoid assumptions that usually accompany parametric tests. Thus, to test the effect of cue, cue labels were shuffled across trials, restricted within each contrast level and participant, and then accuracy was averaged across trials to obtain the mean accuracy for each condition. Then, we performed the ANOVA on this data with the two factors, only testing for the main effects without the interaction term, to obtain the F-value for the main effect of cue. This procedure was repeated 1000 times, yielding a distribution of 1000 F-values under the null hypothesis that cue type did not affect the behavioral accuracy. A p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with the unshuffled labels. The same procedure was repeated to test the main effect of contrast, but shuffling the contrast labels instead of the cue labels.

Next, we tested the interaction effect between cue and contrast. We first performed an ANOVA on the accuracy data with only the main effects and obtained the residual from this test. This was done to exclude the variance explained by the main effects. To test the interaction between cue and contrast, we shuffled the labels of the residual data, restricted within each participant. Then, we performed the two-way ANOVA on the residuals, testing for the two main effects and the two-way interaction effect, to obtain the F-value for the interaction effect between cue and contrast. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null

distribution that exceeded the F-value obtained using the data with unshuffled labels. The same series of the randomization tests was repeated on the tilt offset.

Localizer task

To localize voxels that maximally respond to the target locations in the bottom-up spatial mapping task and the top-down attention task, we presented a flickering checkerboard stimulus at full contrast shaped as a circular annulus, spanning 4.5 – 9.5° radius from the center fixation point (corresponding to the spatial locations of the bottom-up mapping stimuli and the attention target stimuli). The flickering localizer stimulus was presented on a uniform grey background with the same background luminance as in the top-down attention task, and a 0.2° central fixation point was present throughout the task. Each trial consisted of a 13-s stimulus presentation followed by a 13-s blank ITI. The task for the participants was the same as the bottom-up mapping task in which participants reported occasional changes in the contrast of the fixation point. Each run consisted of 12 trials and a 12.8-s blank period at the beginning of the run, adding up to 324.8 s.

Magnetic resonance imaging

All MRI scanning was performed on a General Electric (GE) Discovery MR750 3.0T research-dedicated scanner at the UC San Diego Keck Center for Functional Magnetic Resonance Imaging. Functional echo-planar imaging (EPI) data were acquired using a Nova Medical 32-channel head coil (NMSC075-32- 3GE-MR750) and the Stanford Simultaneous Multi-Slice (SMS) EPI sequence (MUX EPI), with a multiband factor of 8 and 9 axial slices per band (72 total slices; voxel size = 2-mm^3 isotropic; 0-mm gap; matrix = 104 x 104; field of view = 20.8 cm; TR/TE = 800/35 ms; flip angle = 52°; in-plane acceleration = 1). Image

reconstruction procedures and unaliasing procedures were performed on local servers using reconstruction code from CNI (Center for Neural Imaging at Stanford). The initial 16 TRs collected at the onset of each run served as reference images required for the transformation from k-space to image space. Two short (17 s) "topup" data sets were collected during each session, using forward and reverse phase-encoding directions. These images were used to estimate susceptibility-induced off-resonance fields (69) and to correct signal distortion in EPI sequences using FSL (FMRIB Software Library; http://www.fmrib.ox.ac.uk/fsl) topup (70).

During each functional session, we also acquired an accelerated anatomical scan using parallel imaging (GE ASSET on a FSPGR T1-weighted sequence; 172 slices; 1 x 1 x 1-mm^3 voxel size; TR = 8,136 ms; TE = 3,172 ms; flip angle = 8°; 1-mm slice gap; 256 x 192-cm matrix size) using the same 32-channel head coil. We also acquired one additional high-resolution anatomical scan (172 slices; 1 x 1 x 1-mm voxel size; TR = 8,136 ms; TE = 3,172 ms; flip angle = 8°; 1-mm slice gap; 256 x 192-cm matrix size) during a separate retinotopic mapping session using an Invivo eight-channel head coil. This scan produced higher quality contrast between gray and white matter and was used for segmentation, flattening, and visualizing retinotopic mapping data.

Preprocessing

All imaging data were preprocessed using software tools developed and distributed by FreeSurfer (https://surfer.nmr.mgh.harvard.edu) and FSL. Cortical surface gray-white matter volumetric segmentation of the high-resolution anatomical image was performed using the "recon-all" utility in the FreeSurfer analysis suite (71). Segmented T1 data were used to define regions of interest (ROIs) for use in subsequent analyses. The first volume of every functional

run was then coregistered to this common anatomical image. Transformation matrices were generated using FreeSurfer's manual and boundary-based registration tools (72). These matrices were then used to transform each 4D functional volume using FSL FLIRT (73, 74), such that all cross-session data from a single participant was in the same space. Next, motion correction was performed using the FSL tool MCFLIRT (73) without spatial smoothing, a final sinc interpolation stage, and 12 degrees of freedom. Slow drifts in the data were removed last, using a high pass filter (1/40 Hz cutoff). No additional spatial smoothing was applied to the data apart from the smoothing inherent to resampling and motion correction. Signal amplitude time-series were normalized via z-scoring on a voxel-by-voxel and run-by-run basis. Z-scored data were used for all further analyses unless mentioned otherwise. Because trial events were jittered with respect to TR onsets, they were rounded to the nearest TR. The analyses performed after preprocessing were all carried out in MATLAB 9.1 using custom in-house functions (code to be posted on osf.org).

Identifying Regions of Interest (ROIs)

To identify voxels that were visually responsive to the target locations, a general linear model (GLM) was performed on data from the localizer task using FSL FEAT (FMRI Expert Analysis Tool, version 6.00). Individual localizer runs were analyzed using BET brain extraction (75) and data prewhitening using FILM (76). Predicted BOLD responses were generated for blocks of annulus stimulus by convolving the stimulus sequence with a canonical gamma hemodynamic response function (phase = 0 s, sd = 3 s, lag = 6 s). The temporal derivative was also included as an additional regressor to accommodate slight temporal shifts in the waveform to yield better model fits and to increase explained variance. Individual runs were combined using a standard weighted fixed effects model, and spatial smoothing was applied (2-mm

FWHM). Voxels that were significantly activated by the localizer stimulus ($p < 0.05$; FDR corrected) were defined as visually responsive and used in all subsequent analyses (see Supplementary Table 1.2 for the number of selected voxels for each ROI). Note that the voxels in each ROI were identified based on their responsiveness to bottom-up stimulus drive. However, this should not lead to the identification of voxels that are less sensitive to top-down modulations given that top-down and bottom-up factors are known to target the same regions of retinotopically organized visual areas (29, 77–79). Supplementary control analyses were conducted by selecting voxels based on activations during the top-down attention task, and the results are shown in Supplementary Figures 1.8 and 1.9.

Retinotopic mapping stimulus protocol

Retinotopic mapping data were acquired during an independent scanning session, following previously published retinotopic mapping protocols to define the visual areas V1, V2, V3, V3AB, V4, IPS0, IPS1, IPS2, and IPS3 (80–85). The participants viewed a contrast-reversing checkerboard stimulus (4 Hz) configured as a rotating wedge (10 cycles, 36 s/cycle), an expanding ring (10 cycles, 36 s/cycle), or a bowtie (8 cycles, 40 s/cycle). To increase the quality of data from parietal regions, participants performed a covert attention task on the rotating wedge stimulus, which required them to detect contrast-dimming events that occurred occasionally (on average, 1 event occurred every 7.5 s) in a row of the checkerboard. This stimulus was limited to a 22° x 22° field of view.

Anatomical and functional retinotopy analyses were performed using a set of custom wrappers around existing FreeSurfer and FSL functionality. ROIs were combined across left and right hemispheres and across dorsal and ventral areas (for V2–V3) by concatenating voxels.

Only visually responsive voxels, selected using the localizer procedure described in the section above, were included in each of the retinotopically defined ROIs. For MVPA analyses, IPS0, IPS1, IPS2, and IPS3 were combined into a single "IPS" ROI for each participant, because some IPS ROIs of some participants did not have enough number of visually responsive voxels to train and test classifiers (less than 50).

Multivariate pattern analysis (MVPA) decoding

The purpose of MVPA analysis was to compare the precision of bottom-up stimulus-driven representations of spatial location and the precision of top-down attentional modulations based on the amount of information in voxel activation patterns in each ROI.

Before performing MVPA, we first shifted all timepoints by 4 seconds to account for the hemodynamic lag. Then, we averaged voxel activation patterns for each trial across the timepoints of the whole trial from the bottom-up mapping task (3-s duration) and the cue-to-target delay period from the attention task (6-8 s duration). To make sure that we were measuring information related to the relative pattern of activation across voxels in each ROI rather than information solely related to mean signal changes across conditions, we mean-centered each single-trial voxel activation pattern: For each trial and within each ROI, we calculated the mean amplitude across voxels and subtracted that value from the amplitude of each voxel, which resulted in trial-wise voxel activation patterns that have the across-voxel mean value at 0.

We performed all decoding analyses using a linear support-vector machine (SVM). First, to measure the precision of bottom-up representations, the classifier was trained and tested on the bottom-up spatial mapping task data in a leave-one-run-out manner. Critically, data points for two adjacent wedges were grouped together under the same location label to match the 12 target

locations in the attention task. Decoding accuracy was calculated as the proportion of trials where the location predicted by the classifier matched the location where the wedges were presented. Next, to measure the precision of top-down attentional modulations, we used cross-generalization in which the classifier was trained using all data from the bottom-up mapping task and then tested on the top-down attention task data. Decoding accuracy was calculated as the proportion of trials where the location predicted by the classifier matched the location where the target was presented, separately for cue conditions. Finally, as a control analysis, the classifier was trained and tested on the top-down attention task data in four folds, as task conditions were counterbalanced across five runs. That is, the first 15 out of 20 runs were used to train the classifier, using the target location as the labels, and then the classifier predicted the target location for each trial in the five runs that were left out. This procedure was repeated four times leaving out a different fold (five counterbalanced runs) each time. Decoding accuracy was calculated as the proportion of trials where the location predicted by the classifier matched the location where the target was presented, separately for cue conditions. As it was not our focus of the present study, contrast levels in the top-down attention task were not considered as a factor, and data were collapsed across contrast levels (see Supplementary Figures 1.3-1.4 for analysis results with contrast as an additional factor; also see Supplementary Figure 1.2 for univariate BOLD response changes across contrast).

Assessing spatial precision of top-down and bottom-up representations with decoding errors

To make further inferences on the qualitative aspects of the spatial representation based on the decoding errors in the above analyses, we acquired classifier predictions in each trial for each ROI and sorted them by the stimulated (in the bottom-up mapping task) or the cued (in the

top-down attention task) locations and organized the output into 12 x 12 matrices (Figure 1.4). For analysis and visualization purposes, we labeled spatial positions arbitrarily from 1 to 12, 1 being the leftmost position in the first quadrant and the number increasing in clockwise direction. To make quantitative comparisons between these confusion matrices, we performed a linear regression on each matrix with a diagonal and an off-diagonal regressor (Figure 1.5a). Both regressors were 12 x 12 matrices, and the diagonal regressor was created by assigning ones in the cells on the diagonal and zeros in the rest. The off-diagonal regressor was created by first assigning ones in the 3 x 3 groups of cells reflecting the quadrants of the spatial locations and zeros in the rest, and then subtracting the diagonal regressor from it so that there was no overlap between the two regressors. These regressors and a constant term were used as predictors in the linear regression model, which was fit to individual confusion matrices to obtain beta weights for each participant and regressor. Then, to compare these beta weights across ROIs and tasks/conditions, we computed ratios by dividing the diagonal beta weight by the off-diagonal beta weight and then log transforming the result (Figure 1.5b).

Statistical procedures for MVPA

To be conservative, all statistics reported regarding the MVPA results were based on non-parametric randomization tests over 1000 iterations (note that all results reported as significant are also significant using standard parametric approaches). To assess the significance of decoding accuracy in the bottom-up mapping task (Figure 1.2a), within each iteration of the leave-one-run-out procedure, the location labels for the test data were shuffled when computing the decoding accuracy for each ROI. Then, average decoding accuracy for each ROI was obtained by taking the mean across all runs. We repeated this procedure 1000 times to compute a distribution of decoding accuracies for each participant and each ROI under the null hypothesis

that the location of the stimulus had no impact on voxel activation patterns. To assess the significance of decoding accuracy at the group level, we then performed t-tests between the shuffled decoding accuracies across all participants and chance accuracy (1/12) to obtain a single distribution of 1000 t-values. We obtained p-values for each ROI by calculating the proportion of iterations on which the shuffled t-values exceeded the real t-value that was obtained using the

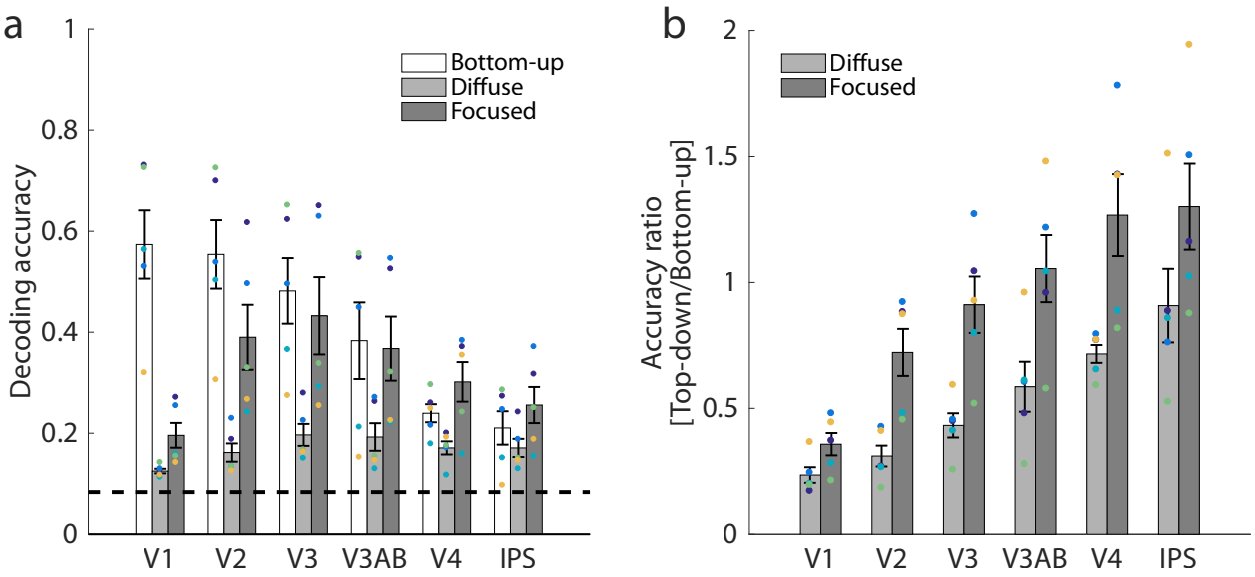


Figure 1.2 (a) Decoding accuracy based on fMRI activation patterns in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused condition; light and dark gray bars). In the bottom-up mapping task, decoding accuracy was highest in V1 and decreased in later visual areas. In the top-down attention task, decoding accuracy was generally higher in the focused than in the diffuse condition. While top-down decoding accuracy was much lower than bottom-up decoding accuracy in V1, accuracy in the bottom-up mapping task and the focused condition was comparable in later areas (e.g. V3AB, V4, IPS), leading to an interaction between task type (bottom-up vs top-down) and visual areas. Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. The dotted line indicates chance performance (1/12 or ~.083). (b) To better visualize the interaction between task type and visual areas, we computed the ratio between bottom-up and top-down decoding accuracies for each ROI. To obtain this ratio, the top-down decoding accuracy was divided by the bottom-up decoding accuracy within each ROI, separately for the diffuse and the focused conditions. A low ratio score indicates that an ROI had higher decoding accuracy in the bottom-up mapping task, consistent with higher precision of bottom-up representations. A high ratio score indicates that an ROI had higher decoding accuracy in the top-down task, consistent with higher precision of top-down representations. While V1 showed higher bottom-up precision, later areas showed comparable level of bottom-up and top-down precision. Colored dots represent data from individual participants. Error bars are ±1SEM.

unshuffled data. To assess whether decoding accuracies were significantly different across ROIs, we shuffled the ROI labels within each participant and performed one-way repeated measures ANOVA on the decoding accuracies with ROI as a factor to obtain 1000 F-values under the null hypothesis that ROI had no impact on decoding performance. The p-value was calculated as the proportion of shuffled F-values that exceeded the real F-value.

To assess the precision of top-down modulations in the attention task (Figure 1.2a), we first assessed the significance of classifier decoding performance in each ROI by collapsing data across cue conditions and comparing it to chance (1/12). We performed a randomization test by shuffling the location labels in the test (attention task) data 1000 times, and then we computed paired t-tests on the decoding accuracy against chance. The p-value for each ROI was obtained by computing the proportion of shuffled t-values that exceed the real t-value.

To evaluate the impact of the attention cue on the top-down precision, we performed a randomization test for three-way ANOVA on the decoding accuracy with cue and ROI as factors. First, we tested the main effect of the two factors. To test the effect of cue type, cue labels were shuffled across trials, restricted within each ROI and participant, and then the classifier correctness for each trial was averaged to obtain decoding accuracy for each condition. Then, we performed an ANOVA on this data, only testing for the two main effects without the interaction terms, to obtain the F-value for the main effect of cue. This procedure was repeated across 1000 iterations, which yields a distribution of 1000 F-values under the null hypothesis that cue conditions did not affect the decoding accuracy. A p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test the main effects of ROI.

Next, we tested the interaction effects between the two factors. We first performed an ANOVA on the decoding accuracy with only the main effects and obtained the residuals. This was done to exclude the variance explained by the main effects. To test the interaction between cue type and ROI, we shuffled the labels of the residual data, restricted within each participant. Then, we performed an ANOVA on the residuals, testing for the two main effects and the interaction effect, to obtain the F-value for the interaction effect between cue type and ROI. This procedure was repeated 1000 times, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels.

To further examine the significant interaction effects between cue type and ROI from the above analysis, we tested the effect of cue type within each ROI. We first shuffled the cue labels across trials within each ROI and participant, and averaged classifier accuracy across trials to obtain decoding accuracy for each cue condition. Next, we performed a paired t-test between the decoding accuracies from each cue condition to obtain the t-value for each ROI. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values for each ROI, under the null hypothesis that cue conditions did not affect decoding accuracy. Then, a p-value for each ROI was calculated by taking the proportion of iterations where shuffled t-values exceed the real t-value, the proportion of iterations where the real t-value exceeded the shuffled t-values, and then taking the minimum value and multiplying it by two for a two-tailed test.

To compare the difference between the precision of bottom-up and top-down modulations across ROIs (Figure 1.2b), we divided the decoding accuracy for each cue condition in the top-down attention task by the decoding accuracy in the bottom-up spatial mapping task, for each ROI. To assess the effect of cue and ROI on these ratio scores, we performed a

randomization test using a two-way ANOVA on the ratio scores with cue type and ROI as factors. First, we tested the main effect of the two factors on the ratio score. To test the effect of cue type, cue labels were shuffled across trials in the attention task data, restricted within each participant and ROI, and then the accuracy on each trial was averaged to obtain average accuracy for each condition. Then, decoding accuracies for each condition in the top-down attention task were divided by the decoding accuracy from the bottom-up mapping task. We performed the two-way ANOVA on this ratio scores, only testing for the main effects of cue and ROI, to obtain the F-value for the main effect of cue. This procedure was repeated 1000 times, yielding a distribution of 1000 F-values under the null hypothesis that cue conditions did not affect the ratio scores. A p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test the main effect of ROI, in which ROI labels for each trial were shuffled within each cue condition and participant, instead of cue labels. Next, we tested the interaction effect between cue and ROI. We first performed a two-way ANOVA on the ratio scores with only the main effects and obtained the residuals. This was done to exclude the variance explained by the main effects. To test the interaction between cue and ROI, we shuffled the labels of the residual data, restricted within each participant. Then, we performed the two-way ANOVA on the residuals, testing for the two main effects and the two-way interaction effect, to obtain the F-value for the interaction effect between cue type and ROI. This procedure was repeated 1000 times, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels.

To assess the significance of difference between task conditions and ROIs on the beta weight ratios from the confusion matrices obtained from the previous section, the same procedure of the two-way ANOVA as described in the above paragraph was applied to the beta weight ratios, but using task labels (Bottom-up, Diffuse, and Focused) instead of cue types.

Additional control and single-subject analyses

In all analyses reported, we performed MVPA using all visually responsive voxels from each ROI as defined using the functional localizer (the circular annulus stimulus, see section *Localizer task*). In the interest of making classification performance more comparable across ROIs, we also repeated all analyses after restricting the voxels to the top 300 voxels that were most responsive to the functional localizer in each ROI (ranked by F-values). Overall, reducing the number of voxels did not lead to significant changes in the pattern of decoding performance across ROIs (see Supplementary Figure 1.5).

In addition, to assess how many individual subjects also showed the same effects that we report in the group data, we also performed the randomization tests described above within each participant. This was done by shuffling the condition labels corresponding to the effect we want to test 1000 times, performing statistical tests (t-test or ANOVA) on the single-trial data, and then calculating the proportion of shuffled t- or F-values that exceed the statistical values from real data. We counted how many participants showed a significant result for each test and reported them in addition to the group-level statistics (Supplementary Table 1.1).

Results

We first report the behavioral results of the top-down attention task. Then, we present fMRI results from the bottom-up mapping task, estimating the precision of stimulus-driven responses in each visual area. Finally, we present fMRI results from the top-down attention task and compare them with the precision of the stimulus-driven responses to estimate the relative precision of both types of modulation.

To be conservative with our statistical tests, we evaluated significance using non-parametric randomization tests over 1000 iterations (see sections *Analysis of behavioral data* and *Statistical procedures for MVPA analysis* in Methods). In addition, we performed statistical analyses on single-subject data and reported the number of participants that showed significant results in the same direction as in the group-level analyses (see section *Additional control and single-subject analyses* in Methods and Supplementary Table 1.1).

Behavior

Participants (n=5), analyzed at the single-subject and group level, performed a top-down spatial attention task in which they were cued to the quadrant (Diffuse condition) or the exact location (Focused condition) of the target. When the target grating appeared, participants indicated whether its orientation was closer to a horizontal or a vertical orientation (Figure 1.1a). During a 6-8 second cue-to-target delay period (2 seconds on catch trials), a spatially uniform noise stimulus was presented. The contrast of the noise stimulus matched the contrast of the target, which varied across 5 different levels (0, 5, 10, 20, 50% Michelson contrast). To equate task difficulty across all cue and contrast conditions, we adjusted the task difficulty (tilt offset of the target relative to 45° or 135°; see section *Top-down spatial attention task* in Methods for

detail). Task difficulty adjustments were performed separately for each condition between every run to maintain accuracy at approximately 75% across all conditions.

Since there was no interaction between contrast and other factors (see Supplementary Figures 1.3-1.4), behavioral data were collapsed across contrast (see section *Analysis of behavioral data* in Methods). In addition to the scanning sessions, participants completed an additional eye-tracking session in which they performed the same attention task, and data from the eye-tracking session and the scanning sessions were combined (see Supplementary Figure 1.1 for behavioral results of individual sessions and Supplementary Figure 1.7 for eye-tracking results). A paired t-test between cue types (Diffuse vs. Focused) was performed separately on the accuracy and the tilt offset. The effect of cue on the tilt offset was significant, as the tilt offset was larger in the diffuse compared to the focused condition ($p < 0.05$; Figure 1.1d). Accuracy in the diffuse condition was also slightly higher compared to the focused condition (by 2% difference, $p < 0.01$; Figure 1.1c). These results indicate that participants were able to detect smaller perceptual changes in the focused condition compared to the diffuse condition. That said, there was a small change in accuracy that may contribute to the observed effect of attention on the size of the tilt offset. However, this tradeoff was not a major concern as the focus of the fMRI analyses was on the comparison of the relative precision between top-down and bottom-up representations as opposed to focused vs. diffuse attention.

Bottom-up spatial mapping task: estimating the precision of bottom-up stimulus-driven representations

To measure the spatial precision of stimulus-driven representations in retinotopically defined regions of interest (ROIs) in occipital and parietal cortex, we performed a separate bottom-up spatial mapping task in which participants viewed a series of small stimuli presented

in one of 24 locations around an imaginary circle centered on fixation (Figure 1.1b). To equate the number of spatial positions across the bottom-up and the top-down attention task, the bottom-up task data from adjacent wedges were collapsed to match the 12 target locations in the top-down attention task (analysis results without collapsing reported in Supplementary Figure 1.6). To control for any differences in the mean signal level across tasks, we calculated and subtracted the mean amplitude across all voxels on a trial-by-trial basis separately for each ROI, which centered the average amplitude across voxels on zero for both tasks. We then trained and tested a linear classifier on the activation patterns extracted from each visual area to decode the spatial location of the stimulus on each trial using leave-one-run-out cross-validation (see section *MVPA decoding* in Methods).

Decoding accuracy was above chance in all ROIs (Figure 1.2a, white bars; one-tailed t-test, all p's < 0.02, p-value computed via randomization test, see section *Statistical procedures for MVPA analysis* in Methods). However, decoding accuracy was significantly different across ROIs, and was numerically highest in V1 and numerically lower across subsequent areas of the visual hierarchy (one-way ANOVA on decoding accuracy across ROIs, $p < 0.001$). This pattern is consistent with prior reports indicating that the spatial precision of bottom-up sensory representations is most precise in V1, less precise in mid-level occipital regions, and lowest in IPS (see 22, 86).

Top-down attention task: estimating the spatial precision of top-down attentional modulations

Next, to estimate the spatial precision of top-down attentional modulations, we trained a classifier on the bottom-up mapping task and cross-generalized that classifier to decode attended

spatial locations during the cue-to-target delay period in the top-down attention task. We performed this cross-generalization analysis with the assumption that training the classifier using precise and focal mapping stimuli would allow us to estimate the spatial precision of top-down attentional modulations *relative* to the precision of bottom-up representations. Thus, a high degree of cross-generalization would suggest that top-down attentional modulations are as precise as bottom-up stimulus-driven representations. In contrast, a lower degree of cross-generalization would suggest that top-down modulations are less precise than the bottom-up representations observed in the bottom-up mapping task. Note that this logic holds only if the neural codes for spatial position share a common format across the bottom-up and top-down tasks. To account for a case in which this is not true, in a later section we report a complementary decoding analysis performed within the top-down attention task.

As with behavioral results reported above, all statistical analyses were performed after collapsing across contrast levels (analysis results without collapsing reported in Supplementary Figures 1.3-1.4). First, all ROIs showed cross-generalized decoding performance that was above chance when collapsed across cue conditions (Figure 1.2a, light and dark gray bars; one-tailed t-test against chance, p's < 0.01; see Supplementary Table 1.1 for the number of participants who showed significant effect when tested within individual participants). A two-way ANOVA on decoding accuracies with cue type (Diffuse vs. Focused) and ROI as factors revealed significant main effects of both factors (both p's < 0.001). There was also a significant interaction between cue type and ROI ($p < 0.001$). To further explore the cause of this interaction, we tested the effect of cue type within each ROI. Decoding accuracy was significantly higher in the focused than in the diffuse condition in all ROIs except V1 (two-tailed t-test, V1: $p = 0.07$, other ROIs: p's < 0.05).

To directly assess the interaction between ROI and the precision of top-down and bottom-up modulations, we next computed the ratio between the bottom-up and top-down decoding accuracy. For this analysis, top-down decoding accuracies were divided by the bottom-up accuracies for each ROI (Figure 1.2b). A low ratio score means higher bottom-up decoding accuracy compared to top-down decoding accuracy, consistent with higher precision of bottom-up representations. A high ratio score means higher top-down decoding accuracy compared to bottom-up decoding accuracy, consistent with higher precision of top-down representations. As shown in Figure 1.2b, V1 had the lowest ratio score, indicating that the bottom-up decoding accuracy was much higher than top-down decoding accuracy in V1. Critically, there was a gradual decrease in the magnitude of ratio scores across the visual hierarchy, with later regions showing a similar level of top-down and bottom-up decoding accuracy. For instance, the ratio scores in V3AB, V4, and IPS were close to or above one, demonstrating that top-down decoding accuracy is comparable to or even slightly higher than the bottom-up decoding accuracy in these areas. We also found that there was a relatively higher top-down decoding accuracy across areas in the focused cue condition. A two-way ANOVA on the ratio scores with cue type and ROI as factors revealed significant main effects of both factors (both p's < 0.001; see Supplementary Table 1.1 for the number of participants who showed significant effect when tested within individual participants), and a significant interaction between these factors ($p < 0.001$). Thus, even though the absolute value of decoding accuracies cannot be compared within a single ROI given differences in task parameters, the relative precision of bottom-up and top-down modulations changes systematically across visual areas.

Finally, to control for the possibility that top-down and bottom-up information about spatial position might be represented in different formats, we performed an additional analysis in

which the classifier was trained and tested using only data from the top-down attention task. V1 still yielded the lowest top-down decoding accuracy (Figure 1.3), and other ROIs followed the similar pattern observed in the cross-generalization results presented in Figure 1.2.

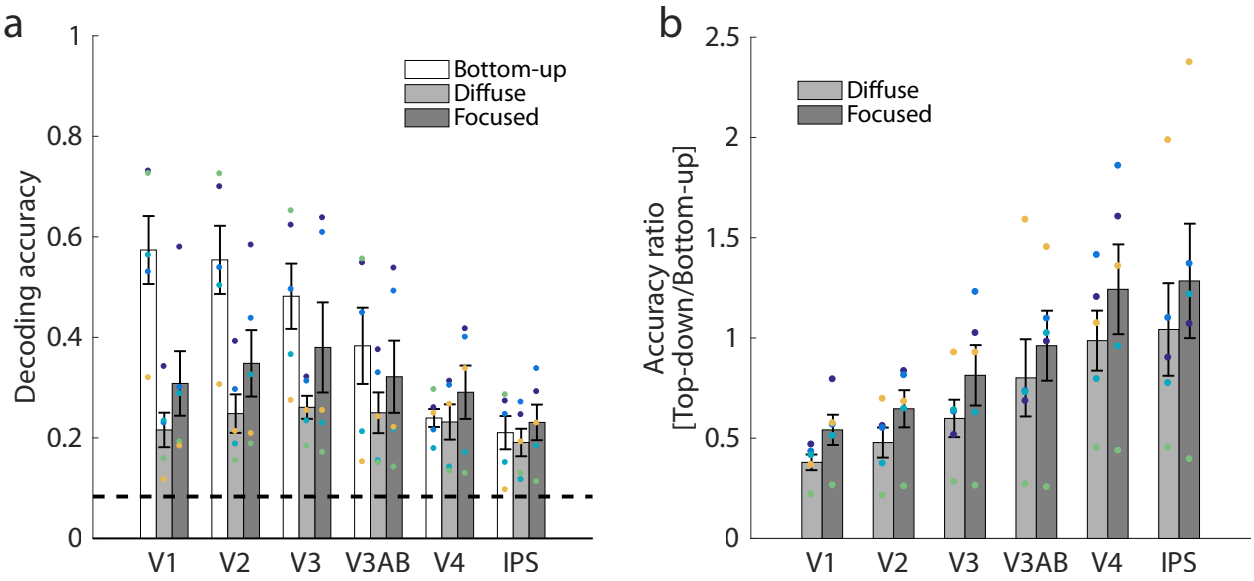


Figure 1.3 Similar to Figure 1.2 but only using data from the top-down attention task to train and test the classifier for the top-down decoding accuracy. (a) Decoding accuracy based on fMRI activation patterns in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused conditions; light and dark gray bars). For comparison, bottom-up decoding accuracies from Figure 1.2a are plotted together. The general pattern of results followed Figure 1.2a: decoding accuracy in the top-down attention task was generally higher in the focused than in the diffuse condition, albeit to a lesser degree. Comparing across ROIs, while V1 showed much lower top-down decoding accuracy relative to bottom-up decoding accuracy, later areas (e.g. V3AB, V4, IPS) showed comparable decoding accuracies across tasks, leading to an interaction between task type and visual areas. Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. The dotted line indicates chance performance (1/12 or ~.083). (b) Ratio of decoding accuracies between top-down attention and bottom-up mapping task. To obtain this ratio score, the top-down decoding accuracy for each ROI was divided by the bottom-up decoding accuracy in that ROI, separately for the diffuse and the focused conditions. A low ratio score indicates that an ROI had higher decoding accuracy in the bottom-up mapping task, consistent with higher bottom-up precision. A high ratio score indicates that an ROI had higher decoding accuracy in the top-down task, consistent with higher top-down precision. While V1 showed relatively higher bottom-up precision, later areas showed comparable level of bottom-up and top-down precision, showing a similar pattern of results to Figure 1.2b. Colored dots represent data from individual participants. Error bars are ±1SEM.

Complementary approach to comparing bottom-up and top-down spatial precision

As an additional assessment of bottom-up and top-down precision, we next quantified the confusability of decoded spatial positions in the bottom-up mapping and the top-down attention tasks because classifier errors should track the similarity between the activation patterns associated with nearby spatial positions. Thus, the pattern of decoding errors provides a more nuanced description of the relationship between voxel activation patterns associated with different spatial positions and the precision of those representations.

Figure 1.4 shows the group-averaged confusion matrices for the actual target locations plotted against the classifier-predicted locations in the bottom-up mapping and the top-down attention tasks. For the attention task, confusion matrices from the cross-generalization analysis (reported in Figure 1.2a) are plotted, and data were collapsed across contrast levels as well. Each cell within a matrix is color-coded to indicate the proportion of trials for which the classifier predicted the spatial position labeled by the row, out of all trials in which the target was presented at the spatial position labeled by the column. For example, for the bottom-up mapping task, in the most upper-left cell of the V1 confusion matrix, the high value marked with light green color (50.17%) indicates that out of all trials in which the target was presented at position 1, the classifier correctly predicted approximately half of those trials as position 1.

In the bottom-up mapping task and the focused condition in the top-down attention task, predictions were centered on and around the main diagonal, consistent with the classifier inferring the correct spatial position. In contrast, predictions in the diffuse condition showed a clustering pattern within the quadrant of the presented target. This clustering indicates that the

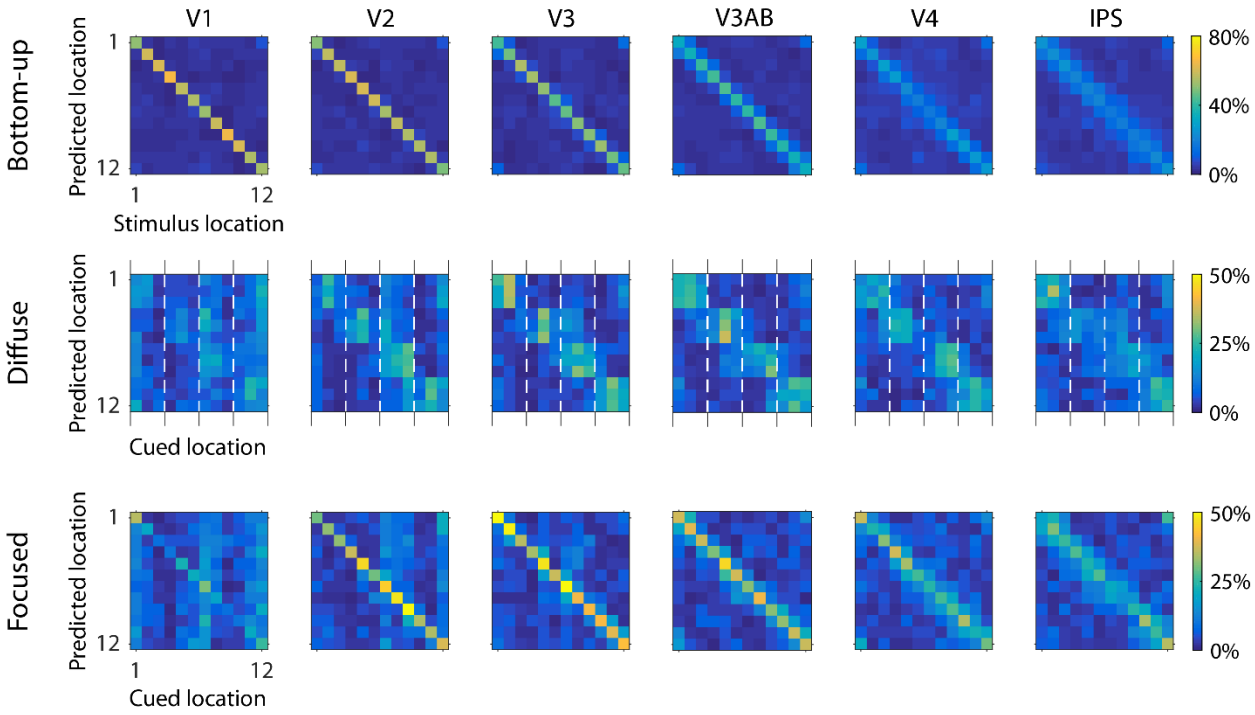


Figure 1.4 Confusion matrices of classifier predictions for the presented stimulus location in the bottom-up mapping task and the cued location in the top-down attention task. For analysis and visualization purposes, spatial locations were arbitrarily labeled from 1 to 12, 1 being the leftmost position in the first quadrant with numbers increasing in the clockwise direction. Each cell within the matrices was colored based on values within the range of 0-80% for the bottom-up mapping task and 0-50% for the top-down attention task conditions, as indicated in the color bars on the right (this was done to make patterns easier to discern). The vertical dotted white lines in the diffuse condition matrices (middle row) divide the 12 spatial locations into the 4 cued quadrants. In the bottom-up mapping task (top row), classifier predictions were clustered on the diagonal, where the predicted location closely tracked the actual stimulus location. In the focused condition of the top-down attention task (bottom row), the diagonal pattern was visible but to a lesser degree than that in the bottom-up mapping task. In the diffuse condition of the top-down attention task (middle row), classifier predictions were clustered within the cued quadrant, indicating that the attentional modulation was spread across the whole quadrant, consistent with subjects using the diffuse cue as intended.

classifier was distributing guesses more uniformly across the three locations within each quadrant, corresponding to the cued locations in the diffuse condition.

To quantify differences in representational precision reflected in the confusion matrices between tasks and task conditions, we used a general linear model with a diagonal and off-

diagonal predictors (Figure 1.5a). The diagonal predictor captures a narrow, precise spatial representation and the off-diagonal predictor captures a broad, quadrant-wide spatial representation. These two predictors were used as regressors on each of the confusion matrices in Figure 1.4 to calculate beta weights in each task/condition, separately for each participant. We then computed a ratio between the two beta weights by dividing the diagonal beta weight by the off-diagonal beta weight and log-transforming the resulting value. Thus, the ratio value should be near zero if both beta weights are similar, larger than zero if the diagonal weight is higher than the off-diagonal weight, and below zero if the diagonal weight is lower than the off-diagonal weight (Figure 1.5b).

The results were consistent with our earlier comparison of bottom-up and top-down decoding accuracies. Representations in the bottom-up mapping task were highly precise, with the beta weight ratios highest in V1 and gradually decreasing in later areas. Compared to the bottom-up representation, top-down spatial representations in the focused condition were less precise in earlier areas, especially in V1, but reached a comparable level of precision in later areas. Finally, top-down representations in the diffuse condition were not precise and spread across a quadrant, showing beta weight ratios close to zero across all ROIs. A two-way ANOVA on the beta weight ratios with task condition (Bottom-up vs. Diffuse vs. Focused) and ROI as factors supports this trend, with significant main effects of task condition and ROI, as well as a significant interaction between these factors (all p's < 0.001).

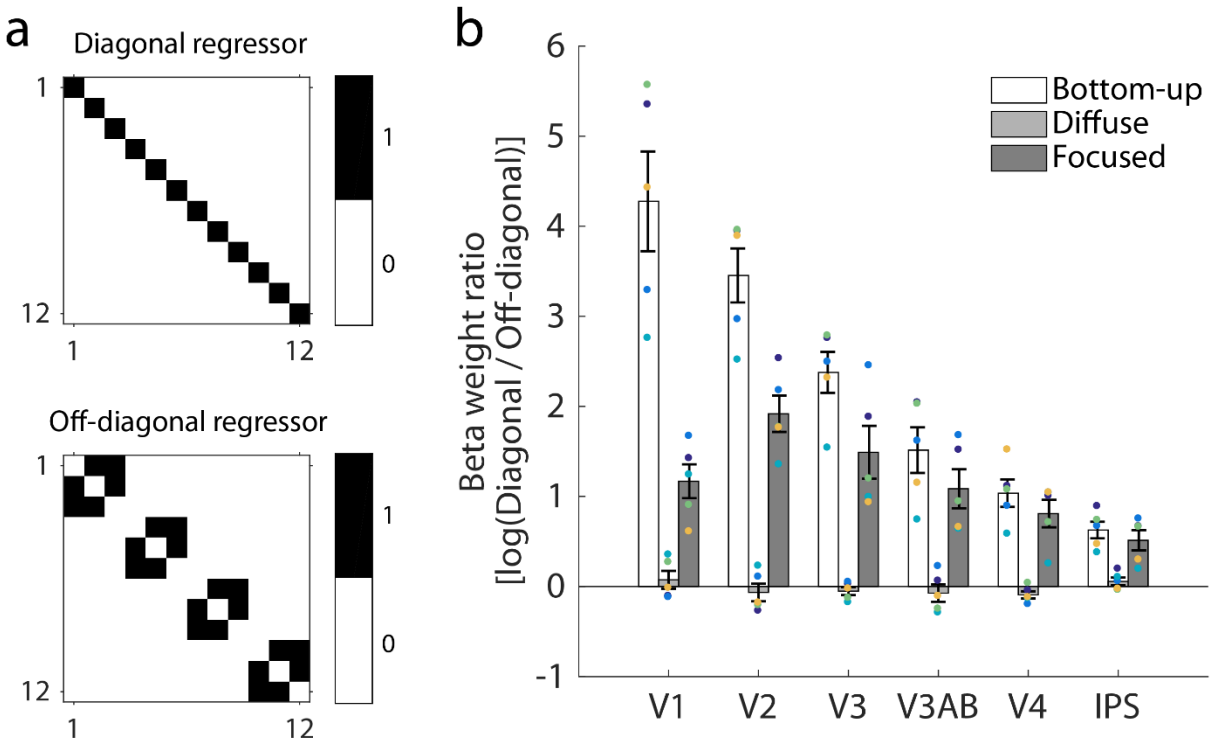


Figure 1.5 (a) Diagonal and off-diagonal regressors used to fit the confusion matrices shown in Figure 1.4. Cells colored black were assigned ones, and cells colored white were assigned zeros. (b) Ratio of beta weights in the bottom-up mapping task and the diffuse and focused conditions from the top-down attention task. The beta weights for the diagonal regressor were divided by the beta weights for the off-diagonal regressor and then plotted on a logarithmic scale. Beta ratios above zero indicate that the diagonal beta weight was larger than the off-diagonal beta weight, and beta ratios below zero indicate that the off-diagonal beta weight was larger than the diagonal beta weight. Beta weight ratios in the bottom-up mapping task were all above zero, with the highest value in V1 and gradually decreasing in later areas. Beta weight ratios in the diffuse condition from the top-down attention task were close to zero across all ROIs. Beta weight ratios in the focused condition from the attention task were above zero for all ROIs, but when compared to the ratios from the bottom-up task, the ratios for earlier areas were much smaller while ratios for later areas were at a similar level. Colored dots represent data from individual participants. Error bars are ±1SEM.

Discussion

Although top-down attention is critical for selecting and enhancing the processing of relevant sensory stimuli, the role of anatomical constraints in frontal and parietal areas (e.g., larger receptive fields) in determining the precision of top-down attentional modulations is not well understood. Here, we used a bottom-up mapping task and a top-down attention task to assess decoding accuracy based on voxel activation patterns in human visual areas. Because the tasks were – by necessity – different, we cannot directly compare decoding accuracies within a single visual area. However, the pattern of relative decoding accuracy across areas is revealing: Bottom-up decoding accuracy was significantly higher than top-down decoding accuracy in V1, but decoding accuracy was comparable in mid-level occipital areas and in IPS. These observations are consistent with the idea that the representational precision of the areas generating top-down control signals, coupled with the precision of mid-level areas that are the primary targets of these signals (61, 87), limits the precision of modulations in early parts of the visual system that have the highest bottom-up fidelity.

As a measure of precision, we used decoding accuracy to quantify the information content of spatial representations and the associated confusion matrices to provide a more descriptive assessment of the representational precision. Note that our measure of precision, decoding accuracy, and the pattern of decoding errors might be impacted by several different single-neuron modulations like a change of gain and/or tuning bandwidth in individual neurons. However, these underlying single-unit modulations jointly contribute to the overall amount of spatial information in the voxel activation patterns, and thus the decoding accuracy provides an overall assessment of aggregate changes in the representational precision in a visual area. In addition, the low degree of cross-generalization from the bottom-up mapping task to the top-

down attention task in V1 might be due to differences in stimulus attributes, such as spatial frequency, between the two tasks. However, Figure 1.3 reveals that the same pattern is observed even when evaluating top-down decoding accuracy using only data from the top-down attention task, in which all low-level stimulus attributes are equated. Furthermore, our decoding methods do not rule out the possibility that differences in the signal-to-noise ratio across brain areas or tasks contributed to the observed pattern of results. That said, we consider noise as one key factor that is partially responsible for the differences in spatial precision across ROIs. On this account, it should be noted that lower signal and/or higher noise could both contribute to lower overall spatial precision in this study.

Our initial hypothesis was that the spatial resolution of attentional modulations in early visual cortex would be limited due to the large receptive field sizes in attentional control areas where the top-down signals presumably originate. However, the precision of a spatial representation in a population of neurons is not just limited by the size of receptive fields. Instead, the spatial precision of a given area can be substantially higher than that of individual neurons via pooling information across many neurons (88–93) . However, even if precision gets higher with pooling, the precision of neural codes in areas with smaller RFs (and more neurons, such as V1) is still likely to reach a higher theoretical asymptote than areas with larger RFs (and fewer neurons). That said, this logic relies on assumptions about how the information is pooled and utilized by subsequent stages of processing (e.g., pooling may be sub-optimal, thus undercutting the theoretical upper limit of precision).

At first glance, our results may seem contradictory to previous studies that have reported evidence for highly detailed representations in V1 that rely on top-down feedback. For example, studies of visual working memory have shown that mnemonic representations held over a long

period of time can be highly precise in V1 and less so in later cortical regions, which is opposite to the pattern we observed here (e.g., 94–97). However, in these working memory studies, the initial presentation of the to-be-remembered sample stimulus likely evoked a highly precise bottom-up response that could be maintained with high fidelity during the delay period via top-down modulatory signals. In contrast, the present task provides a more direct assessment of the precision of purely top-down modulations, as there was no initial stimulus that evoked a spatially specific bottom-up response. Instead, participants were only given a symbolic central cue to guide the deployment of spatial attention, which enabled us to measure top-down attention that is not contaminated by bottom-up stimulus drive.

Our results likewise provide more information about the precision of top-down modulations by requiring attention to one of 12 contiguous locations, which is a more fine-grained manipulation of top-down attention than commonly used. For example, previous studies that focused on pure top-down processes have reported univariate BOLD modulations in early visual cortex, even before any physical stimulus is presented (26, 29). However, in these studies, univariate BOLD responses were averaged over a relatively broad region of visual cortex and participants attended to one of only few possible locations (e.g., left visual field vs right visual field, or one of the four quadrants). Thus, these prior studies do not directly reveal information about the precision of top-down modulations in situations where highly focused attention is required. Similarly, other studies of mental imagery have reported that multivariate activation patterns in early areas like V1 carry item-specific information even in the absence of a bottom-up stimulus (35, 38, 41–43). However, these mental imagery studies typically used a small set of categorical images, which means that the activation patterns only need to be sufficiently precise to successfully discriminate coarsely related exemplars. More recently, Breedlove et al. (98) used

an encoding model and found that identification accuracy for 64 visual images was higher for presented versus imagined images in early visual areas and comparable in later areas of the visual hierarchy such as IPS. Furthermore, Favila et al. (99) modelled population receptive fields (pRF) to map evoked responses during cued memory recall of spatial locations and found that recalled representations had lower spatial precision compared to representations during stimulus presentation in earlier areas of the visual hierarchy, while their precision was comparable in areas such as hV4 and later. These recent reports from studies on mental imagery and memory, which also involve top-down feedback signals, appear to converge with the present results that focus on spatial attention to specific locations in the visual field.

Finally, for the decoding analysis of spatial position on the diffuse cue trials, we used the presented target locations as the "correct" labels to compute the decoding accuracy. This would naturally lead to lower decoding performance for the diffuse cue condition, given that the attention manipulation results in a broad modulation across the whole quadrant, and the participant, unlike the classifier, does not know the target location until the target is presented. Thus, it is not surprising, due to the setup of the classifier, that the decoding accuracy for the exact target position is lower in the diffuse compared to the focused condition. However, we do not view this as problematic for several reasons. First, the results indicate that the classifier confused locations within a quadrant with about equal probability, consistent with participants spreading spatial attention appropriately across a quadrant in response to the diffuse cues (see Figure 1.4, middle row). Therefore, lower decoding accuracy in the diffuse condition accurately reflects lower spatial precision of the top-down representation. Second, our primary focus was on the comparison between bottom-up and top-down precision across visual areas, so the key findings rely primarily on the bottom-up mapping task and the focused cue trials.

In conclusion, our findings reveal an interaction between visual areas and the relative precision of bottom-up and top-down modulations, suggesting that top-down modulations are not as precise as bottom-up modulations in early visual areas compared to later areas. This is important because these early areas, particularly V1, have small receptive fields and are thus ideal candidates for high-fidelity top-down modulations. This comparative lack of top-down precision may be the result of relatively coarse spatial coding in frontal and parietal cortex and the anatomical pattern of feedback projections that primarily target mid-level visual areas as opposed to earlier areas with the highest potential for spatially focal representations.

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Journal of Neurophysiology, 127(2), 504-518, 2022, Park, Sunyoung; Serences, John T. The dissertation author was the primary investigator and author of this paper.

Supplementary Methods and Results

Quantifying univariate BOLD response changes in the attention task

To measure the change in the average magnitude of the BOLD response across all voxels in each ROI as a function of stimulus contrast and cue condition, we averaged BOLD responses from attended quadrants only. First, we selected voxels within each ROI that responded significantly to the stimulus locations in each quadrant. Using the mapping task data, we labeled the trials based on the quadrant that the wedge was presented in, and then using a similar procedure for identifying ROIs from the localizer task, performed a GLM with FSL FEAT. We contrasted beta estimates for one quadrant against estimates for all other quadrants and repeated this across all quadrants to select significant voxels for each quadrant ($p < 0.05$; FDR corrected). Next, we shifted all timepoints in the attention task by 4 seconds to account for the hemodynamic lag. Then, we averaged voxel responses during the cue-to-target delay period (6-8 s window after cue offset) for each trial in the attention task from the voxels that are selective for the cued quadrant, within each retinotopically defined ROI (Supplementary Figure 1.2).

To evaluate the impact of the attention cue and the contrast of the noise stimulus on the mean amplitude, we performed a randomization test for a three-way ANOVA on the mean BOLD responses with cue type, contrast level, and ROI as factors. Cue and ROI were considered as categorical factors, and contrast was considered as a continuous factor. First, we tested the main effect of the three factors. To test the effect of cue, cue labels were shuffled across trials, restricted within each contrast condition, ROI, and participant, and then averaged for each condition. Then, we performed ANOVA on this data, only testing for the three main effects without the interaction terms, to obtain the F-value for the main effect of cue. This procedure

was repeated across 1000 iterations, yielding a distribution of 1000 F-values under the null hypothesis that cue conditions did not affect the mean amplitude. A p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test the main effects of contrast and ROI.

Next, we tested the interaction effects between the three factors. We first performed an ANOVA with only the main effects and obtained the residuals. This was done to exclude the variance explained by the main effects. To test the interaction between cue and contrast, we shuffled the labels of the residual data, restricted within each ROI and participant. Then, we performed ANOVA on the residuals, testing for the three main effects and the three two-way interaction effects, to obtain the F-value for the interaction effect between cue and contrast. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test interaction effects between cue and ROI, and contrast and ROI. Finally, to test the three-way interaction between cue type, contrast level, and ROI, we first performed an ANOVA with the main effects and the three two-way interactions and obtained the residuals. Then, we shuffled the labels of the residual data without restriction and subjected the data to a full three-way ANOVA with the three main effects, three two-way interactions, and the three-way interaction. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels.

To further examine the interaction effect between cue type and contrast level, we performed a one-way ANOVA on the mean response amplitude with contrast level as a continuous factor, separately for each cue condition and within each ROI. We evaluated the statistical significance of this effect with a randomization test by shuffling the contrast label within each participant and each ROI and repeating the one-way ANOVA to obtain an F-value. Repeating this procedure 1000 times yielded a distribution of F-values under the null hypothesis that contrast had no impact on the responses. P-values were calculated by taking the proportion of F-values from the shuffled data that exceeded the F-value obtained using the unshuffled data. We also tested the effect of cue within each ROI by performing paired t-tests after collapsing across all contrast levels. We then repeated the t-test after shuffling the cue labels across 1000 iterations to obtain 1000 t-values under the null hypothesis that the attention-cue label had no effect on the mean BOLD responses. We then obtained p-values by calculating the proportion of shuffling iterations on which the shuffled t-value exceeded the real t-value, and the proportion on which the real t-value exceeded the shuffled t-value. We then took the minimum value between the two and multiplied it by two for a two-tailed test.

Assessing the effect of stimulus contrast on the top-down attentional representation

While not our primary manipulation of interest, we included the noise stimulus during the cue-to-target delay period in the attention task to determine if the precision of top-down attentional modulations interacted with the strength of concurrent bottom-up sensory drive. For example, a focused attention cue is expected to produce a more precise spatial representation than a diffuse cue. However, this relative increase in precision with focal attention might be further amplified (or dampened), if the top-down attention signals interact with the strength of bottom-up stimulus drive.

To evaluate the impact of the attention cue and the contrast of the noise stimulus on the top-down precision, we performed a randomization test for three-way ANOVA on the decoding accuracy with cue, contrast, and ROI as factors. Cue and ROI were considered as categorical factors, and contrast was considered as a continuous factor. First, we tested the main effect of the three factors. To test the effect of cue type, cue labels were shuffled across trials, restricted within each contrast condition, ROI, and participant, and then the classifier correctness for each trial was averaged to obtain decoding accuracy for each condition. Then, we performed an ANOVA on this data, only testing for the three main effects without the interaction terms, to obtain the F-value for the main effect of cue. This procedure was repeated across 1000 iterations, which yields a distribution of 1000 F-values under the null hypothesis that cue conditions did not affect the decoding accuracy. A p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test the main effects of contrast and ROI.

Next, we tested the interaction effects between the three factors. We first performed an ANOVA on the decoding accuracy with only the main effects and obtained the residuals. This was done to exclude the variance explained by the main effects. To test the interaction between cue type and contrast level, we shuffled the labels of the residual data, restricted within each ROI and participant. Then, we performed an ANOVA on the residuals, testing for the three main effects and the three two-way interaction effects, to obtain the F-value for the interaction effect between cue type and contrast level. This procedure was repeated 1000 times, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels. The same procedure was repeated to test the other interaction effects between cue type

and ROI, and contrast level and ROI. To test the three-way interaction between cue, contrast, and ROI, we first performed ANOVA on the decoding accuracy with the main effects and the three two-way interactions and obtained the residuals. Then, we shuffled the labels of the residual data without restriction and subjected the data to a full three-way ANOVA with the three main effects, three two-way interactions, and the three-way interaction. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values, and a p-value was obtained by taking the proportion of F-values from this null distribution that exceeded the F-value obtained using the data with unshuffled labels.

To further examine the significant interaction effects (between cue type and ROI, contrast level and ROI) from the above analysis, we tested the effect of cue type and contrast level within each ROI. To test the effect of cue in individual ROIs, we first shuffled the cue labels for across trials within each ROI and participant, and averaged classifier accuracy across trials to obtain decoding accuracy for each cue condition. Next, we performed a paired t-test between the decoding accuracies from each cue condition to obtain the t-value for each ROI. This procedure was repeated across 1000 iterations, yielding a distribution of 1000 F-values for each ROI, under the null hypothesis that cue conditions did not affect decoding accuracy. Then, a p-value for each ROI was calculated by taking the proportion of iterations where shuffled t-values exceed the real t-value, the proportion of iterations where the real t-value exceeded the shuffled t-values, and then taking the minimum value and multiplying it by two for a two-tailed test. A similar procedure was repeated to test the effect of contrast level within each ROI, but we shuffled the contrast labels and used a one-way ANOVA with contrast as a continuous variable. In this case, the p-value for each ROI was calculated as the proportion of shuffled F-values that exceeded the real F-value.

Eye-tracking session

Eye-tracking data were collected in an entirely separate session to the MRI scanning session from four out of five fMRI participants who were available to come into the lab. An SR Eyelink 1000+ eye-tracker (SR Research) was used to collect continuous gaze-position data at a sampling rate of 1000 Hz. Participants performed only the top-down attention task and completed the same number of trials as in the scanner. All aspects of the task were the same, except that it was self-paced, which means that each trial terminated when the response was made, and participants initiated the next trial. Tilt offset for each condition was manually adjusted after every 27 trials to keep accuracies around 75%.

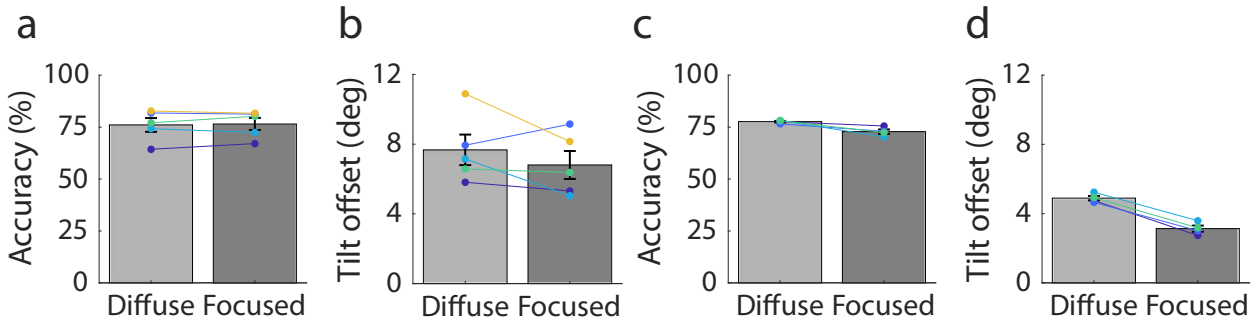
Analysis of eye-tracking data

Gaze positions during each run were preprocessed by first removing data points during blinks and filling in those data points by interpolating between eye positions before and after the blink. Then, separately for each run, gaze positions were detrended, high-pass filtered at 0.1Hz, and low-pass filtered at 10Hz. Next, gaze positions for each trial were corrected by subtracting the mean position during the last 200-ms in the pre-cue period, to account for task-unrelated drifts between trials and runs.

To examine eye movement away from fixation, average gaze positions during pre-cue, cue, delay period, and target presentation was calculated for each trial. To test if the eye movement away from fixation was in the direction of the cued location in each trial, the gaze position vector was projected onto the vector of the cued location to attain the magnitude of the projected vector (Supplementary Figure 1.7a).

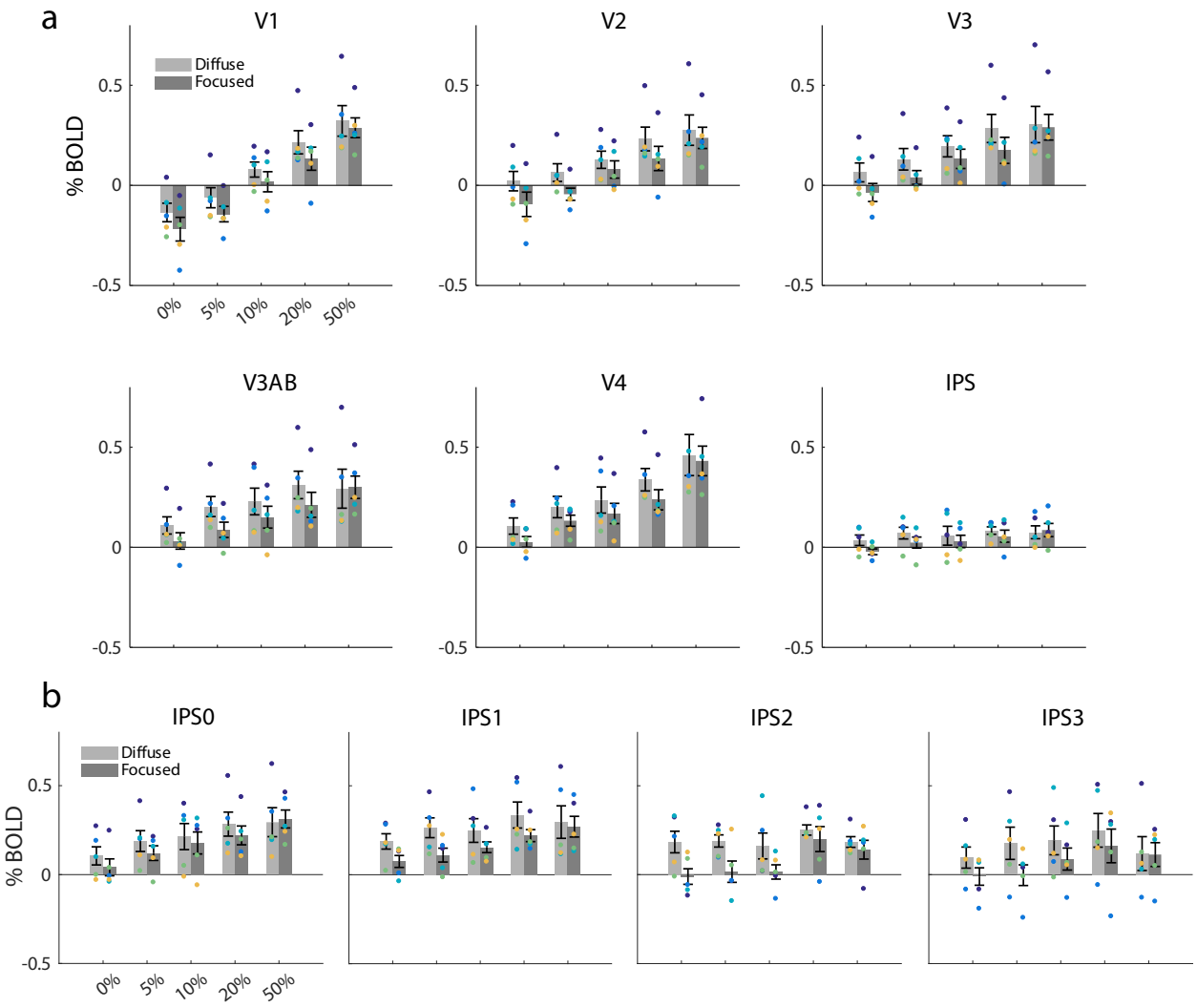
Control analysis for voxel selection using data from the top-down attention task

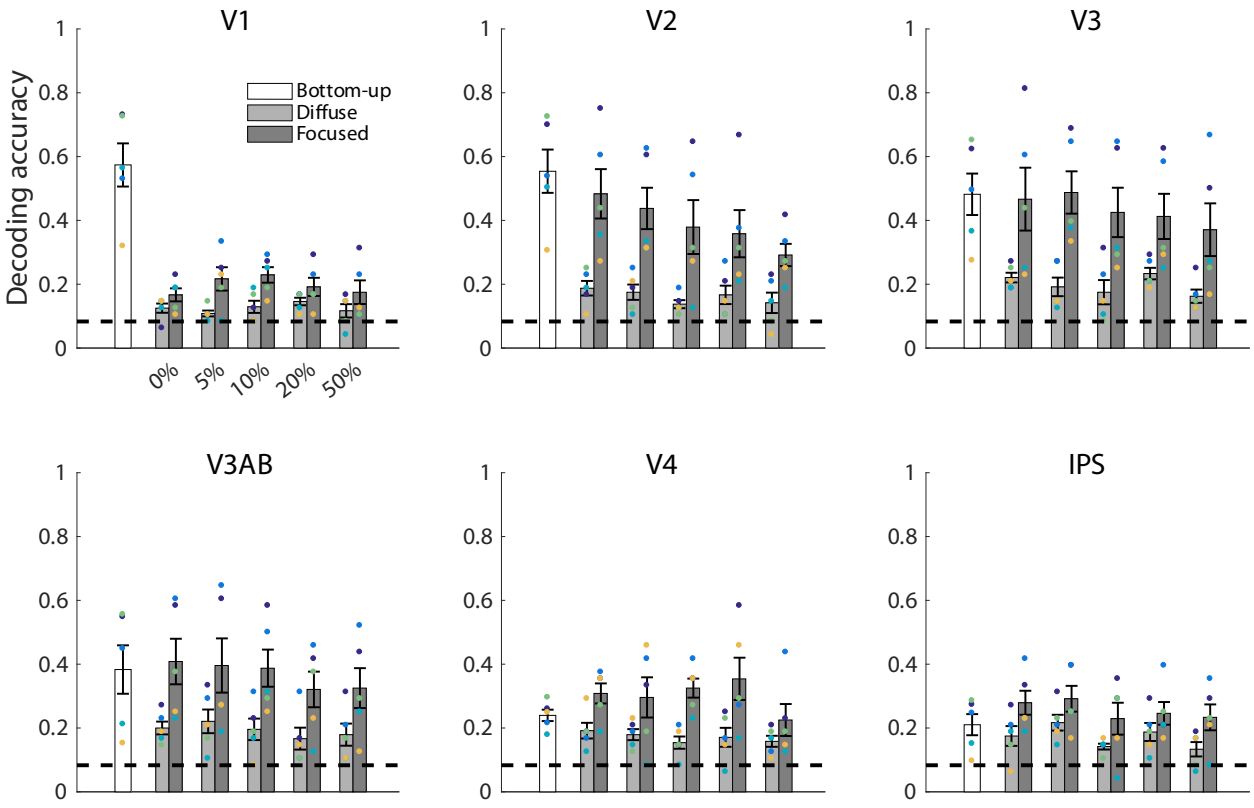
Our main analysis used a separate localizer task to select visually responsive voxels for the decoding analyses. Because the localizer task used flickering checkerboard stimuli, there is a chance that this process might have biased our pool of selected voxels in favor of bottom-up processing/representations that could have inflated bottom-up decoding accuracy in early visual areas. To address this possibility, we conducted control analyses with two alternative voxel selection methods using only the top-down attention task to select the voxels. First, we re-labeled the trials of the top-down task based on the quadrant that was cued and used all runs to run one-way ANOVA for individual voxels within each retinotopically defined ROI and found voxels that were significantly modulated across the cued quadrants during the delay period between the cue and the target ($p < 0.05$; FDR corrected; see Supplementary Table 1.2 for the number of selected voxels). After a single set of voxels has been selected for each ROI, the same procedure for the decoding analysis described in the main text was carried out. Note that this method is circular because data from all runs were used in the ANOVA to select the most selective voxels. This was done intentionally as we wanted to bias this analysis in favor of finding high-precision representations in the top-down attention task (see Supplementary Figure 1.8 for decoding results). In addition to this circular analysis, we also used a second method of voxel selection that avoided this circularity by using a leave-one-run-out procedure: All but one run of the top-down task was used in the voxel selection process and the decoder was tested only on the left out run, and this was repeated leaving out a different run on each iteration (20 iterations total). The bottom-up decoding, in this case, was done on every set of selected voxels and the resulting decoding accuracy was averaged across iterations (see Supplementary Figure 1.9 for decoding results).



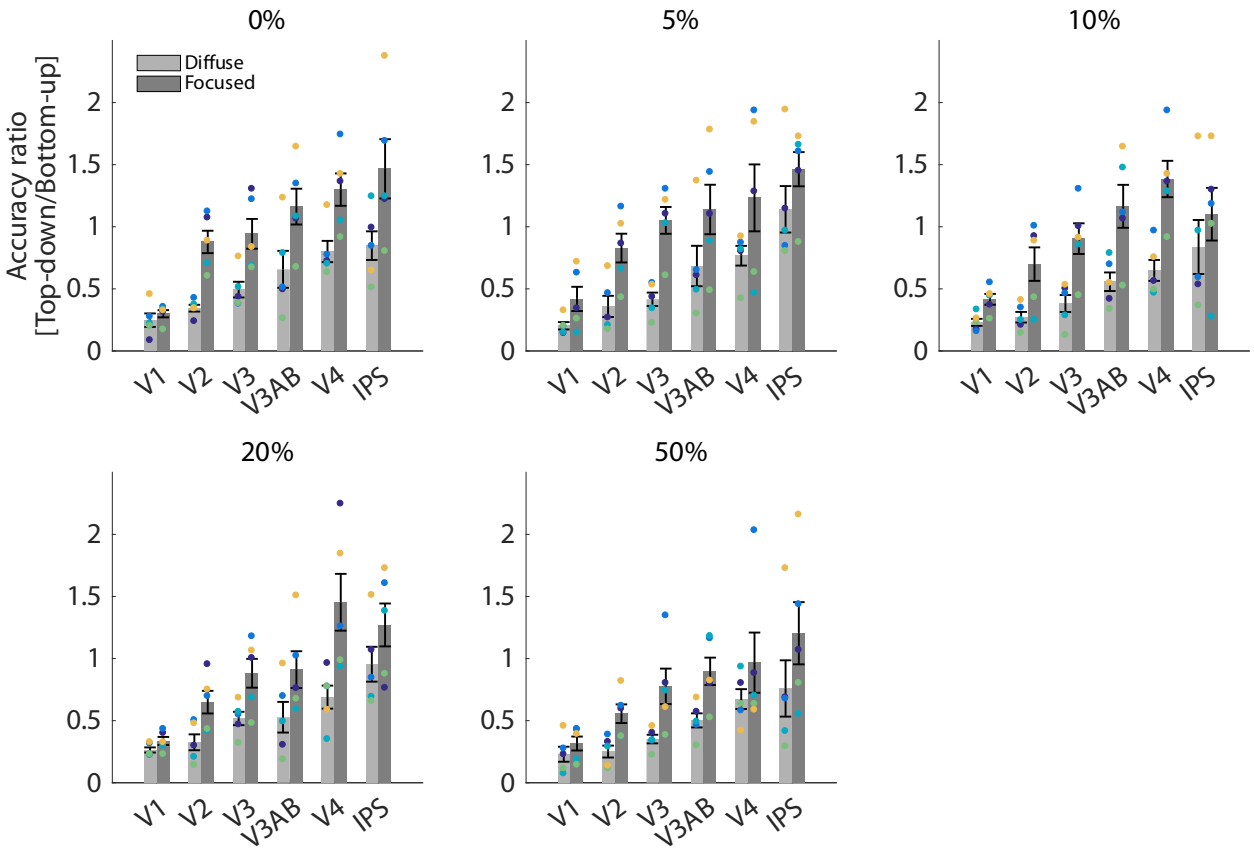
Supplementary Figure 1.1 Behavioral data for the top-down spatial attention task during the scanning and the eye-tracking session plotted separately. (a, b) Behavioral performance for the top-down spatial attention task during the scanning session. (a) Mean behavioral accuracy. Accuracy was comparable between the diffuse and the focused conditions. (b) Mean tilt offset. Tilt offsets were higher in the diffuse than in the focused condition. The effect of cue type on accuracy ($p = 0.33$) and on the tilt offset ($p = 0.12$) was not statistically significant. (c, d) Behavioral performance for the top-down spatial attention task during the eye-tracking session. (c) Mean behavioral accuracy. Accuracy was comparable between the diffuse and the focused conditions. (d) Mean tilt offset. Tilt offsets were higher in the diffuse than in the focused condition. The effect of cue type on accuracy and on the tilt offset were both significant (p's < 0.001). Colored dots represent data from individual participants. Error bars represent ±1 standard error of the mean (SEM).

Supplementary Figure 1.2 (a) Univariate BOLD amplitude changes in the top-down spatial attention task. Mean BOLD amplitude was higher in the diffuse than in the focused condition, and it increased as a function of the contrast level of the noise stimulus that was presented during the cue-to-target interval. Mean response amplitude increased as a function of stimulus contrast. A three-way ANOVA on the mean BOLD amplitude with cue condition, contrast level, and ROI as factors revealed significant main effects of all three factors (all p's < 0.01, p-values computed via randomization test, see Supplementary Methods, section *Quantifying univariate BOLD response changes in the attention task*). There were also significant interaction effects between cue and contrast ($p < 0.001$) and contrast and ROI ($p < 0.001$), but other interaction effects did not reach significance (cue X ROI: $p = 0.13$; cue X contrast X ROI: $p = 0.81$). To determine the direction of the interaction between cue and contrast, we performed a one-way ANOVA with contrast level as a continuous factor, separately for each cue condition within each ROI. This revealed a significant effect of contrast in all ROIs for both diffuse and focused conditions, except for IPS in the diffuse condition (all ROIs except for IPS in both cue conditions: p's < 0.001; IPS in diffuse: $p = 0.16$; IPS in focused: $p < 0.01$). Colored dots represent data from individual participants. Error bars represent ±1SEM. (b) Univariate BOLD amplitude changes in the top-down spatial attention task for IPS sub-regions. When collapsed across cue conditions, the effect of contrast was significant in IPS0 and IPS1 in both cue conditions (p's < 0.05) and IPS2 and IPS3 in the focused condition (p's < 0.05). However, stimulus contrast did not have a significant effect in IPS2 and IPS3 in the diffuse condition ($p = 0.67$ and $p = 0.51$, respectively). When collapsed across contrast levels, the effect of cue was significant in IPS1 ($p < 0.001$) but did not reach significance in the rest (IPS0: $p = 0.13$, IPS2: $p = 0.12$, IPS3: $p = 0.13$). Colored dots represent data from individual participants. Error bars represent ±1SEM.

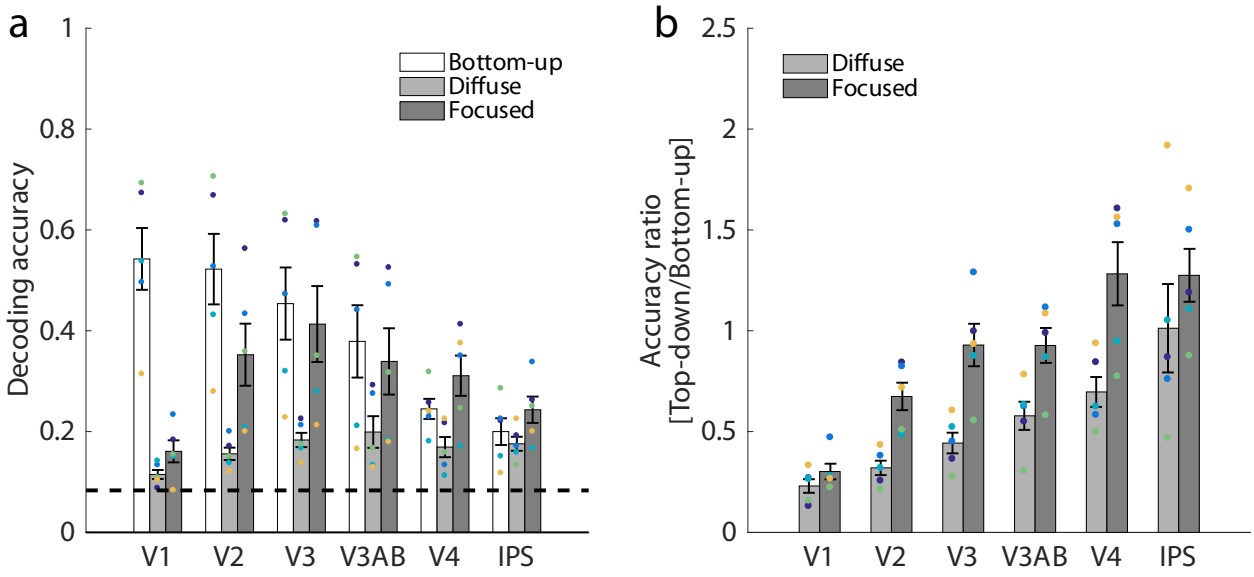

a
V1
V2
V3
V3AB
V4
IPS
Diffuse
Focused
% BOLD
0.5
0
-0.5
0%
5%
10%
20%
50%
b
IPS0
IPS1
IPS2
IPS3



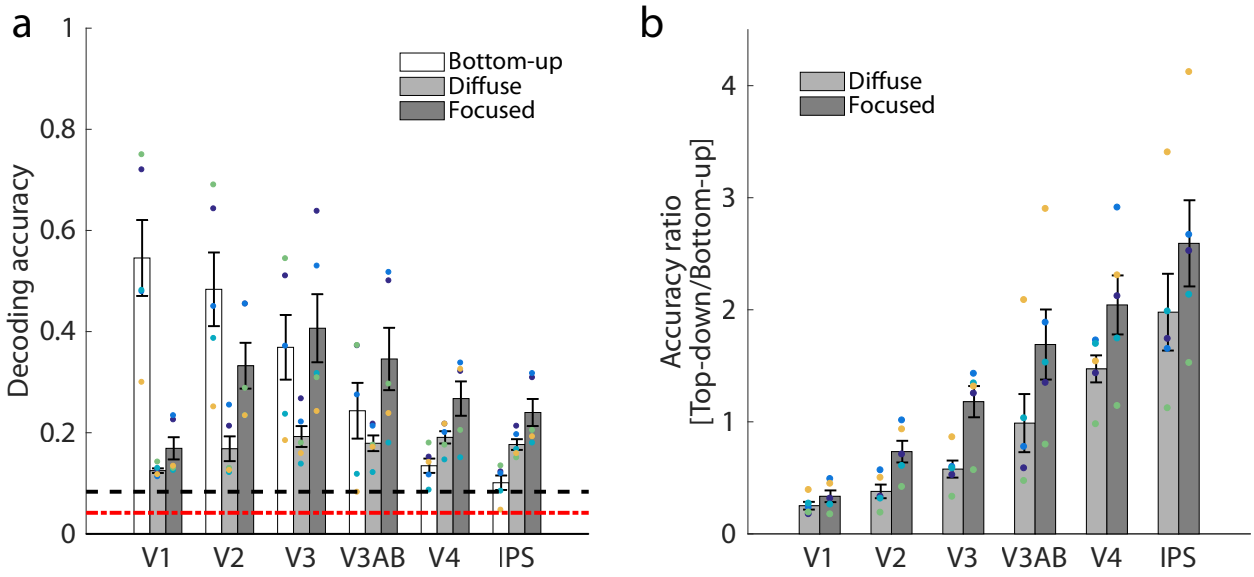
Supplementary Figure 1.3 Decoding accuracy in the top-down spatial attention task, plotted separately in each area for each contrast level and cue condition. For comparison, the same values for bottom-up decoding accuracies from Figure 1.2a are plotted together (white bars). Decoding accuracy in the attention task were generally higher in the focused than in the diffuse condition, and decreased as a function of contrast in V2, V3, V3AB, and IPS. While top-down decoding accuracy was much lower than bottom-up decoding accuracy in V1, accuracy in the mapping task and the focused condition were comparable in later areas (e.g. V3AB, V4, IPS). A three-way ANOVA on the mean decoding accuracies with cue condition, contrast level, and ROI as factors revealed significant main effects of all three factors (all p's < 0.01). There were also significant interactions between cue type and ROI ($p < 0.01$) and between contrast level and ROI ($p < 0.05$). Other interaction effects did not reach significance (cue X contrast: $p = 0.10$; cue X contrast X ROI: $p = 0.12$). To further explore the cause of the significant interactions, we tested the effect of cue and contrast within each ROI. When collapsed across contrast levels, decoding performance was significantly higher in the focused than in the diffuse condition in all ROIs except V1 (two-tailed t-test, V1: $p = 0.07$, other ROIs: p's < 0.05). Next, we evaluated the impact of the noise stimulus contrast presented during the cue-to-target delay period, by performing a one-way ANOVA with contrast level as a continuous variable after collapsing across cue conditions. This revealed a significant decrease in decoding accuracy as a function of contrast in V2, V3, V3AB, and IPS (p's < 0.05). This effect may have been due to higher contrast noise stimuli functioning as distractors that interfered with the maintenance of attention at the cued location(s). Colored dots represent data from individual participants. Error bars represent ±1SEM. The dotted line indicates chance performance (1/12, or ~.083).



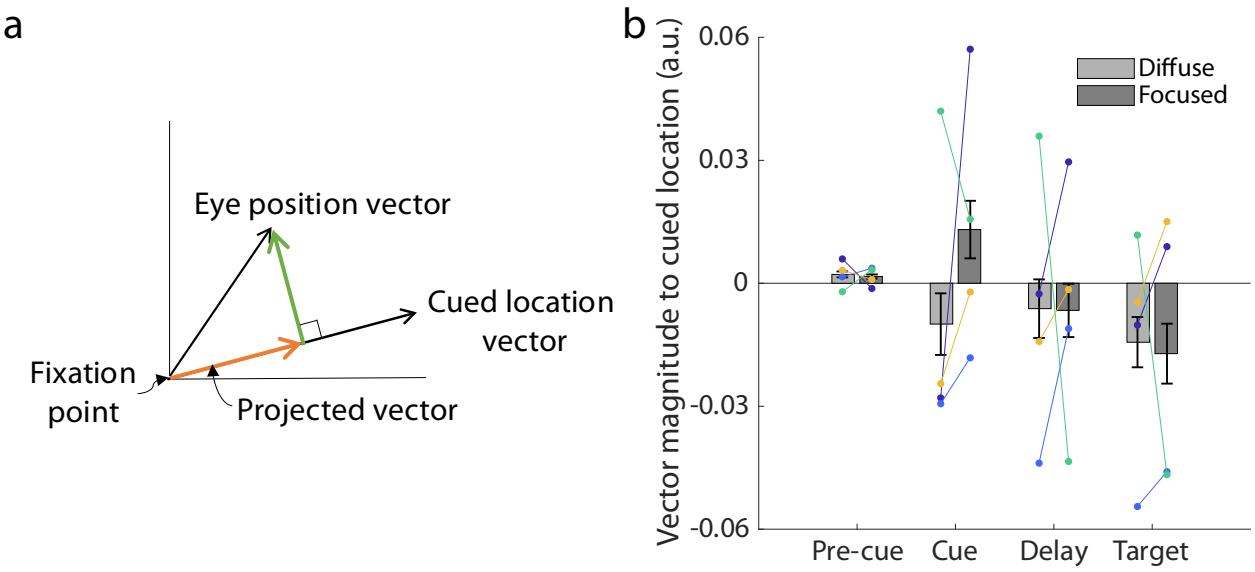
Supplementary Figure 1.4 Ratio of decoding accuracies between top-down attention and bottom-up mapping task, plotted separately for each contrast level. Ratio scores were obtained following the same procedure as described in Figure 1.2b but done separately for each contrast level instead of collapsing across the contrast levels. Low ratio scores indicate that the ROI had higher decoding accuracy in the mapping task, consistent with higher bottom-up precision. High ratio scores indicate that the ROI had higher decoding accuracy in the top-down task, consistent with higher top-down precision. Ratio scores for each contrast level showed a similar pattern to the collapsed result (shown in Figure 1.2). Colored dots represent data from individual participants. Error bars are ±1SEM.



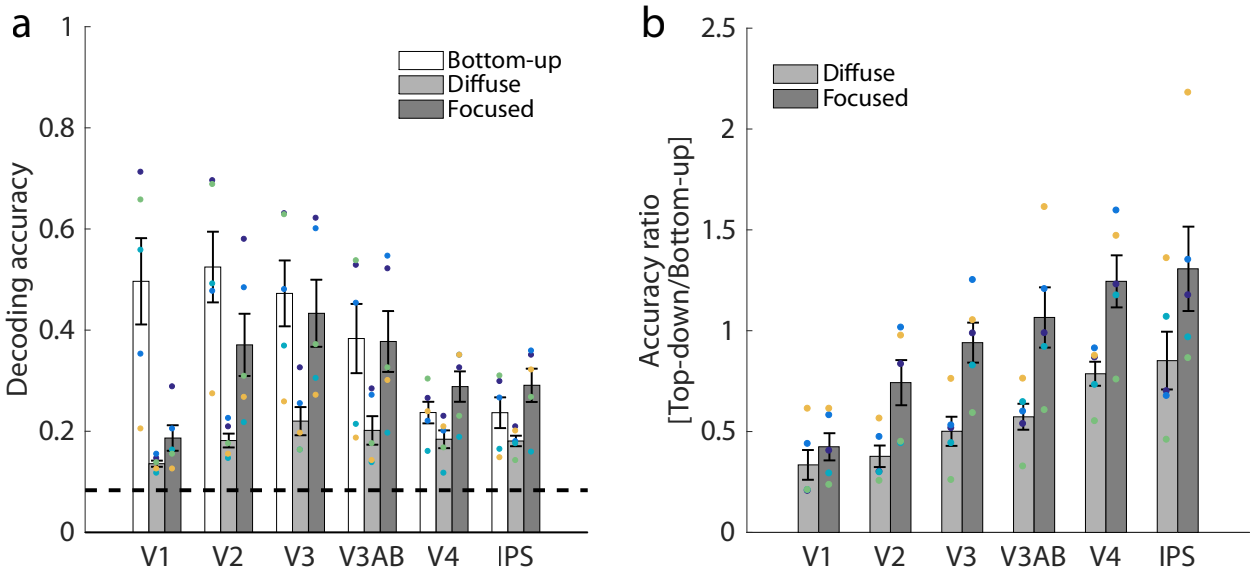
Supplementary Figure 1.5 (a) Decoding accuracy from top 300 voxels in each ROI in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused conditions; light gray and dark gray bars). Decoding accuracies showed a similar pattern to when the number of voxels were not restricted (shown in Figure 1.2a). Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. The dotted line indicates chance performance (1/12 or ~.083). (b) Ratio of decoding accuracies from top 300 voxels in each ROI between the bottom-up mapping task and the top-down attention task in each ROI, obtained using the same procedure as in Figure 1.2b. While V1 showed higher bottom-up precision, later areas showed comparable level of bottom-up and top-down precision, similar to Figure 1.2b. Colored dots represent data from individual participants. Error bars are ±1SEM.



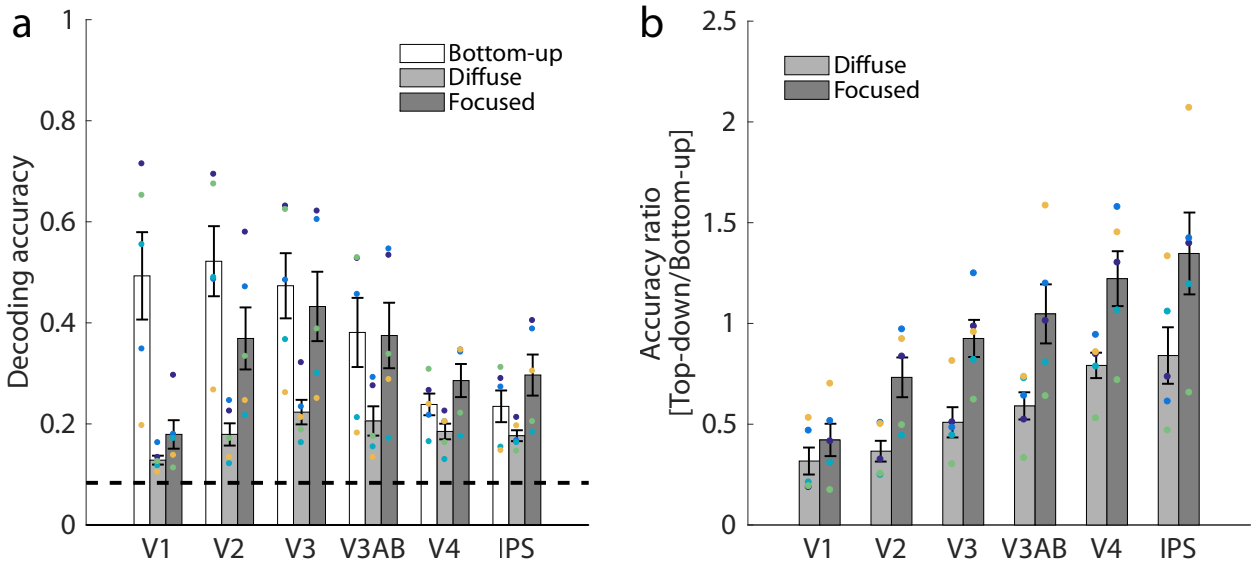
Supplementary Figure 1.6 Similar results to Figure 1.2 when using 24-way decoding of the mapping task. Classifiers were trained on the mapping task data without combining the wedges. Decoding accuracy for the mapping task was calculated by comparing the predicted location and the presented location. Decoding accuracy for the attention task was calculated by comparing the predicted location and the attended location, and a given trial was marked as correct if the predicted location from the 24-way decoder was included in the attended location. The general pattern of results was similar to Figure 1.2. (a) Decoding accuracy based on fMRI activation patterns in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused; gray and dark gray bars). Note that the chance level of decoding accuracy is different across tasks: 1/24 for mapping (red, dash-dot line), 1/12 for attention task (black, dashed line). While top-down decoding accuracy was much lower than bottom-up decoding accuracy in V1, accuracy in the mapping task and the focused condition was comparable in later areas (e.g. V3AB, V4, IPS), leading to an interaction between task type (bottom-up vs top-down) and visual area. Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. (b) Ratio of decoding accuracies between top-down attention and bottom-up mapping task, obtained using the same procedure as in Figure 1.2b. While V1 showed higher bottom-up precision, later areas showed top-down precision that is comparable or even higher than bottom-up precision. Colored dots represent data from individual participants. Error bars are ±1SEM.



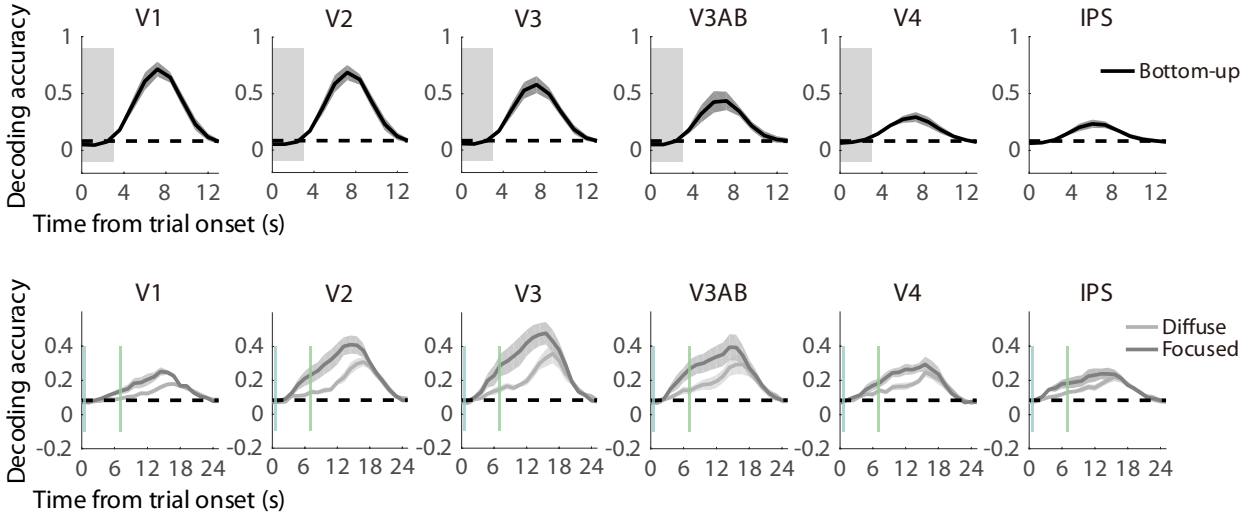
Supplementary Figure 1.7 Eye-tracking results from an independent eye-tracking session conducted outside of the scanner. Eye-tracking data were collected form four out of five fMRI participants. (a) An example of the vector projection calculation. A vector for average eye position was projected onto the vector for the cued location for that trial. The magnitude of the projected vector (orange) was calculated for each period (pre-cue, cue, delay, and target) within each trial. (b) Average magnitude of the projected vector towards the cued location. A larger value means that the gaze moved away from fixation in the direction of the cued location. Although overall eye movement away from fixation was larger for focused condition compared to diffuse condition during delay and target presentation, these eye movements were not systematically related to the cued location (all p's > 0.1).



Supplementary Figure 1.8 Similar results to Figure 1.2 when using another way to select voxels based on the top-down task. To bias the selection in favor of the top-down decoding as much as possible, all runs of the top-down task were used to select voxels that changed their response significantly across the cued quadrants during the delay period. The general pattern of results was similar regardless. (a) Decoding accuracy based on fMRI activation patterns in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused; gray and dark gray bars). While top-down decoding accuracy was much lower than bottom-up decoding accuracy in V1, accuracy in the mapping task and the focused condition was comparable in later areas (e.g. V3AB, V4, IPS), leading to an interaction between task type (bottom-up vs top-down) and visual area. Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. The dotted line indicates chance performance (1/12 or ~.083). (b) Ratio of decoding accuracies between top-down attention and bottom-up mapping task, obtained using the same procedure as in Figure 1.2b. While V1 showed higher bottom-up precision, later areas showed comparable level of bottom-up and top-down precision, similar to Figure 1.2b. Colored dots represent data from individual participants. Error bars are ±1SEM.



Supplementary Figure 1.9 Similar results to Figure 1.2 when selecting voxels based on the top-down task. As a control analysis, instead of using the separate localizer task to define visually responsive voxels, voxels that changed their response significantly across the cued quadrants during the delay period in the top-down attention task were used in the decoding analysis. To avoid circularity in the top-down decoding accuracy, a leave-one-run-out method was used, in which the run that was not used in the voxel selection step was used as a test dataset in the decoding analysis (see section Control analysis for an alternative voxel selection method in the Supplementary Methods). (a) Decoding accuracy based on fMRI activation patterns in the bottom-up spatial mapping task (white bars) and the top-down spatial attention task (Diffuse and Focused; gray and dark gray bars). While top-down decoding accuracy was much lower than bottom-up decoding accuracy in V1, accuracy in the mapping task and the focused condition was comparable in later areas (e.g. V3AB, V4, IPS), leading to an interaction between task type (bottom-up vs top-down) and visual area. Filled, colored dots represent data from individual participants, and error bars represent ±1SEM. The dotted line indicates chance performance (1/12 or ~.083). (b) Ratio of decoding accuracies between top-down attention and bottom-up mapping task, obtained using the same procedure as in Figure 1.2b. While V1 showed higher bottom-up precision, later areas showed comparable level of bottom-up and top-down precision, similar to Figure 1.2b. Colored dots represent data from individual participants. Error bars are ±1SEM.



Supplementary Figure 1.10 Average timecourse of decoding accuracy during bottom-up mapping task and top-down attention task. For the decoding analyses presented in the main text, we opted to shift the timecourse of the BOLD response by 4 seconds to account for the hemodynamic delay of the BOLD signal based on prior fMRI studies, based on the assumption that a canonical HRF peaks around 4-6 seconds after the stimulus onset. However, as actual HRF may differ by task and brain area, it is possible that our choice of the time window might have not captured relevant signals. Therefore, we looked at the un-shifted timecourse of the decoding accuracy for both bottom-up and top-down tasks. The timecourse was calculated by performing the decoding analyses on a sliding window of three TRs (2.4 s) and averaging across all trials (excluding short-delay catch trials for top-down task). In the upper panels, the grey shaded area indicates the duration of wedge stimulus presentation (0-3 s from trial onset) in the bottom-up spatial mapping task. In the lower panels, the blue bar indicates the onset of the cue (0.8 s from trial onset), and the green bar indicates the earliest onset of the target stimulus (6.95 s from trial onset). Shaded area around each line indicates ±1SEM.

Supplementary Table 1.1 Number of participants (out of 5 total) who showed significant effects ($p < 0.05$) from the main statistical tests reported when analyzed at a single-subject level with trial as the unit of variance.

Two-way ANOVA on decoding accuracy

Cue	5
ROI	5
Cue X ROI	3

Effect of cue within each ROI

	V1	V2	V3	V3AB	V4	IPS
Cue	2	5	5	5	3	3

Two-way ANOVA on accuracy ratio scores

Cue	5
ROI	0
Cue X ROI	2

Supplementary Table 1.2 Average number of voxels selected for the main analysis based on the functional localizer task and the supplementary analyses based on the top-down attention task. Values are mean number of voxels selected for each retinotopic ROI across all subjects ±1SEM.

ROI	*defined using localizer task (used in all main analyses)*	*defined using all runs of top-down task (Supp. Fig. 8)*	*defined using top-down task with the leave-one-run-out procedure (Supp. Fig. 9)*
V1	776.2±103.32	754±323.06	723.07±320.82
V2	723.6±59.41	804.6±218.24	772.41±219.11
V3	714±121.98	808.6±129.46	777.4±131.57
V3AB	608.6±250.11	629.2±175.08	609.32±173.6
V4	342.2±47.38	321.4±72.35	309.99±73.24
IPS	513.4±162.44	407.6±117.03	385.75±114.93

References

1. **Egeth HE**, **Yantis S**. VISUAL ATTENTION: Control, Representation, and Time Course. *Annu Rev Psychol* 48: 269–297, 1997. doi: 10.1146/annurev.psych.48.1.269.

2. **Giordano AM**, **McElree B**, **Carrasco M**. On the automaticity and flexibility of covert attention: A speed-accuracy trade-off analysis. *J Vis* 9: 30–30, 2009.

3. **Jonides J**. Further toward a model of the mind's eye's movement. *Bull Psychon Soc* 21: 247–250, 1983.

4. **Carrasco M**. Visual attention: the past 25 years. *Vision Res* 51: 1484–1525, 2011. doi: 10.1016/j.visres.2011.04.012.

5. **Serences JT**, **Kastner S**. A Multi-level Account of Selective Attention. .

6. **Ling S**, **Carrasco M**. Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision Res* 46: 1210–1220, 2006.

7. **Liu T**, **Abrams J**, **Carrasco M**. Voluntary attention enhances contrast appearance. *Psychol Sci* 20: 354–362, 2009.

8. **Abrams J**, **Barbot A**, **Carrasco M**. Voluntary attention increases perceived spatial frequency. .

9. **Yeshurun Y**, **Montagna B**, **Carrasco M**. On the flexibility of sustained attention and its effects on a texture segmentation task. *Vision Res* 48: 80–95, 2008. doi: 10.1016/j.visres.2007.10.015.

10. **Barbot A**, **Carrasco M**. Attention Modifies Spatial Resolution According to Task Demands. .

11. **Awh E**, **Matsukura M**, **Serences JT**. Top-down control over biased competition during covert spatial orienting. *J Exp Psychol Hum Percept Perform* 29: 52, 2003.

12. **Serences JT**, **Yantis S**, **Culberson A**, **Awh E**. Preparatory activity in visual cortex indexes distractor suppression during covert spatial orienting. *J Neurophysiol* 92: 3538–3545, 2004.

13. **Brefczynski JA**, **DeYoe EA**. A physiological correlate of the'spotlight'of visual attention. *Nat Neurosci* 2: 370–374, 1999.

14. **Cohen MR**, **Maunsell JHR**. Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* 12: 1594–1600, 2009. doi: 10.1038/nn.2439.

15. **Colby CL**. The neuroanatomy and neurophysiology of attention. *J Child Neurol* 6: S90–S118, 1991.

16. **Kastner S**, **Ungerleider L**. Mechanisms of Visual Attention in the Human Cortex. *Annu Rev Neurosci* 23: 315–341, 2000. doi: 10.1146/annurev.neuro.23.1.315.

17. **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* 77: 24–42, 1997. doi: 10.1152/jn.1997.77.1.24.

18. **Mitchell JF**, **Sundberg KA**, **Reynolds JH**. Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron* 55: 131–141, 2007. doi: 10.1016/j.neuron.2007.06.018.

19. **Moran J**, **Desimone R**. Selective attention gates visual processing in the extrastriate cortex. *Science* 229: 782–784, 1985.

20. **Reynolds JH**, **Pasternak T**, **Desimone R**. Attention Increases Sensitivity of V4 Neurons. *Neuron* 26: 703–714, 2000. doi: 10.1016/S0896-6273(00)81206-4.

21. **Sundberg KA**, **Mitchell JF**, **Reynolds JH**. Spatial Attention Modulates Center-Surround Interactions in Macaque Visual Area V4. *Neuron* 61: 952–963, 2009. doi: 10.1016/j.neuron.2009.02.023.

22. **Vo VA**, **Sprague TC**, **Serences JT**. Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. *J Neurosci* 37: 3386–3401, 2017. doi: 10.1523/JNEUROSCI.3484-16.2017.

23. **Desimone R**, **Duncan J**. Neural Mechanisms of Selective Visual Attention. *Annu Rev Neurosci* 18: 193–222, 1995. doi: 10.1146/annurev.ne.18.030195.001205.

24. **Reynolds JH**, **Chelazzi L**. Attentional Modulation of Visual Processing. *Annu Rev Neurosci* 27: 611–647, 2004. doi: 10.1146/annurev.neuro.26.041002.131039.

25. **Ganis G**, **Thompson WL**, **Kosslyn SM**. Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cogn Brain Res* 20: 226–241, 2004.

26. **Hopfinger JB**, **Buonocore MH**, **Mangun GR**. The neural mechanisms of top-down attentional control. *Nat Neurosci* 3: 284–291, 2000.

27. **Hopfinger JB**, **Woldorff MG**, **Fletcher EM**, **Mangun GR**. Dissociating top-down attentional control from selective perception and action. *Neuropsychologia* 39: 1277–1291, 2001.

28. **Kaas A**, **Weigelt S**, **Roebroeck A**, **Kohler A**, **Muckli L**. Imagery of a moving object: the role of occipital cortex and human MT/V5+. *Neuroimage* 49: 794–804, 2010.

29. **Kastner S**, **Pinsk MA**, **De Weerd P**, **Desimone R**, **Ungerleider LG**. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* 22: 751–761, 1999.

30. **Klein I**, **Paradis A-L**, **Poline J-B**, **Kosslyn SM**, **Le Bihan D**. Transient activity in the human calcarine cortex during visual-mental imagery: an event-related fMRI study. *J Cogn Neurosci* 12: 15–23, 2000.

31. **Knauff M**, **Kassubek J**, **Mulack T**, **Greenlee MW**. Cortical activation evoked by visual mental imagery as measured by fMRI. *Neuroreport* 11: 3957–3962, 2000.

32. **Le Bihan D**, **Turner R**, **Zeffiro TA**, **Cuenod CA**, **Jezzard P**, **Bonnerot V**. Activation of human primary visual cortex during visual recall: a magnetic resonance imaging study. *Proc Natl Acad Sci* 90: 11802–11805, 1993.

33. **O'Craven KM**, **Kanwisher N**. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J Cogn Neurosci* 12: 1013–1023, 2000.

34. **Slotnick SD**, **Thompson WL**, **Kosslyn SM**. Visual mental imagery induces retinotopically organized activation of early visual areas. *Cereb Cortex* 15: 1570–1583, 2005.

35. **Albers AM**, **Kok P**, **Toni I**, **Dijkerman HC**, **de Lange FP**. Shared Representations for Working Memory and Mental Imagery in Early Visual Cortex. *Curr Biol* 23: 1427–1431, 2013. doi: 10.1016/j.cub.2013.05.065.

36. **Horikawa T**, **Kamitani Y**. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun* 8: 15037, 2017. doi: 10.1038/ncomms15037.

37. **Kamitani Y**, **Tong F**. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8: 679–685, 2005.

38. **Koenig-Robert R**, **Pearson J**. Decoding the contents and strength of imagery before volitional engagement. *Sci Rep* 9: 3504, 2019. doi: 10.1038/s41598-019-39813-y.

39. **Kok P**, **Failing MF**, **de Lange FP**. Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. *J Cogn Neurosci* 26: 1546–1554, 2014. doi: 10.1162/jocn_a_00562.

40. **Kok P**, **Mostert P**, **de Lange FP**. Prior expectations induce prestimulus sensory templates. *Proc Natl Acad Sci* 114: 10473–10478, 2017. doi: 10.1073/pnas.1705652114.

41. **Lee S-H**, **Kravitz DJ**, **Baker CI**. Disentangling visual imagery and perception of real-world objects. *Neuroimage* 59: 4064–4073, 2012.

42. **Naselaris T**, **Olman CA**, **Stansbury DE**, **Ugurbil K**, **Gallant JL**. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage* 105: 215–228, 2015. doi: 10.1016/j.neuroimage.2014.10.018.

43. **Thirion B**, **Duchesnay E**, **Hubbard E**, **Dubois J**, **Poline J-B**, **Lebihan D**, **Dehaene S**. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage* 33: 1104–1116, 2006. doi: 10.1016/j.neuroimage.2006.06.062.

44. **Vetter P**, **Smith FW**, **Muckli L**. Decoding Sound and Imagery Content in Early Visual Cortex. *Curr Biol* 24: 1256–1262, 2014. doi: 10.1016/j.cub.2014.04.020.

45. **Hubel DH**, **Wiesel TN**. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *J Comp Neurol* 158: 295–305, 1974.

46. **Hubel DH**, **Wiesel TN**. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195: 215–243, 1968. doi: 10.1113/jphysiol.1968.sp008455.

47. **Livingstone M**, **Hubel D**. Segregation of Form, Color, Movement, and Depth: Anatomy, Physiology, and Perception. *Sci New Ser* 240: 740–749, 1988.

48. **Victor JD**, **Purpura K**, **Katz E**, **Mao B**. Population encoding of spatial frequency, orientation, and color in macaque V1. .

49. **Desimone R**, **Schein SJ**. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *J Neurophysiol* 57: 835–868, 1987.

50. **Nandy AS**, **Mitchell JF**, **Jadi MP**, **Reynolds JH**. Neurons in macaque area V4 are tuned for complex spatio-temporal patterns. *Neuron* 91: 920–930, 2016.

51. **Pasupathy A**, **Connor CE**. Responses to contour features in macaque area V4. *J Neurophysiol* 82: 2490–2502, 1999.

52. **Roe AW**, **Chelazzi L**, **Connor CE**, **Conway BR**, **Fujita I**, **Gallant JL**, **Lu H**, **Vanduffel W**. Toward a Unified Theory of Visual Area V4. *Neuron* 74: 12–29, 2012. doi: 10.1016/j.neuron.2012.03.011.

53. **Van Essen DC**, **Gallant JL**. Neural mechanisms of form and motion processing in the primate visual system. *Neuron* 13: 1–10, 1994.

54. **Wilson HR**, **Wilkinson F**. From orientations to objects: Configural processing in the ventral stream. *J Vis* 15: 4–4, 2015.

55. **Zeki SM**. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J Physiol* 236: 549–573, 1974.

56. **Blatt GJ**, **Andersen RA**, **Stoner GR**. Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. *J Comp Neurol* 299: 421–445, 1990. doi: 10.1002/cne.902990404.

57. **Cavanaugh J**, **Joiner WM**, **Wurtz RH**. Suppressive surrounds of receptive fields in monkey frontal eye field. *J Neurosci* 32: 12284–12293, 2012.

58. **Mayo JP**, **DiTomasso AR**, **Sommer MA**, **Smith MA**. Dynamics of visual receptive fields in the macaque frontal eye field. *J Neurophysiol* 114: 3201–3210, 2015. doi: 10.1152/jn.00746.2015.

59. **Mohler CW**, **Goldberg ME**, **Wurtz RH**. Visual receptive fields of frontal eye field neurons. .

60. **Anderson JC**, **Kennedy H**, **Martin KAC**. Pathways of Attention: Synaptic Relationships of Frontal Eye Field to V4, Lateral Intraparietal Cortex, and Area 46 in Macaque Monkey. *J Neurosci* 31: 10872–10881, 2011. doi: 10.1523/JNEUROSCI.0622-11.2011.

61. **Barone P**, **Batardiere A**, **Knoblauch K**, **Kennedy H**. Laminar Distribution of Neurons in Extrastriate Areas Projecting to Visual Areas V1 and V4 Correlates with the Hierarchical Rank and Indicates the Operation of a Distance Rule. *J Neurosci* 20: 3263–3281, 2000. doi: 10.1523/JNEUROSCI.20-09-03263.2000.

62. **Schall J**, **Morel A**, **King D**, **Bullier J**. Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J Neurosci* 15: 4464–4487, 1995. doi: 10.1523/JNEUROSCI.15-06-04464.1995.

63. **Stanton GB**, **Bruce CJ**, **Goldberg ME**. Topography of projections to posterior cortical areas from the macaque frontal eye fields. *J Comp Neurol* 353: 291–305, 1995. doi: 10.1002/cne.903530210.

64. **Moore T**, **Armstrong KM**. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373, 2003. doi: 10.1038/nature01341.

65. **Buffalo EA**, **Fries P**, **Landman R**, **Liang H**, **Desimone R**. A backward progression of attentional effects in the ventral stream. *Proc Natl Acad Sci* 107: 361–365, 2010. doi: 10.1073/pnas.0907658106.

66. **Brainard DH**. The Psychophysics Toolbox. *Spat Vis* 10: 433–436, 1997. doi: 10.1163/156856897X00357.

67. **Kleiner M**, **Brainard D**, **Pelli D**, **Ingling A**, **Murray R**, **Broussard C**. What is new in psychtoolbox 3. *Perception* 36: 1–16, 2007.

68. **Stigliani A**, **Jeska B**, **Grill-Spector K**. Encoding model of temporal processing in human visual cortex. In: *Proceedings of the National Academy of Sciences*.

69. **Andersson JL**, **Skare S**, **Ashburner J**. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20: 870–888, 2003.

70. **Jenkinson M**, **Beckmann CF**, **Behrens TE**, **Woolrich MW**, **Smith SM**. Fsl. *Neuroimage* 62: 782–790, 2012.

71. **Dale AM**, **Fischl B**, **Sereno MI**. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* 9: 179–194, 1999.

72. **Greve DN**, **Fischl B**. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 48: 63–72, 2009.

73. **Jenkinson M**, **Bannister P**, **Brady M**, **Smith S**. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17: 825–841, 2002.

74. **Jenkinson M**, **Smith S**. A global optimisation method for robust affine registration of brain images. *Med Image Anal* 5: 143–156, 2001.

75. **Smith SM**. Fast robust automated brain extraction. *Hum Brain Mapp* 17: 143–155, 2002.

76. **Woolrich MW**, **Ripley BD**, **Brady M**, **Smith SM**. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* 14: 1370–1386, 2001.

77. **Gandhi SP**, **Heeger DJ**, **Boynton GM**. Spatial attention affects brain activity in human primary visual cortex. *Proc Natl Acad Sci* 96: 3314–3319, 1999.

78. **Silver MA**, **Ress D**, **Heeger DJ**. Topographic maps of visual spatial attention in human parietal cortex. *J Neurophysiol* 94: 1358–1371, 2005.

79. **Sprague TC**, **Serences JT**. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat Neurosci* 16: 1879–1887, 2013.

80. **Engel SA**, **Glover GH**, **Wandell BA**. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb Cortex N Y NY 1991* 7: 181–192, 1997.

81. **Jerde TA**, **Curtis CE**. Maps of space in human frontoparietal cortex. *J Physiol-Paris* 107: 510–516, 2013.

82. **Sereno MI**, **Dale A**, **Reppas J**, **Kwong K**, **Belliveau J**, **Brady T**, **Rosen B**, **Tootell R**. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268: 889–893, 1995.

83. **Swisher JD**, **Halko MA**, **Merabet LB**, **McMains SA**, **Somers DC**. Visual topography of human intraparietal sulcus. *J Neurosci* 27: 5326–5337, 2007.

84. **Wandell BA**, **Dumoulin SO**, **Brewer AA**. Visual field maps in human cortex. *Neuron* 56: 366–383, 2007.

85. **Winawer J**, **Witthoft N**. Human V4 and ventral occipital retinotopic maps. *Vis Neurosci* 32: E020, 2015.

86. **Sheremata SL**, **Silver MA**. Hemisphere-Dependent Attentional Modulation of Human Parietal Visual Field Representations. *J Neurosci* 35: 508–517, 2015. doi: 10.1523/JNEUROSCI.2378-14.2015.

87. **Felleman DJ**, **Van Essen DC**. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb Cortex* 1: 1–47, 1991. doi: 10.1093/cercor/1.1.1.

88. **Bosking WH**, **Crowley JC**, **Fitzpatrick D**. Spatial coding of position and orientation in primary visual cortex. *Nat Neurosci* 5: 874–882, 2002. doi: 10.1038/nn908.

89. **Prinzmetal W**. Location perception: TheX-Files parable. *Percept Psychophys* 67: 48–71, 2005. doi: 10.3758/BF03195012.

90. **Seriès P**, **Latham PE**, **Pouget A**. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat Neurosci* 7: 1129–1135, 2004. doi: 10.1038/nn1321.

91. **Shapley R**, **Victor J**. Hyperacuity in cat retinal ganglion cells. *Science* 231: 999–1002, 1986.

92. **Westheimer G**. Visual hyperacuity. In: *Progress in sensory physiology*. Springer, 1981, p. 1–30.

93. **Westheimer G**, **McKee SP**. Integration regions for visual hyperacuity. *Vision Res* 17: 89–93, 1977.

94. **Harrison SA**, **Tong F**. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458: 632–635, 2009.

95. **Serences JT**, **Ester EF**, **Vogel EK**, **Awh E**. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychol Sci* 20: 207–214, 2009. doi: 10.1111/j.1467-9280.2009.02276.x.

96. **Sprague TC**, **Ester EF**, **Serences JT**. Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Curr Biol* 24: 2174–2180, 2014. doi: 10.1016/j.cub.2014.07.066.

97. **Sprague TC**, **Ester EF**, **Serences JT**. Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron* 91: 694–707, 2016. doi: 10.1016/j.neuron.2016.07.006.

98. **Breedlove JL**, **St-Yves G**, **Olman CA**, **Naselaris T**. Generative feedback explains distinct brain activity codes for seen and mental images. *Curr Biol* 30: 2211–2224, 2020.

99. **Favila SE**, **Kuhl BA**, **Winawer J**. Perception and memory have distinct spatial tuning properties in human visual cortex. *bioRxiv* 811331, 2020.

Chapter 2 Consequences of relaying top-down attentional modulations via neurons with high-dimensional selectivity

Abstract

Visual search is often guided by top-down knowledge of the spatial location or features that define a behaviorally relevant object in the environment. In turn, sensory neurons in early visual cortex undergo highly selective gain modulations that reflect the location or the identity of the attended feature. These top-down gain modulations are thought to be mediated via feedback from areas of prefrontal cortex. However, neurons in many areas of prefrontal cortex (PFC) have high-dimensional tuning for multiple features (i.e. mixed-selectivity). This poses a fundamental challenge: how can PFC neurons that are simultaneously tuned to many features propagate highly selective top-down feedback to early sensory neurons that are tuned to specific relevant features? To test this, we used a spiking neural network consisting of two layers of neurons that have reciprocal, random projections between the layers. The first layer consisted of eight ring-like sub-networks, in which neurons were topographically arranged by stimulus selectivity in a circular feature space. The second layer consisted of neurons that were randomly and reciprocally connected to multiple first layer neurons, which gave rise to linear mixed selectivity to multiple features in the second layer neurons. As a result, we could simulate top-down modulations by applying gain to the second layer, which excite/inhibit first layer neurons across many sub-networks. In spatial and feature-based attention simulations, we found that the top-down feedback signals spread to sub-networks that did not receive any stimulus input and modulated their activities in a consistent, but idiosyncratic way that does not resemble stimulus-related responses. This demonstrates the importance of random projections in canceling out spurious signals, preserving the fidelity of relevant top-down feedback without generating illusory representations. Our findings highlight previously unrecognized consequences of

relaying top-down feedback via neurons with high-dimensional tuning functions and how selective modulation can be achieved in a hierarchical neural model.

Introduction

Efficient visual search requires selectively prioritizing some stimulus inputs over others based on top-down knowledge about the spatial location and features that define currently relevant stimuli (1, 2). Although it is well-established that behavioral goals can modulate the gain and covariance structure of sensory codes, fundamental questions remain about how such top-down modulations back-propagate from higher-order areas of prefrontal and parietal cortex to influence perception (3, 4). For instance, basic visual features such as colors or orientations are encoded by neurons in the striate and extrastriate visual cortex (EVC) that are tuned to specific feature values and that have relatively small and precise spatial receptive fields (RFs; Refs. 5–9). In contrast, neurons in prefrontal cortex (PFC) have much larger spatial RFs and encode more abstract information about behavioral goals (10–16). Moreover, neurons in PFC also exhibit flexible and high-dimensional tuning functions that can encode multiple feature values (and/or combinations of feature values, termed *mixed selectivity*; Refs. 17, 18). Nevertheless, both causal and correlational studies suggest that PFC neurons with high-dimensional selectivity can generate highly selective modulations that target specific sub-populations of neurons in EVC to enhance the response evoked by relevant features (19–24). Here we investigate how PFC neurons with high-dimensional tuning can generate precise top-down modulatory signals to facilitate efficient information processing in the visual cortex.

Classic studies in visual information processing demonstrate that EVC neurons are highly selective to specific spatial locations and visual features. For example, invasive recordings in non-human animal models reveal that neurons in the striate cortex and EVC exhibit small, highly selective spatial RFs that are retinotopically organized (7, 9, 25). Other studies have found that these neurons show increased firing rate for specific orientations, colors, motion directions and

other sets of features (7, 8, 26–32). Moreover, anatomical studies reveal a functional-anatomical alignment such that areas of visual cortex are organized according to circular feature dimensions such as orientation of motion direction (5, 6, 8, 33). Thus, visual features in different locations are encoded by different subsets of feature-selective neurons in EVC, and responses in these neurons are thought to collectively form the basis for generating increasingly complex representations of real-world objects and where they are in the visual scene (34–39).

When a visual feature within the scene is behaviorally relevant, top-down attentional modulations can selectively increase the gain of neurons that are selective for the spatial positions and feature values that align with the attentional target to enhance the overall speed and accuracy of information processing (Refs. 23, 24, 40–54; reviewed in Refs. 2, 55). For example, when top-down attention is directed to a spatial location, the firing rate of neurons in EVC that have corresponding spatial RFs increases compared to when the attention is directed outside of their spatial RFs, even in the absence of an external stimulus (44, 52, 56). Similarly, searching for a specific feature, such as a particular direction of motion, will increase the firing rate of neurons in MT that are selective for that feature while simultaneously decreasing the firing rate of neurons tuned to other features (23, 42, 57).

These highly specific top-down modulations of space- and feature-selective responses in EVC are thought to be generated in regions of frontal and parietal cortex, such as the lateral intraparietal area (LIP), the frontal eye fields (FEFs), and lateral prefrontal cortex (LPFC; Refs. 1–4, 14, 58–64). For example, microstimulation of neurons in the FEF leads to a spatially selective increase in the firing rates in V4 neurons, mimicking the effects of spatial attention (19). In addition, while FEF neurons are not inherently selective for stimulus features such as color or orientation (13, 65, 66), they have been found to flexibly encode whatever feature is

currently behaviorally relevant (66–69). More recently, the ventral prearcuate (VPA) region of the prefrontal cortex in nonhuman primates was suggested to compute feature-selective signals that are relayed to FEF to guide the spatial selection of relevant features, as optogenetic inactivation of this area reduced feature-based selection effects in FEF and V4 (70, 71). Similarly, neurons in LPFC are thought to integrate sensory input and internal states (task rules, behavioral goals) to flexibly control selective attention as a function of current behavioral demands (14, 72–74). Lesion studies have long implicated LPFC in cognitive control, as damage to this general area leads to deficits in task switching and the use of complex rules to guide behavior (75–82). Temporarily disrupting activity in LPFC with transcranial magnetic stimulation also impairs behavioral performance in tasks that require top-down control over the response to the bottom-up stimulus (83, 84), and similar findings were reported for optogenetic inactivation of posterior LPFC (85). Human imaging studies further suggest many prefrontal areas appear to be involved - to varying degrees - in both spatial and feature-based attention (86–89). For example, whereas FEF is traditionally thought to play an important role in spatial attention, inferior frontal junction (IFJ) has been implicated in feature-based attentional modulation, which is analogous to the function of VPA in non-human primates (90–95).

It is important to note that these prefrontal areas that are suggested to be the source of top-down modulations in earlier visual areas have relatively low spatial selectivity compared to the target earlier visual areas. For example, while striate cortex and EVC neurons have quite small spatial RFs (e.g. in V1: smaller than 1° at fovea and ∼1° at 10° periphery; Ref. 9), parietal and frontal cortex neurons have much larger spatial RFs (e.g. in FEF: ∼10–20° with a large variance at 10° periphery; Refs. 10–13). That is, PFC neurons that are inherently less selective nevertheless relay feedback signals in a highly selective manner to EVC neurons.

Moreover, PFC neurons are found to have high-dimensional selectivity to different stimulus features. Some neurons exhibit linear mixed selectivity, characterized by responses to multiple features that linearly combine to predict the final output of the neuron. Other neurons exhibit non-linear mixed selectivity, where sensitivity to different stimulus features varies as a function of task context (17, 18). Previous empirical and theoretical studies suggest that high-dimensional tuning supports flexible representations of different external stimuli or task demands, and increases the encoding capacity of population responses (17, 18, 96). That said, these complex, high-dimensional responses stand in contrast to the highly selective, lower-dimensional tuning properties typically observed in EVC neurons.

This divergence in receptive field complexity between PFC and EVC poses a challenge for understanding top-down attentional control for visual features: How can neurons with high-dimensional tuning in PFC evoke highly selective attentional modulations that enhance the signal for a single relevant feature for a specific spatial location in EVC? In the domain of spatial attention, this problem might be partially overcome, as many neurons in PFC have topographical organization and thus there might be some consistency across the visual processing hierarchy that supports mapping feedback signals to specific subsets of EVC neurons (97). This broad feedback gain can then interact with the spatially specific bottom-up stimulus input to exhibit a spatially localized attentional modulation. In contrast, the backpropagation of feature-selective feedback signals is seemingly more complicated. For example, feedback signals emanating from a neuron with high-dimensional tuning, such as a PFC neuron sensitive to red and vertical, may cause spurious modulations of task irrelevant features if only the color red happens to be behaviorally relevant. Such unintended modulations may in turn interfere with the specificity of

top-down modulations of the target feature or even cause perceptual interference in the form of illusory percepts.

One way this correspondence problem could be solved is if the neurons that control top-down modulations collapse their response profiles from high- to low-dimensional before back-propagating modulatory signals to EVC. However, this solution is circular as the neurons generating the control signals would themselves need an additional control signal about how to modulate their activity. Here, we propose an alternative hypothesis that randomness in the feedforward and feedback connections between EVC and higher-order regions can cancel out spurious signals that might be propagated as the result of the high-dimensional selectivity in the higher-order areas, leaving intact only the desired modulatory signal targeting the relevant feature. These random projections could be initialized from matching random, unique labels that are tagged by cell adhesion molecules such as protocadherin (98, 99). To test this hypothesis, we used a spiking neural network model to simulate a neural architecture in which the connections between lower-level 'sensory' neurons and higher-order 'control' neurons were random. We first investigate the efficacy of top-down modulatory signals in terms of targeting a relevant spatial location and/or a feature using spatial and feature attention tasks. Our model reproduced the contrast gain effects associated with top-down attention, aligning with previous experimental results (2, 100). Notably, feature-based attentional gain caused modulations in unstimulated sub-networks that did not resemble coherent sensory responses, highlighting how neurons with high-dimensional tuning can propagate idiosyncratic signals throughout the network. Our findings suggest that random projections in a hierarchical network can naturally enhance representations of relevant features while simultaneously cancelling out irrelevant signals.

Results

We adapted a spiking neural network model that was previously used to model storage capacity and patterns of errors in visual working memory (101). The model consists of two layers of Poisson spiking neurons where the first layer has eight sub-networks with neurons that are selective for different feature values in a circular stimulus space and the second layer receives converging input from neurons in all eight sub-networks (Figure 2.1). Each neuron in a sub-network has short-range excitatory connections and long-range inhibitory connections to other neurons within the sub-network, which gives rise to a response profile that follows a circular normal distribution across the feature dimension. In the second layer, each neuron is randomly connected via reciprocal feedforward and feedback connections to neurons in different first layer sub-networks, with overall excitatory and inhibitory connections balanced to prevent runaway excitation in the network. This architecture gives rise to two important properties of the model: 1) neurons in the second layer can have high-dimensional tuning for multiple features through converging inputs from first layer neurons with different tuning preferences, and 2) even though the first layer sub-networks are not directly connected to each other, neurons in different sub-networks indirectly interact via overlapping feedforward and feedback connections with neurons in the second layer. This latter point is key to testing our central hypothesis, whether randomness in between-layer connections contribute to targeted feedback modulations, as the indirect interactions between neurons in different sub-networks can give rise to spurious sensory representations with the application of top-down signals that back-propagate from the second layer.

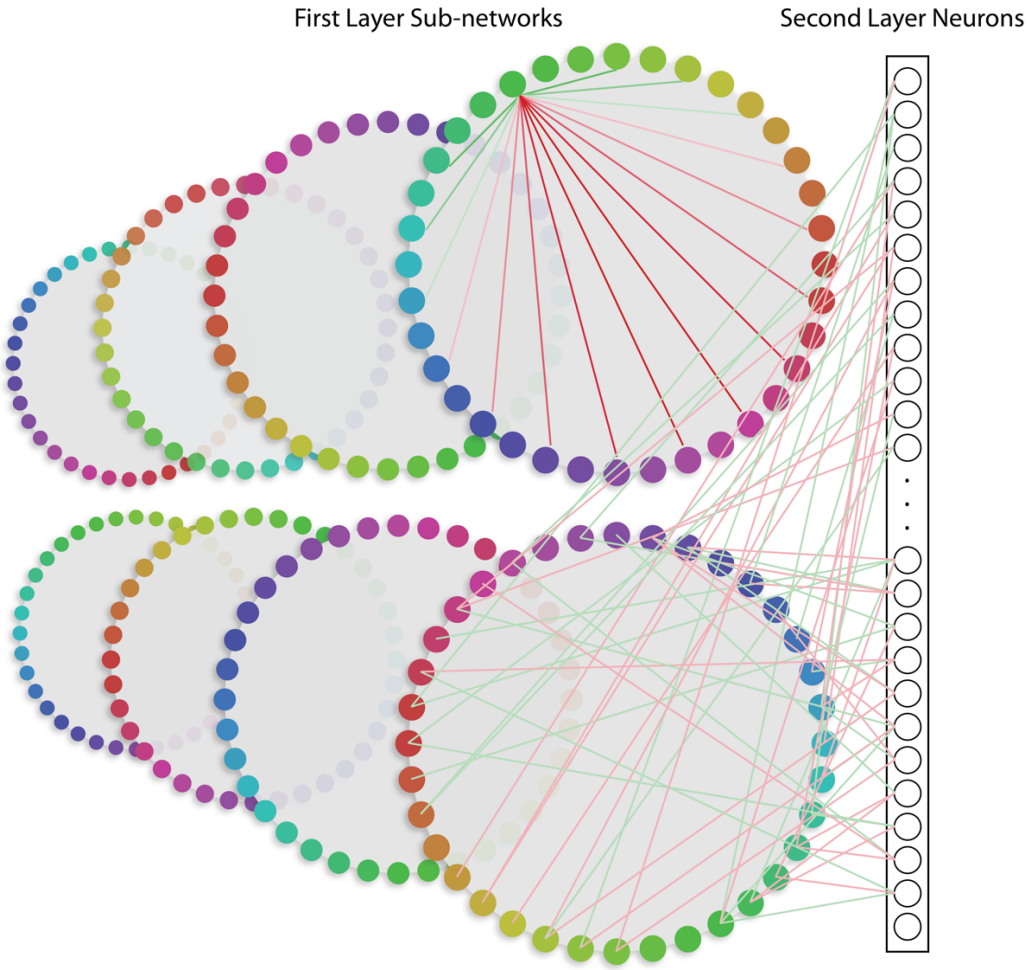


Figure 2.1. Spiking network model overview. The model comprised of two layers of Poisson spiking neurons. The first layer contained 8 ring-like "sensory" sub-networks consisting of 512 neurons, simulating sensory networks that encode information in circular stimulus spaces. Every neuron in each sub-network had a preferred stimulus input, due to short-range excitatory connections (green lines within the top-right sub-network) and long-range inhibitory connections (red lines within the top-right sub-network) between each neuron and its neighbors around the ring. We conceptualized each sub-network to encode independent sensory inputs such as orientations presented in different spatial positions around the visual field. The second layer is composed of 1024 neurons that are randomly and reciprocally connected to the neurons in the first layer (red and green lines connecting to the bottom-right sub-network), with the connections balanced between excitatory (green lines) and inhibitory (red lines). While the first layer sub-networks were not directly connected to each other, they can indirectly interact via feed-forward and feedback projections to the second layer. Stimulus inputs were a circular normal distribution centered on an angle between 1° and 360°, which we defined as the circular stimulus space that the sub-networks encode.

Applying spatially specific gain to the second layer shifts response functions in first layer neurons

After setting up the network, we tested to see if the network is capable of showing stereotyped effects of spatial attention as reported in prior empirical work. To simulate top-down spatial attention, we picked top 20% of the second layer neurons that have the highest weights to each sub-network based on the rank-order of the feedback connection weights (i.e. connections from the second layer neurons to the first layer neurons) and then applied varying levels of gain to those. This approach was based on the premise that each of the first layer sub-networks represents a group of neurons encoding a visual feature presented in a different spatial position in the visual field. Then, we presented a single stimulus to one of the sub-networks and measured the average firing rates of the neurons with the highest selectivity for the presented feature (Figure 2.1A). We averaged these firing rates across trials when attentional gain was targeted to the stimulated sub-network, in which case the stimulus location matches the attended location (Attended condition), and when attentional gain was targeted on a different, unstimulated sub-network, in which case the stimulus location did not match the attended location (Unattended condition).

Figure 2.2B shows the average firing rates as a function of stimulus strength in order to visualize attention-related shifts in the contrast response function (CRF; Figure 2.2B, 2.2D; Refs. 2, 56, 100, 102, 103). We found that as stronger attention gain was applied, the response function in the Attended condition shifted leftward, showing a stereotypical contrast gain effect (Figure 2.2B). As expected, there were no systematic shifts in the Unattended condition (Figure 2.2D). To quantify the attention-related shift, we fit a Naka-Rushton equation to each response

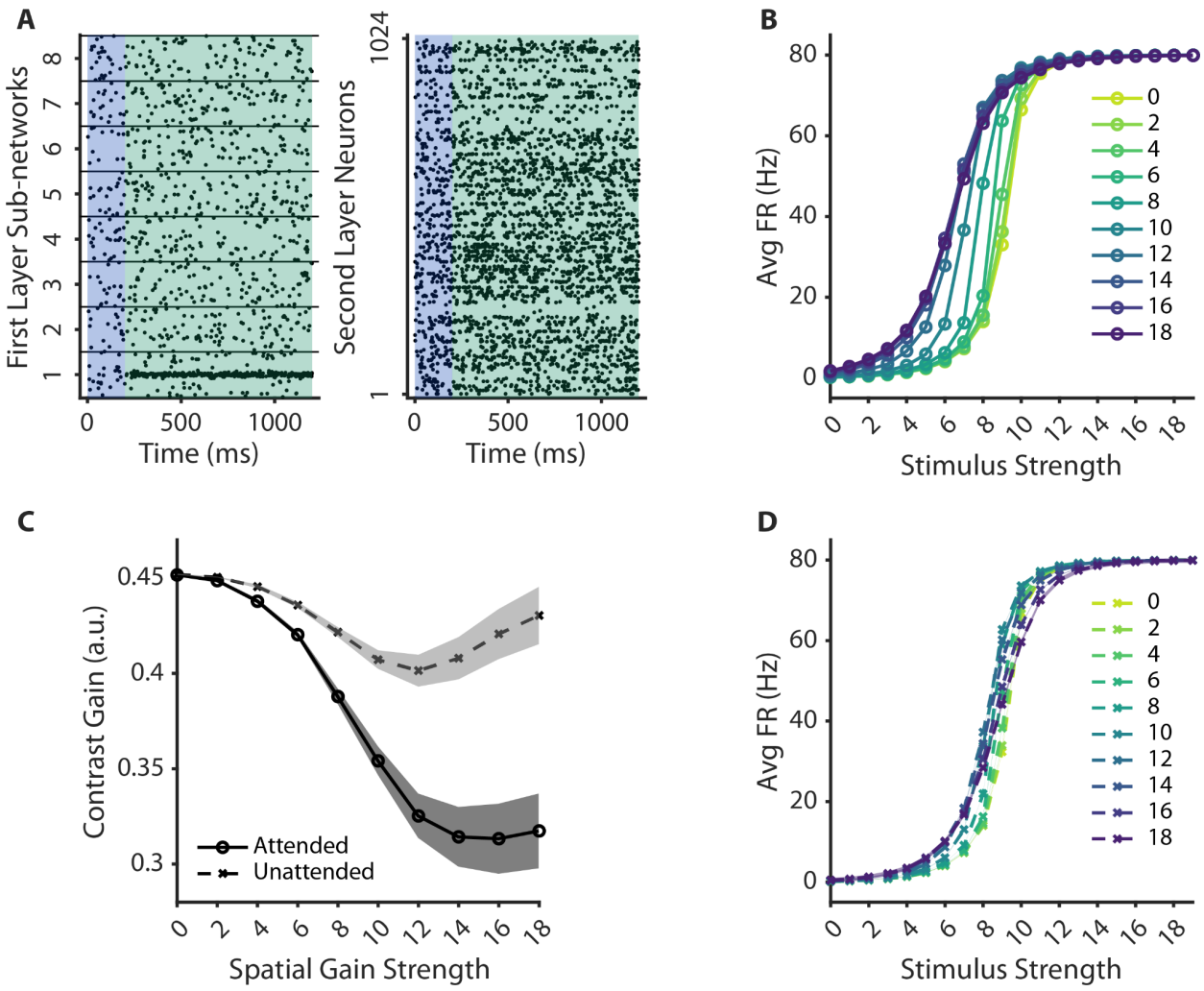


Figure 2.2 Spatial attention simulations. A. An example trial from the *attention* task in which sub-network 1 was attended. Left panel shows spikes for the first layer sub-networks and right panel shows spikes for the second layer neurons. Spatial attention gain was applied to a subset of second layer neurons that have the highest connectivity weights to sub-network 1 in the first layer for 200ms (shaded blue; gain strength: 8). Then, a single stimulus input at 180° was presented to sub-network 1 for 1000ms (shaded green; stimulus strength: 7). B, D. Contrast response functions (CRFs) for when the stimulus was presented in the attended sub-network (B) and in the unattended sub-network (D). Each line represents a different level of spatial attention gain strength. C. Estimated contrast gain parameters from Attended (solid black line) and Unattended (dotted gray line) conditions. Shaded areas in B, C, D represent standard error of mean across network initializations.

function to estimate the contrast gain parameter. The estimated contrast gain, for which a lower value indicates larger leftward shift in the response function, decreased with stronger attention

gain in the Attended condition, while the change was minimal in the Unattended condition (Figure 2.2C). This demonstrates that our model is able to produce enhanced stimulus representations with spatial attention.

Applying feature specific gain to the second layer enhances responses to relevant over irrelevant stimuli in first layer neurons

Next, we evaluated the effect of feature-based attention on stimulus representations in the first layer of our model. We simultaneously presented two stimuli to one sub-network and then applied varying levels of gain to a subset of second layer neurons to simulate top-down feature-based attention to one of the two presented stimuli (Figure 2.3B). We then examined the stimulus-evoked response of second layer neurons when each stimulus was presented individually to that sub-network. Based on these evoked responses, we identified 20% of the second layer neurons that responded maximally to each stimulus. Depending on which of the two presented stimuli was the attentional target, the corresponding subset of second layer neurons received attentional gain that then back-propagated to the sensory relevant sensory neurons in the first layer. While this method of determining which neurons in the second layer to target with top-down gain required some knowledge about network connectivity, it was inspired by empirical demonstrations that global neuromodulators like norepinephrine interact with local 'hot-spots' of stimulus-driven activity to amplify neural gain and feedback signals to early sensory areas (termed the *GANE* model; Ref. 104).

We next measured the average firing rates of neurons that maximally prefer the attended stimulus (Attended condition) or the unattended stimulus (Unattended condition) for every trial from the stimulated sub-network and plotted them as a function of stimulus strength (Figure

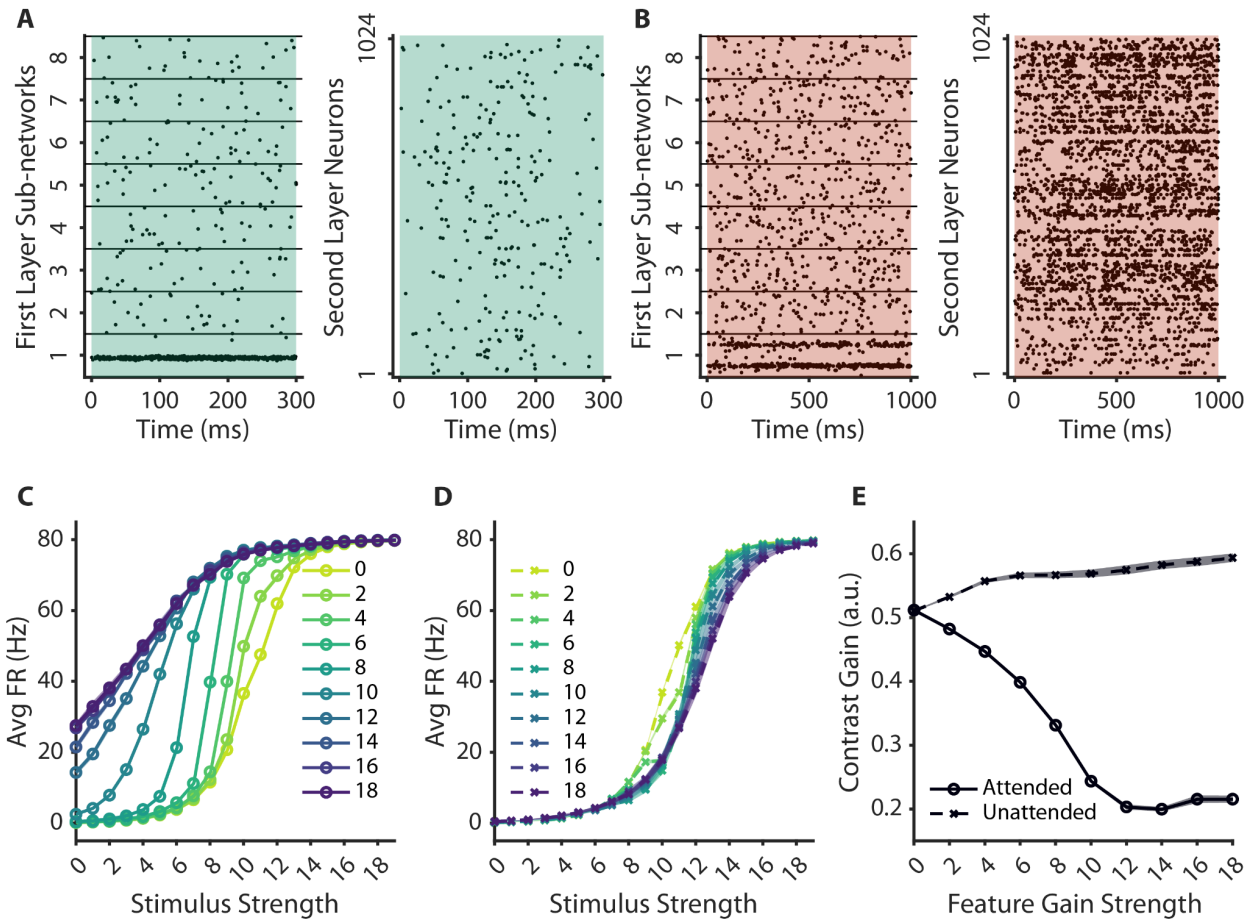


Figure 2.3 Feature-based attention simulations. A. An example trial from the *sensory* task in which sub-network 1 was presented with a stimulus input at 157.5° for 300ms. B. An example trial from the *attention* task. Two stimulus inputs, at 90° and 270°, were presented to sub-network 1 and at the same time (stimulus strength: 6), feature-based attention gain was applied to a subset of second layer neurons that have the highest selectivity to the 90° stimulus in sub-network 1 (lower cluster of spikes; gain strength: 8) for 1000ms. C, D. CRFs for when the stimulus was attended (C) and unattended (D). Each line represents a different level of feature-based attention gain strength. E. Estimated contrast gain parameters from Attended (solid black line) and Unattended (dotted gray line) conditions. Shaded areas in C, D, E represent standard error of mean across network initializations.

2.3C, 2.3D). As with spatial attention, we found a contrast gain modulation such that response functions were systematically shifted leftward with increasing feature-based attentional gain. For

the unattended stimulus, the response function did not show the same pattern and instead shifted rightward with stronger top-down gain applied to the other stimulus. This pattern gave rise to diverging contrast gain parameters that decreased with higher feature attention gain for the attended stimulus and slightly increased for the unattended stimulus (Figure 2.3E). These results demonstrate that our network can capture feature-based attention effects that bias representations of relevant features over representations of irrelevant features that are simultaneously present.

Having established basic space- and feature-based attentional modulations in the network, we next investigated the effects of backpropagated top-down signals in the unstimulated sub-networks to better understand the consequences of propagating gain signals through second layer neurons with high-dimensional tuning and connections with many neurons across sub-networks in the first layer. To this end, we presented the network with an additional *sensory* task, in which a single stimulus was presented to each first layer sub-network to drive feature-specific activity patterns across neurons in each of the eight sub-networks (Figure 2.3A). We then used these stimulus-driven activity patterns in each sub-network to train decoding models to predict the attended feature in the stimulated and unstimulated sub-networks. This approach allowed us to confirm that attentional gain amplified stimulus-specific representations of the relevant feature in the stimulated sub-network, and also if stimulus-specific representations incidentally arose in other unstimulated sub-networks which might be predicted based on studies of global feature-based attention effects (23, 105, 106).

We first ensured that we could decode the identity of a single feature presented to a sub-network in the *sensory* task. We trained a circular ridge regression model on the activity patterns across the neurons separately for each sub-network and quantified the decoding accuracy as the mean absolute error (MAE) between the presented and the predicted feature value. The average

MAE was 0.34°, 95% CI [0.33, 0.34], indicating that the network supports highly accurate and stable sensory-driven representations.

We then cross-generalized the same regression model that was trained using data from the *sensory* task to predict stimulus identity in the *attention* task. We observed that when no top-down gain was applied to the second layer, the regression model was equally likely to predict that either of the two presented stimuli was behaviorally relevant and attended (Figure 2.4A, top-left panel). However, as we increased the strength of top-down gain, the model was increasingly accurate at predicting which of the two stimuli was the attentional target (Figure 2.4A), and this increase in prediction accuracy was reflected in systematically decreasing MAEs (Figure 2.4B, solid black line). This change was not observed in the unstimulated sub-networks (Figure 2.4B, dotted gray line), confirming that attentional gain of stimulus-driven signals is largely contained within the sub-network that is presented with bottom-up stimuli and that attention biases representations in a way that resembles enhanced response to the bottom-up sensory input.

As a secondary point, we note that the stimulus strength was fixed at 10 for the *sensory* task but varied from 0 to 18 in the *attention* task. Thus, we next computed MAE separately for each stimulus strength level in the *attention* task (Figure 2.5). For stimulus strengths 0 to 10, prediction accuracy increased as a function of feature-based attention gain and with higher stimulus strength, as can be seen in decreasing MAE. However, when the stimulus strength was higher than 10, we observed that MAE decreased with higher feature-based attentional gain, eventually rising above chance level of 90°. This lower-than-chance prediction accuracy indicates that the regression model is systematically guessing the value of the unattended stimulus, and occurs because when a strong stimulus signal is combined with even more gain, it

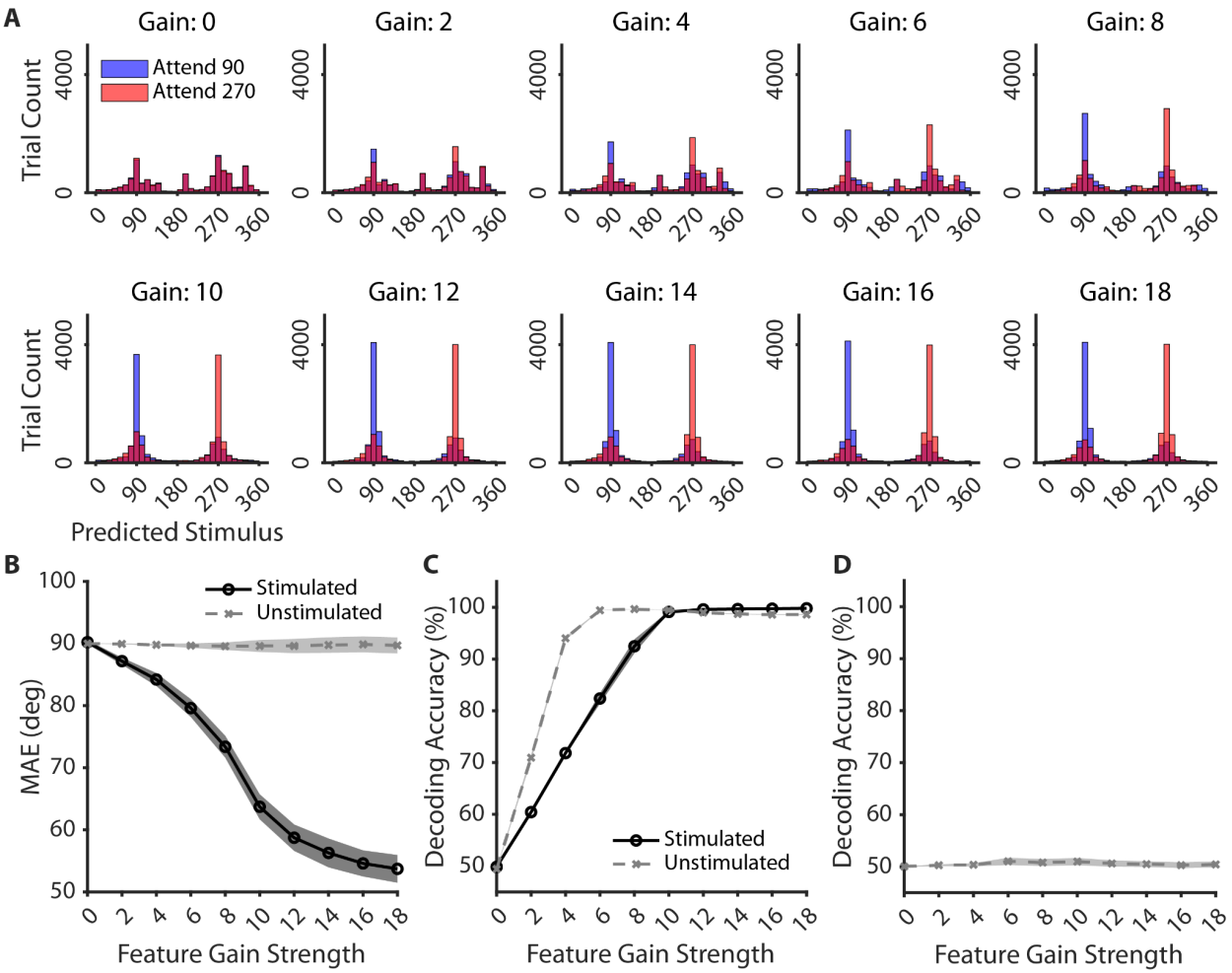


Figure 2.4 Decoding results from feature-based attention simulations. A. Histograms of predicted identity of the attended stimulus for every trial in the *attention* task for the stimulated sub-network, from a circular ridge regression model trained on the *sensory* task, collapsed across all network initializations. Each histogram represents predictions from different levels of feature-based attention gain strength, as labeled above each plot. B. Average MAE between the predicted and actual stimulus input for Stimulated (solid black line) and Unstimulated (dotted gray line) sub-networks based on regression model predictions. C. Decoding accuracy of support vector machines (SVMs) trained and tested on the *attention* task for Stimulated (solid black line) and Unstimulated (dotted gray line) sub-networks. D. Decoding accuracy of support vector machines (SVMs) based on the *attention* task, training on one unstimulated sub-network and testing on another unstimulated sub-network. Shaded areas in B, C, D represent standard error of mean across network initializations.

can diverge substantially from the sensory signal that the regression model is trained on (as shown in the raster plots in Supplementary Figure 2.1). This divergence from the sensory-evoked

response at high stimulus strengths thus led the model to predict that the unattended stimulus was more similar to the sensory signals that were used to train the model.

Propagated top-down feedback signals in the unstimulated sub-networks are consistent, but do not resemble stimulus-driven response patterns

In the last section, we determined that feature-based attention did not lead to modulations in unstimulated sub-networks that resembled highly coherent stimulus-driven responses. However, the high-dimensional selectivity and diffuse connectivity of second layer neurons might lead to idiosyncratic, but nevertheless consistent, modulations in unstimulated sub-networks. Thus, we tested whether feedback signals induced incidental, but consistent, activity patterns by training a support vector machine (SVM) to decode the value of the attended stimulus in the stimulated sub-network based on 1) activity patterns in the stimulated sub-network, and 2) activity patterns from each of the unstimulated sub-networks (i.e. separate SVMs trained/tested for each sub-network). In the stimulated sub-network, decoding accuracy increased with stronger top-down gain (Figure 2.4C, solid black line), consistent with the regression-based MAE results above. Critically, activity patterns in the unstimulated sub-networks also supported decoding of the attended feature, with the average decoding accuracy across all unstimulated sub-networks increasing as a function of increasing top-down gain (Figure 2.4C, dotted gray line). Indeed, decoding performance based on the activity patterns in unstimulated sub-networks was better than the decoding performance in the stimulated sub-network at lower feature gain levels. This is likely due to the competition between the attended and unattended inputs in the stimulated sub-network, while the unstimulated sub-networks do not have any bottom-up signals that might disrupt the feedback signals. Collectively, these findings demonstrate that top-down

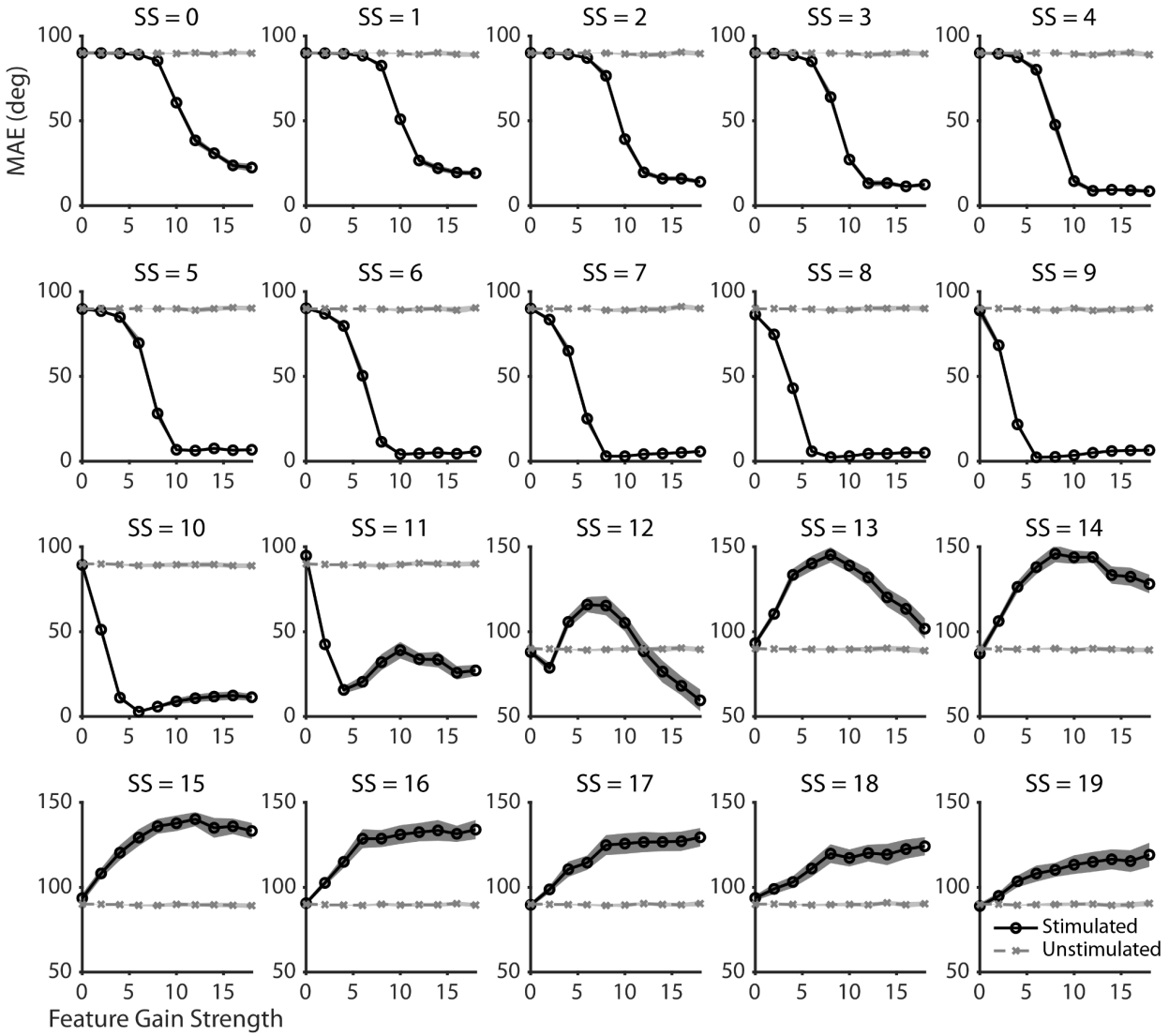


Figure 2.5 Average MAE in Figure 2.4B separated by stimulus strength level, as labeled above each plot. Solid black lines represent Stimulated sub-network and dotted gray lines represent Unstimulated sub-network. Shaded areas represent standard error of mean across network initializations.

feedback that is relayed via neurons with high-dimensional tuning functions leads to systematic changes that spread globally throughout the network depending on the focus of attention, but in a way that is idiosyncratic and different from stimulus-driven responses.

To confirm that this idiosyncratic change in activity from feedback signals in the unstimulated sub-networks are indeed structurally unrelated to the stimulus-driven responses and are solely driven by random patterns of feedback modulations, we further examined whether the attention-dependent activity patterns are consistent across different unstimulated sub-networks. We trained a SVM on the activity patterns from one unstimulated sub-network and cross-generalized that classifier to a different unstimulated sub-network, repeating over all possible training and testing pairs. The average decoding accuracy was close to chance, even with stronger top-down gain (Figure 2.4D). This lack of robust cross-generalization demonstrates that while each unstimulated sub-network exhibited systematic attentional modulations within itself, the modulatory patterns were not shared across the sub-networks. Thus, relaying top-down modulations through high-dimensional neurons leads to systematic modulations in unstimulated networks, while the randomness of the feedback seems to prevent the unintended emergence of stimulus-like representations that might interfere with information processing.

Applying both spatial and feature attention gain results in additive shifts in response functions

Our network model was able to capture standard spatial attention and feature-based attention effects. Based on these results, we took a further step to investigate how the two attention effects would interact within our network. We presented the network with the feature-based attention task while providing spatial attention concurrently to the sub-network where the stimuli were presented (Figure 2.6A). More specifically, spatial attention gain was given to the top 20% of second layer neurons with the highest weights for the first sub-network, then the two stimuli were presented, and then feature-based attention gain was given to the top 20% of the

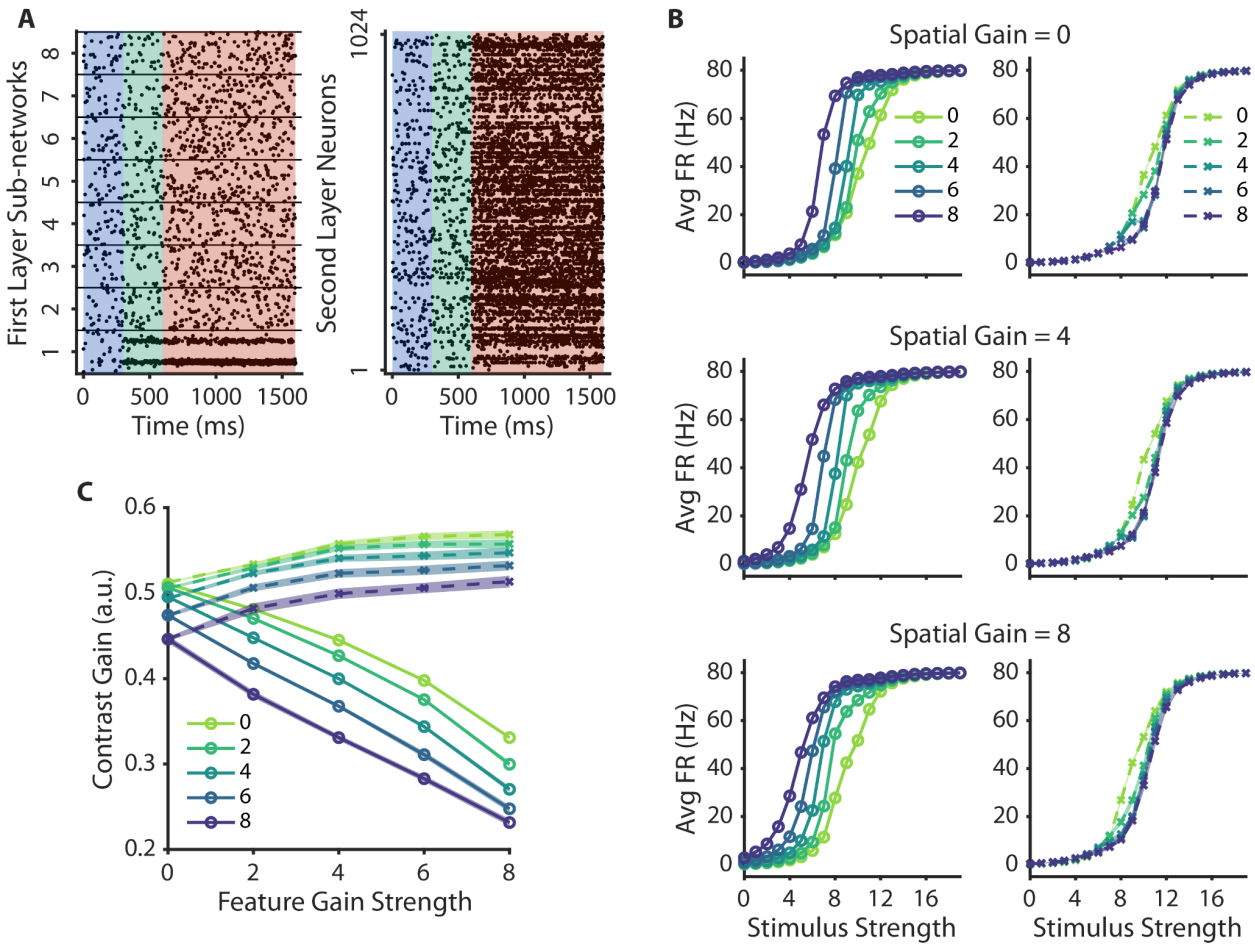


Figure 2.6 Concurrent spatial and feature-based attention simulations. A. An example trial from the *attention* task. First, spatial attention gain was applied to a subset of second layer neurons that have the highest connectivity weights to sub-network 1 in the first layer for 300ms (shaded blue; spatial attention gain strength: 6). Next, two stimulus inputs, at 90° and 270°, were presented to sub-network 1 for 300ms (shaded green; stimulus strength: 7). Then, feature-based attention gain was applied to a subset of second layer neurons that have the highest selectivity to the 90° stimulus in sub-network 1 (lower cluster of spikes) for 1000ms (shaded red; feature-based attention gain strength: 6). B. CRFs for when the stimulus was attended (left panels) and unattended (right panels). Each line represents a different level of feature-based attention gain strength, and each row of panels represents different level of spatial attention gain. C. Estimated contrast gain parameters from Attended (solid lines) and Unattended (dotted lines) conditions. Colors of the lines represent different levels of spatial attention gain. Shaded areas in B, C represent standard error of mean across network initializations.

second layer neurons that maximally respond to the attentional target stimulus. The intensity of the gain for spatial and feature attention was independently modulated, and when both were to be

applied to the same second layer neuron, the intensity values were added and then applied. To keep the network from saturating, we limited the range of both spatial and feature attention gain strength to the lower half of the range we used for previous simulations (0-8 out of 0-18). We measured the average firing rates of neurons that maximally prefer the attended stimulus (Attended condition) or the unattended stimulus (Unattended condition) from the first sub-network for every trial and plotted them as a function of stimulus strength (Figure 2.6B).

When there was no spatial attention gain and only the feature-based attention was modulated, we replicated the results from feature-based attention simulations: We found the contrast gain effect with stronger top-down gain, in which response functions were shifted leftward when higher gain strength was applied for the attended stimulus (Figure 2.6B, top row, left panel). When we applied spatial attention on top of the feature-based attention, we could see that the response functions shift leftward overall with higher spatial attention gain (Figure 2.6B, middle and bottom row, left panels). In contrast, for the unattended stimulus, we could see a slight rightward shift with higher feature-based attention gain and an overall leftward shift with higher spatial attention gain (Figure 2.6B, right panels). To better visualize this interaction, we estimated the contrast gain parameter as previously described and plotted it as a function of spatial and feature-based attention gain (Figure 2.6C). Consistent with previous results and the response function plots, the contrast gain parameter decreased with stronger feature attention gain in no spatial attention condition. Importantly, this decreasing pattern shifted downward in a linear fashion with stronger spatial attention gain. This downward shift was also visible for the unattended stimulus albeit at a smaller scale. This is likely due to spatial attention increasing the overall excitability of the neurons in the targeted sub-network, regardless of which feature the neurons are tuned to, and therefore requiring weaker stimulus strength to show the same level of

firing rates for both attended and unattended stimuli. In all, this suggests that in our network model, spatial attention and feature attention interact additively in enhancing stimulus-driven signals for the attentional target.

Conclusion

While it is widely believed that top-down modulatory signals originate from the prefrontal and parietal cortex, neurons in these areas typically have high-dimensional tuning. As a result, the ability of these neurons to send highly specific feedback signals to early sensory areas is not well understood. Here, we used a spiking neural network to test the hypothesis that random connections between sensory and high-order areas support targeted top-down feedback. Our model recapitulated contrast gain effects with top-down attention, replicating prior empirical findings (2, 100). Importantly, feature-based attentional gain also led to idiosyncratic modulations in unstimulated sub-networks. However, these spurious signals did not resemble coherent sensory responses that can be mistakenly read out as a stimulus input, which is likely due to the random connections between the first "sensory" layer and the second "control" layer. Thus, in this chapter, we demonstrate that feedback signals propagated by neurons with high-dimensional selectivity can spread and cause spurious signals in the network and suggest that connectivity randomness can prevent the rise of "illusory" stimulus representations that can interfere with perceptual processing. In Chapter 3, we test our hypothesis that random projections are crucial for targeted top-down modulations without disruptive spurious signals by parametrically modulating the randomness of the between-layer connections in the spiking network model.

Methods

Computational model

The model was comprised of two layers of Poisson spiking neurons, randomly and reciprocally connected to each other, with parameters adopted from Burak and Fiete (Ref. 107; see also Ref. 101). Each neuron generates spikes based on its firing rate $r_i(t)$, which is determined by a nonlinear function of the weighted sum of all pre-synaptic inputs:

$$r_i(t) = \Phi\left(\sum_j W_{ij} s_j(t)\right)$$

$$spikes(t + dt) \sim Poisson(r_i(t))$$

where i indexes each neuron and j indexes all pre-synaptic neurons. Thus, *Wij* is the synaptic strength from pre-synaptic neuron j to post-synaptic neuron i; $s_j(t)$ is the activation of pre-synaptic neuron j; and Φ is a baseline-shifted hyperbolic tangent: $\tau\Phi(g) = 0.4(1 + tanh(0.25g - 3))$ (107), where g is the total synaptic input, the sum of current synaptic activation, bias, and the stimulus. This function is strictly positive and saturates at an upper bound, just like biological neurons. The synaptic activation of neuron j is determined by convolving the spike train with an exponential function:

$$s_j + \frac{s_j}{\tau} = \sum_\alpha \delta\left(t = t_j^\alpha\right)$$

where t_j^{α} are the spike times of neuron j and the synaptic time constant τ was set to 10ms for all connections (107, 108). While synaptic connections could be either excitatory or inhibitory, this model does not contain different cell-types and neurons do not have refractory periods following each spike. No additional noise was added as the inhomogeneous Poisson processes that generate spikes in our model naturally leads to spiking variability.

The first layer contained 8 ring-like "sensory" sub-networks consisting of 512 neurons, simulating sensory networks that encode information in circular stimulus spaces such as color, orientation or motion. Every neuron in each sub-network had a preferred stimulus input, operationalized here as an angle $\theta_i = 2\pi i/N_{first}$. This tuning arises due to short-range excitatory connections and long-range inhibitory connections between each neuron and its neighbors around the ring, where the synaptic weight between any pair of neurons i, j within a sub-network depends only on $\theta = \theta_i - \theta_j$ through a difference of circular normal functions:

$$W_{ij}^{first} = w(\theta) = \lambda + Aexp(k_1(cos\theta - 1)) - Aexp(k_2(cos\theta - 1))$$

where k_1= 1 determines the width of the excitation kernel, k_2 = 0.83 determines the width of the suppression kernel, A = 2 is the amplitude and λ = 0.28 is the baseline. Self-excitation is set to 0 ($w(0) = 0$). The parameters were adopted from Bouchacourt and Buschman (101), with the exception that we used a higher value for k_2, which allowed a single sub-network to maintain representations for two stimuli simultaneously, a crucial feature for simulating feature-based attention effects with two competing stimuli in our model.

For present purposes, we conceptualize each sub-network to encode independent sensory inputs such as orientations presented in different spatial positions around the visual field. Note

that sub-networks were not directly connected to each other. However, neurons in the 8 sub-networks indirectly interacted via feed-forward and feedback projections to the second layer.

The second layer is composed of 1024 "control" neurons that are connected to the neurons in the 8 sub-networks, leading to a four-fold convergence. All sub-networks converged onto the second layer through a weight matrix (W^{FF}) that determined the feedforward connectivity. Then, each neuron in the second layer network fed back into the same subset of neurons in the first layer that provided excitatory inputs, through the feedback connectivity matrix (W^{FB}), which was the transpose of the feedforward connectivity matrix. This made the connections bi-directional. For a given pair of first layer and second layer neurons, the likelihood of an excitatory connection was 0.35.

The strength of the excitatory feedforward and feedback connections were defined by the parameters α and β, respectively. We adopted α = 2100 and β = 200 from Bouchacourt and Buschman (101) that yields optimal memory performance that maximizes the number of maintained memory items and minimizes the number of spurious memories in their task.

The between-layer connectivity weights were balanced in two ways. First, to balance the total excitatory drive across neurons, the feedforward and feedback weights were scaled by the number of total inter-network excitatory connections (N_{exc}) for each neuron. Second, an equal inhibitory weight was applied to all inputs for each neuron ($-\alpha/(8 * N_{first})$) so that each neuron receives an equal amount of excitatory and inhibitory drive (i.e., $\Sigma_j W_{ij}^{FF} = \Sigma_j W_{ij}^{FB} = 0$). This was to mimic the effect of local inhibitory interneurons causing a broad inhibition in the target network.

After the balancing process, excitatory feedforward connections from the sensory to second layer neuron i had weight $W_{i,exc} = (\alpha/N_{exc_i}) - (\alpha/(8 * N_{first}))$ while inhibitory feedforward connections had weight $W_{i,inh} = -\alpha/(8 * N_{first})$. Similarly, neuron j in the sensory network will receive excitatory feedback connections with weight $W_{j,exc} = (\beta/N_{exc_j}) - (\beta/N_{second})$ and inhibitory feedback connections with weight $W_{j,inh} = -\beta/N_{second}$.

Presentation of stimuli to the sensory network

Sensory stimuli were provided as synaptic drive (s_{ext}) to the first layer sub-networks. So, the response of a given first layer neuron i at a given time t was:

$$r_i^{first}(t) = \Phi\left(\sum_{j \in second} W_{ij}^{FB} s_j(t) + \sum_{j \in first} W_{ij}^{first} s_j(t) + s_i^{ext}(t)\right)$$

Stimulus inputs were presented for varying durations, indicated by the green and red shaded region in Figures 2.2A, 2.3A, 2.3B, 2.6A. Inputs were a circular normal distribution centered on an angle (μ) between 1° and 360° that was chosen to best visualize the spikes in the raster plot. Specifically, the angle was 180° when one stimulus was presented to a sub-network, and 90° and 270°, when two stimuli were presented to a sub-network. The stimulus input to neuron i was

$$s_i^{ext} = \left\{S^{ext} \frac{exp(\kappa\, cos(x - \mu))}{2\pi I_0(\kappa)}\ if\ i \leq [3\sigma], 0\ if\ i > [3\sigma]\right\}$$

where κ = 14, approximating the shape of the gaussian function used in Bouchacourt and Buschman (101) for stimulus inputs. Inputs above and below three standard deviations were set to zero. S^{ext} was the strength of external sensory input that was modulated based on the task between 0 and 19. In cases where two stimuli were presented simultaneously, the two circular

normal distributions centered at different angles were summed and provided as the stimulus input. In all simulations, stimuli were presented to a single sub-network at a time.

Applying Top-down attention to the second layer

Top-down attention gain was provided as synaptic drive (s_{att}) to the second layer. So, the response of a given second layer neuron i at a given time t was:

$$r_i^{second}(t) = \Phi\left(\sum_{j \in first} W_{ij}^{FF} s_j(t) + S_i^{att}(t)\right)$$

where S^{att} was the strength of top-down attention gain that was modulated based on the task between 0 and 18, in increments of 2.

Simulating spatial attention

To simulate and test the effects of spatial attention, we chose a subset of the second layer neurons that showed preference for a given sub-network, as we consider each sub-network as encoding a visual feature in different spatial locations. This was done by taking the feedback connectivity weights between the second layer neurons and a sub-network (1024 X 512), averaging the weights for each second layer neuron, and then rank ordering based on the average weights each second layer neuron has for this sub-network. The top 20 percent of the second layer neurons were defined to have a preference for this sub-network.

In the *attention* task for simulating spatial attention, attentional gain was applied to the second layer neurons with the preference for one of the sub-networks (spatial attention target) for 200ms before a single stimulus was presented to either the attended sub-network (Attended condition) or an unattended sub-network (Unattended condition) for 1000ms. We varied the

stimulus input strength between 0 and 20 in increments of 1, and spatial attentional gain between 0 and 20 in increments of 2. Note that these values are in arbitrary units, and these ranges were intentionally chosen, as they allow us to observe the attentional effects of interest within our model. We simulated 50 trials for each condition.

Simulating feature-based attention

To simulate and test the effects of feature-based attention, we chose a subset of the second layer neurons that showed preference for the attentional target. In a separate simulation, the to-be-attended stimulus feature was presented to a single sub-network for 300ms. Then, the firing rates of the second layer neurons during stimulus presentation were averaged for each second layer neuron, rank-ordered, and then top 20 percent of the second layer neurons that showed the highest firing rates for the presented stimulus was chosen as preferring this stimulus feature.

In the *attention* task for simulating feature-based attention, two stimuli were presented to a single sub-network, and the attentional gain was concurrently applied to the second layer neurons with the preference for one of the two stimuli, for 1000ms. We varied the stimulus input strength between 0 and 20 in increments of 1 and feature attentional gain between 0 and 20 in increments of 2. We simulated 50 trials for each condition.

To evaluate the similarity of representations between attention signals and sensory signals, we conducted simulations to record stimulus-driven signals. In the *sensory* task, we presented one of 16 different stimulus features equally spaced in the stimulus space for 300ms to each of the eight sub-networks. We simulated 100 trials for each stimulus and each sub-network.

Simulating concurrent spatial and feature-based attention

To simulate the interaction between spatial and feature-based attention, we chose two sets of second layer neurons, each for spatial attention and feature-based attention following the procedure described above. Then attentional gain for spatial and feature-based attention was applied to each set. In the case where there was an overlap between the two sets of neurons, the sum of the two attentional gain intensities was applied to those neurons.

In the *attention* task for simulating spatial and feature-based attention, spatial attention gain was applied to one sub-network for 300ms, then two stimuli were presented to that sub-network for 300ms, as in the *attention* task for feature-based attention simulation. Next, feature-based attention gain was applied to one of the two presented stimuli for 1000ms. We varied the stimulus input strength between 0 and 20 in increments of 1, and spatial attentional gain and feature attentional gain were modulated independently between 0 and 8 in increments of 2. Note that we chose a lower range of attentional gain strength to prevent the network from saturating. We simulated 50 trials for each condition.

Evaluating the effects of modulation in stimulus strength and top-down attention

To quantify and compare the stimulus- or attention-related signals in each of the eight sub-networks, the firing rates for each neuron was averaged across the last 250ms of each trial of every simulation. We defined that the neurons that the presented stimulus input (a circular normal distribution) was centered on as having the highest preference for that stimulus based on the ring-like structure of the sub-networks, and thus, averaged their firing rates to plot the CRFs. Specifically, to plot the CRFs for the Attended condition in the spatial attention stimulations, we took the average firing rates of the two neurons that have the highest preference for the presented

stimulus for trials in which the spatial attention was given to the stimulated sub-network, and then averaged across trials for each stimulus strength and attention gain level. For the Unattended condition in the spatial attention simulations, the same was done over trials in which the spatial attention was given to an unstimulated sub-network, analogous to a situation in which there is a mismatch between the attended spatial location and the stimulus location. For feature-based attention and concurrent spatial and feature-based attention simulations, the Attended condition CRFs were drawn from averaging firing rates of the two neurons that have the highest preference for the attended stimulus (out of the two stimulus inputs) for every trial for each level of stimulus strength and attention gain. For the Unattended condition CRFs, we averaged firing rates of the two neurons that have the highest preference the unattended stimulus.

To estimate the contrast gain of the CRFs, we fit a Naka-Rushton function to the CRFs (109):

$$R(c) = G_r \frac{c^n}{c^n + G_c^n} + b_c$$

where G_r is the multiplicative response gain, G_c is the contrast gain that controls the horizontal shift of the CRFs, b is the response baseline offset, and n is the exponent that controls the speed at which the CRF rises and reaches asymptote. The fitting procedure was done with G_c and n as the two free parameters and G_r and b as fixed parameters: G_r fixed to the difference between the highest and lowest firing rates and b fixed to the lowest firing rates for each CRF. There were two steps in the curve fitting procedure. First, we performed an initial grid search to find the best set of the two free parameters that yields the minimum root mean squared error. Then, to fine-tune the search, we used SciPy optimize to minimize the root mean squared error between the data and the fit function, with the best parameter set from the grid search as the initial seed for

the search. The free parameters were restricted to the range of -10 and 10 for n, with the step size 0.1 for the grid search, and -1 and 1 for G_c, with the step size 0.01 for the grid search.

Decoding analysis

To understand the nature of the feedback signal that is propagated to the first layer sub-networks in the feature-based attention simulations, we used a decoding method with a circular regression model to compare bottom-up stimulus-driven signals and signals modulated by top-down attentional feedback.

First, to ensure that the circular regression model that we use is able to learn the signals for different stimulus features and make accurate predictions on which stimulus was presented, we trained and tested the model on the *sensory* task data for the stimulated sub-network. The training and test data was split in such a way that one trial for each presented stimulus feature was assigned to the test data and the rest of the trials were assigned to the training data. Then, the circular regression model with empirical Bayesian ridge regression was trained on the training data to predict which stimulus was presented in the test data. We took the absolute difference between the prediction and the actual stimulus value/angle and averaged across the trials to get the mean absolute error (MAE) value. We repeated this procedure, looping over different subsets of trials for the test data so that all trials had a chance of being included in the test data. The resulting MAE values were averaged over these cross-validation iterations.

Next, to see if the top-down attentional modulations result in stimulus-specific representational changes for the attended stimulus, we used a cross-generalization method, in which we trained the regression model on the *sensory* task data and tested on the *attention* task data to predict which stimulus feature was attended in the stimulated sub-network. This is based

on the logic that if the attentional feedback signals lead to changes in signals that are stimulus-specific, resembling enhanced stimulus responses, the regression model trained on the presented stimulus in the *sensory* task will successfully predict the attended stimulus in the *attention* task. For this, we trained the model on all of the *sensory* task data and then generated predictions for each trial in the *attention* task data. Average MAE was calculated from taking the average absolute difference between the predicted stimulus and attended stimulus value for each trial and averaging across all trials. Since the stimulus strength was fixed at 10 for the *sensory* task but varied from 0 to 18 in the *attention* task, we additionally examined if this discrepancy would lead to any changes in the decoding performance, by averaging MAE for each stimulus strength level separately (Figure 2.5).

In addition to this sensory-to-attention comparison, we wanted to understand what kind of signal is getting propagated to the unstimulated sub-networks with the top-down attention modulations. First, we wanted to test if these feedback signals in the unstimulated sub-networks are consistent for a given attended feature, regardless of if these signals resemble stimulus-driven responses. To achieve this, we trained a linear support vector machine (SVM) to classify the attended stimulus in the *attention* task. We performed cross-validations by leaving two trials out, one each for which of the two stimuli was attended, training the classifier on the rest of the data and then testing on the two left out trials, repeating until every trial had been trained and tested on, and then averaging the classifier performance across all iterations. This procedure was done separately for each of the 8 sub-networks.

Second, we wanted to test if the attentional feedback signals were systematic across the unstimulated sub-networks. To do this, we used cross-generalization between the unstimulated sub-networks: We trained the classifier on a given unstimulated sub-network to classify which

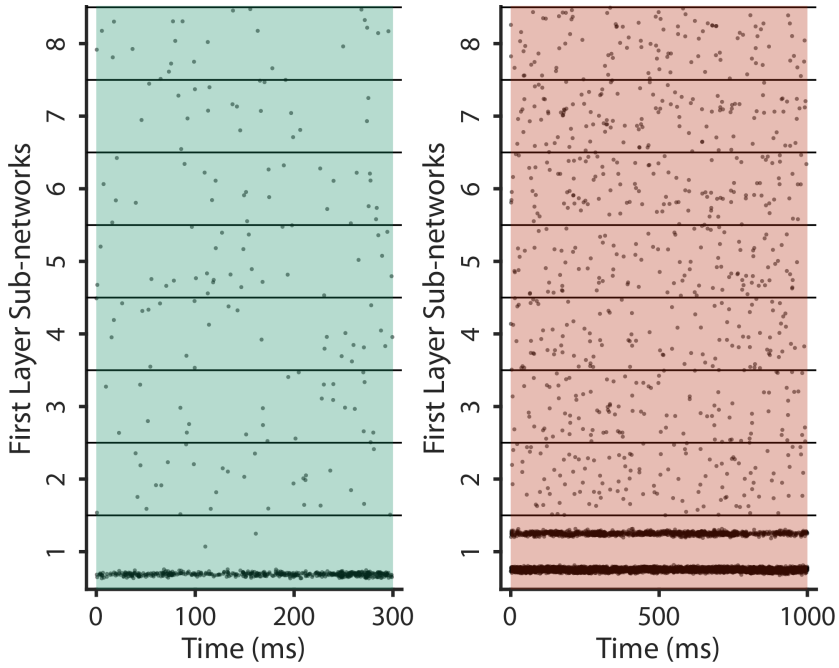
stimulus was attended in the stimulated sub-network, and then tested on the data from a different unstimulated sub-network, to see how well the attention-related activity pattern learned by the classifier is generalizable between two unstimulated sub-networks. We looped over the seven unstimulated sub-networks, using data from each sub-network as the training dataset and testing on each of the remaining six, and then averaged the decoding accuracy across all combinations.

Network Initializations

To ensure that the effects we observe in the simulations are not due to a specific wiring or a connectivity weight matrix, we repeated our simulations across ten iterations using different random seeds to initialize each network. Importantly, this gave rise to different connectivity patterns and weights between the first and the second layer for each initialization. All of the plots show averaged data across these ten iterations with the shaded region as the standard error of mean.

Acknowledgements

Chapter 2, in part, is currently being prepared for submission for publication of the material. Park, Sunyoung; Serences, John. T. The dissertation author was the primary researcher and author of this material.



Supplementary Figure 2.1 Comparison of raster plots for *sensory* task with stimulus strength at 10 and *attention* task with stimulus strength at 12 and feature-based attention gain strength at 6. The presented stimulus in *sensory* task and the attended stimulus in the *attention* task is 90°.

References

1. **Desimone R**, **Duncan J**. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18: 193–222, 1995. doi: 10.1146/annurev.ne.18.030195.001205.

2. **Reynolds J**, **Chelazzi L**. Attentional Modulation of Visual Processing. *Annu Rev Neurosci* 27: 611–47, 2004. doi: 10.1146/annurev.neuro.26.041002.131039.

3. **Corbetta M**, **Shulman GL**. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3: 201–215, 2002. doi: 10.1038/nrn755.

4. **Moore T**, **Zirnsak M**. Neural Mechanisms of Selective Visual Attention. *Annu Rev Psychol* 68: 47–72, 2017. doi: 10.1146/annurev-psych-122414-033400.

5. **Livingstone M**, **Hubel D**. Segregation of Form, Color, Movement, and Depth: Anatomy, Physiology, and Perception. *Sci New Ser* 240: 740–749, 1988.

6. **Victor JD**, **Purpura K**, **Katz E**, **Mao B**. Population encoding of spatial frequency, orientation, and color in macaque V1. .

7. **Hubel DH**, **Wiesel TN**. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160: 106-154.2, 1962.

8. **Hubel DH**, **Wiesel TN**. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195: 215–243, 1968. doi: 10.1113/jphysiol.1968.sp008455.

9. **Hubel DH**, **Wiesel TN**. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *J Comp Neurol* 158: 295–305, 1974.

10. **Blatt GJ**, **Andersen RA**, **Stoner GR**. Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. *J Comp Neurol* 299: 421–445, 1990. doi: 10.1002/cne.902990404.

11. **Cavanaugh J**, **Joiner WM**, **Wurtz RH**. Suppressive surrounds of receptive fields in monkey frontal eye field. *J Neurosci* 32: 12284–12293, 2012.

12. **Mayo JP**, **DiTomasso AR**, **Sommer MA**, **Smith MA**. Dynamics of visual receptive fields in the macaque frontal eye field. *J Neurophysiol* 114: 3201–3210, 2015. doi: 10.1152/jn.00746.2015.

13. **Mohler CW**, **Goldberg ME**, **Wurtz RH**. Visual receptive fields of frontal eye field neurons. .

14. **Miller EK**, **Cohen JD**. An Integrative Theory of Prefrontal Cortex Function. *Annu Rev Neurosci* 24: 167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167.

15. **Funahashi S**. Space representation in the prefrontal cortex. *Prog Neurobiol* 103: 131–155, 2013. doi: 10.1016/j.pneurobio.2012.04.002.

16. **D'Esposito M**, **Postle BR**. The Cognitive Neuroscience of Working Memory. *Annu Rev Psychol* 66: 115–142, 2015. doi: 10.1146/annurev-psych-010814-015031.

17. **Fusi S**, **Miller EK**, **Rigotti M**. Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol* 37: 66–74, 2016. doi: 10.1016/j.conb.2016.01.010.

18. **Rigotti M**, **Barak O**, **Warden MR**, **Wang X-J**, **Daw ND**, **Miller EK**, **Fusi S**. The importance of mixed selectivity in complex cognitive tasks. *Nature* 497: 585–590, 2013. doi: 10.1038/nature12160.

19. **Moore T**, **Armstrong KM**. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373, 2003. doi: 10.1038/nature01341.

20. **Mendoza-Halliday D**, **Torres S**, **Martinez-Trujillo JC**. Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat Neurosci* 17: 1255–1262, 2014. doi: 10.1038/nn.3785.

21. **Zaksas D**, **Pasternak T**. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J Neurosci Off J Soc Neurosci* 26: 11726–11742, 2006. doi: 10.1523/JNEUROSCI.3420-06.2006.

22. **McAdams CJ**, **Maunsell JH**. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci* 19: 431–441, 1999.

23. **Treue S**, **Trujillo JCM**. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399: 575–579, 1999. doi: 10.1038/21176.

24. **Motter BC**. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J Neurosci* 14: 2178–2189, 1994. doi: 10.1523/JNEUROSCI.14-04-02178.1994.

25. **Pollen DA**, **Przybyszewski AW**, **Rubin MA**, **Foote W**. Spatial receptive field organization of macaque V4 neurons. *Cereb Cortex N Y N 1991* 12: 601–616, 2002. doi: 10.1093/cercor/12.6.601.

26. **Movshon JA**, **Adelson EH**, **Gizzi MS**, **Newsome WT**. The Analysis of Moving Visual Patterns. In: *Pattern Recognition Mechanisms*, edited by Chagas C, Gattass R, Gross C. Springer Berlin Heidelberg, p. 117–151.

27. **Zeki SM**. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J Physiol* 236: 549–573, 1974.

28. **Pasupathy A**, **Connor CE**. Responses to contour features in macaque area V4. *J Neurophysiol* 82: 2490–2502, 1999.

29. **Pasupathy A**, **Connor CE**. Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol* 86: 2505–2519, 2001. doi: 10.1152/jn.2001.86.5.2505.

30. **De Valois RL**, **Albrecht DG**, **Thorell LG**. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res* 22: 545–559, 1982. doi: 10.1016/0042-6989(82)90113-4.

31. **De Valois RL**, **Yund EW**, **Hepler N**. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Res* 22: 531–544, 1982. doi: 10.1016/0042-6989(82)90112-2.

32. **Movshon JA**. The velocity tuning of single units in cat striate cortex. *J Physiol* 249: 445–468, 1975. doi: 10.1113/jphysiol.1975.sp011025.

33. **Albright TD**, **Desimone R**, **Gross CG**. Columnar organization of directionally selective cells in visual area MT of the macaque. *J Neurophysiol* 51: 16–31, 1984. doi: 10.1152/jn.1984.51.1.16.

34. **Rust NC**, **Dicarlo JJ**. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci Off J Soc Neurosci* 30: 12978–12995, 2010. doi: 10.1523/JNEUROSCI.0179-10.2010.

35. **Yamins DLK**, **Hong H**, **Cadieu CF**, **Solomon EA**, **Seibert D**, **DiCarlo JJ**. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci* 111: 8619–8624, 2014. doi: 10.1073/pnas.1403112111.

36. **Felleman DJ**, **Van Essen DC**. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb Cortex* 1: 1–47, 1991. doi: 10.1093/cercor/1.1.1.

37. **Kobatake E**, **Tanaka K**. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol* 71: 856–867, 1994. doi: 10.1152/jn.1994.71.3.856.

38. **Van Essen DC**, **Anderson CH**, **Felleman DJ**. Information processing in the primate visual system: an integrated systems perspective. *Science* 255: 419–423, 1992. doi: 10.1126/science.1734518.

39. **Riesenhuber M**, **Poggio T**. Hierarchical models of object recognition in cortex. *Nat Neurosci* 2: 1019–1025, 1999. doi: 10.1038/14819.

40. **Maunsell JHR**, **Sclar G**, **Nealey TA**, **DePriest DD**. Extraretinal representations in area V4 in the macaque monkey. *Vis Neurosci* 7: 561–573, 1991. doi: 10.1017/S095252380001035X.

41. **Maunsell JH**, **Van Essen DC**. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J Neurophysiol* 49: 1127–1147, 1983. doi: 10.1152/jn.1983.49.5.1127.

42. **Martinez-Trujillo JC**, **Treue S**. Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. *Curr Biol* 14: 744–751, 2004. doi: 10.1016/j.cub.2004.04.028.

43. **Treue S**, **Maunsell JH**. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* 382: 539–541, 1996. doi: 10.1038/382539a0.

44. **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* 77: 24–42, 1997. doi: 10.1152/jn.1997.77.1.24.

45. **Chelazzi L**, **Miller EK**, **Duncan J**, **Desimone R**. A neural basis for visual search in inferior temporal cortex. *Nature* 363: 345–347, 1993. doi: 10.1038/363345a0.

46. **Chelazzi L**, **Duncan J**, **Miller EK**, **Desimone R**. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiol* 80: 2918–2940, 1998. doi: 10.1152/jn.1998.80.6.2918.

47. **Haenny PE**, **Maunsell JHR**, **Schiller PH**. State dependent activity in monkey visual cortex. *Exp Brain Res* 69: 245–259, 1988. doi: 10.1007/BF00247570.

48. **Haenny PE**, **Schiller PH**. State dependent activity in monkey visual cortex. *Exp Brain Res* 69: 225–244, 1988. doi: 10.1007/BF00247569.

49. **Motter BC**. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J Neurophysiol* 70: 909–919, 1993. doi: 10.1152/jn.1993.70.3.909.

50. **Motter BC**. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J Neurosci* 14: 2190–2199, 1994. doi: 10.1523/JNEUROSCI.14-04-02190.1994.

51. **Connor CE**, **Gallant JL**, **Preddie DC**, **Van Essen DC**. Responses in area V4 depend on the spatial relationship between stimulus and attention. *J Neurophysiol* 75: 1306–1308, 1996. doi: 10.1152/jn.1996.75.3.1306.

52. **Connor CE**, **Preddie DC**, **Gallant JL**, **Essen DCV**. Spatial Attention Effects in Macaque Area V4. *J Neurosci* 17: 3201–3214, 1997. doi: 10.1523/JNEUROSCI.17-09-03201.1997.

53. **Serences JT**, **Ester EF**, **Vogel EK**, **Awh E**. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychol Sci* 20: 207–214, 2009. doi: 10.1111/j.1467-9280.2009.02276.x.

54. **Ho T**, **Brown S**, **Maanen L van**, **Forstmann BU**, **Wagenmakers E-J**, **Serences JT**. The Optimality of Sensory Processing during the Speed–Accuracy Tradeoff. *J Neurosci* 32: 7992–8003, 2012. doi: 10.1523/JNEUROSCI.0340-12.2012.

55. **Rust NC**, **Cohen MR**. Priority coding in the visual system. *Nat Rev Neurosci* 23: 376–388, 2022. doi: 10.1038/s41583-022-00582-9.

56. **Reynolds JH**, **Pasternak T**, **Desimone R**. Attention Increases Sensitivity of V4 Neurons. *Neuron* 26: 703–714, 2000. doi: 10.1016/S0896-6273(00)81206-4.

57. **Maunsell JHR**, **Treue S**. Feature-based attention in visual cortex. *Trends Neurosci* 29: 317–322, 2006. doi: 10.1016/j.tins.2006.04.001.

58. **Kastner S**, **Ungerleider L**. Mechanisms of Visual Attention in the Human Cortex. *Annu Rev Neurosci* 23: 315–341, 2000. doi: 10.1146/annurev.neuro.23.1.315.

59. **Yantis S**, **Serences JT**. Cortical mechanisms of space-based and object-based attentional control. *Curr Opin Neurobiol* 13: 187–193, 2003. doi: 10.1016/S0959-4388(03)00033-3.

60. **Chun MM**, **Marois R**. The dark side of visual attention. *Curr Opin Neurobiol* 12: 184–189, 2002. doi: 10.1016/S0959-4388(02)00309-4.

61. **Schall JD**. The neural selection and control of saccades by the frontal eye field. *Philos Trans R Soc Lond B Biol Sci* 357: 1073–1082, 2002. doi: 10.1098/rstb.2002.1098.

62. **Buschman TJ**, **Miller EK**. Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science* 315: 1860–1862, 2007. doi: 10.1126/science.1138071.

63. **Squire RF**, **Noudoost B**, **Schafer RJ**, **Moore T**. Prefrontal Contributions to Visual Selective Attention. .

64. **Noudoost B**, **Chang MH**, **Steinmetz NA**, **Moore T**. Top-down control of visual attention. .

65. **Schall J**, **Morel A**, **King D**, **Bullier J**. Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J Neurosci* 15: 4464–4487, 1995. doi: 10.1523/JNEUROSCI.15-06-04464.1995.

66. **Thompson KG**, **Bichot NP**. A visual salience map in the primate frontal eye field. *Prog Brain Res* 147: 251–262, 2005. doi: 10.1016/S0079-6123(04)47019-8.

67. **Bichot NP**, **Schall JD**, **Thompson KG**. Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature* 381: 697–699, 1996. doi: 10.1038/381697a0.

68. **Bichot NP**, **Schall JD**. Effects of similarity and history on neural mechanisms of visual selection. *Nat Neurosci* 2: 549–554, 1999. doi: 10.1038/9205.

69. **Thompson KG**, **Schall JD**. The detection of visual signals by macaque frontal eye field during masking. *Nat Neurosci* 2: 283–288, 1999. doi: 10.1038/6398.

70. **Bichot NP**, **Heard MT**, **DeGennaro EM**, **Desimone R**. A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron* 88: 832–844, 2015. doi: 10.1016/j.neuron.2015.10.001.

71. **Bichot NP**, **Xu R**, **Ghadooshahy A**, **Williams ML**, **Desimone R**. The role of prefrontal cortex in the control of feature attention in area V4. *Nat Commun* 10: 5727, 2019. doi: 10.1038/s41467-019-13761-7.

72. **Stokes MG**, **Buschman TJ**, **Miller EK**. Dynamic Coding for Flexible Cognitive Control. In: *The Wiley Handbook of Cognitive Control*. John Wiley & Sons, Ltd, p. 221–241.

73. **Panichello MF**, **Buschman TJ**. Shared mechanisms underlie the control of working memory and attention. *Nature* 592: 601–605, 2021. doi: 10.1038/s41586-021-03390-w.

74. **Badre D**. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends Cogn Sci* 12: 193–200, 2008. doi: 10.1016/j.tics.2008.02.004.

75. **Milner B**. Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. *Arch Neurol* 9: 90–100, 1963. doi: 10.1001/archneur.1963.00460070100010.

76. **Dias R**, **Robbins TW**, **Roberts AC**. Primate analogue of the Wisconsin card sorting test: Effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behav Neurosci* 110: 872–886, 1996. doi: 10.1037/0735-7044.110.5.872.

77. **Dias R**, **Robbins TW**, **Roberts AC**. Dissociable Forms of Inhibitory Control within Prefrontal Cortex with an Analog of the Wisconsin Card Sort Test: Restriction to Novel Situations and Independence from "On-Line" Processing. *J Neurosci* 17: 9285–9297, 1997. doi: 10.1523/JNEUROSCI.17-23-09285.1997.

78. **Chao LL**, **Knight RT**. Contribution of human prefrontal cortex to delay performance. *J Cogn Neurosci* 10: 167–177, 1998. doi: 10.1162/089892998562636.

79. **Rossi AF**, **Bichot NP**, **Desimone R**, **Ungerleider LG**. Top–Down Attentional Deficits in Macaques with Lesions of Lateral Prefrontal Cortex. *J Neurosci* 27: 11306–11314, 2007. doi: 10.1523/JNEUROSCI.2939-07.2007.

80. **Rossi AF**, **Pessoa L**, **Desimone R**, **Ungerleider LG**. The prefrontal cortex and the executive control of attention. *Exp Brain Res* 192: 489–497, 2009. doi: 10.1007/s00221-008-1642-z.

81. **Miller BT**, **Vytlacil J**, **Fegen D**, **Pradhan S**, **D'Esposito M**. The Prefrontal Cortex Modulates Category Selectivity in Human Extrastriate Cortex. *J Cogn Neurosci* 23: 1–10, 2011. doi: 10.1162/jocn.2010.21516.

82. **Petrides M**. Monitoring of selections of visual stimuli and the primate frontal cortex. *Proc Biol Sci* 246: 293–298, 1991. doi: 10.1098/rspb.1991.0157.

83. **Feredoes E**, **Heinen K**, **Weiskopf N**, **Ruff C**, **Driver J**. Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. *Proc Natl Acad Sci* 108: 17510–17515, 2011. doi: 10.1073/pnas.1106439108.

84. **Nee DE**, **D'Esposito M**. Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *eLife* 6: e28040, 2017. doi: 10.7554/eLife.28040.

85. **Mendoza-Halliday D**, **Xu H**, **Azevedo FAC**, **Desimone R**. Dissociable neuronal substrates of visual feature attention and working memory. *Neuron* 112: 850-863.e6, 2024. doi: 10.1016/j.neuron.2023.12.007.

86. **Bressler SL**, **Tang W**, **Sylvester CM**, **Shulman GL**, **Corbetta M**. Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex in Anticipatory Visual Spatial Attention. *J Neurosci* 28: 10056–10061, 2008. doi: 10.1523/JNEUROSCI.1776-08.2008.

87. **Egner T**, **Monti JMP**, **Trittschuh EH**, **Wieneke CA**, **Hirsch J**, **Mesulam M-M**. Neural Integration of Top-Down Spatial and Feature-Based Information in Visual Search. *J Neurosci* 28: 6141–6151, 2008. doi: 10.1523/JNEUROSCI.1262-08.2008.

88. **Gazzaley A**, **Nobre AC**. Top-down modulation: bridging selective attention and working memory. *Trends Cogn Sci* 16: 129–135, 2012. doi: 10.1016/j.tics.2011.11.014.

89. **Giesbrecht B**, **Woldorff MG**, **Song AW**, **Mangun GR**. Neural mechanisms of top-down control during spatial and feature attention. *NeuroImage* 19: 496–512, 2003. doi: 10.1016/S1053-8119(03)00162-9.

90. **Zanto TP**, **Rubens MT**, **Bollinger J**, **Gazzaley A**. Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage* 53: 736–745, 2010. doi: 10.1016/j.neuroimage.2010.06.012.

91. **Zanto TP**, **Rubens MT**, **Thangavel A**, **Gazzaley A**. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat Neurosci* 14: 656–661, 2011. doi: 10.1038/nn.2773.

92. **Baldauf D**, **Desimone R**. Neural Mechanisms of Object-Based Attention. *Science* 344: 424–427, 2014. doi: 10.1126/science.1247003.

93. **Gong M**, **Liu T**. Biased Neural Representation of Feature-Based Attention in the Human Frontoparietal Network. *J Neurosci* 40: 8386–8395, 2020. doi: 10.1523/JNEUROSCI.0690-20.2020.

94. **Bedini M**, **Baldauf D**. Structure, function and connectivity fingerprints of the frontal eye field versus the inferior frontal junction: A comprehensive comparison. *Eur J Neurosci* 54: 5462–5506, 2021. doi: 10.1111/ejn.15393.

95. **Meyyappan S**, **Rajan A**, **Mangun GR**, **Ding M**. Role of Inferior Frontal Junction (IFJ) in the Control of Feature versus Spatial Attention. *J Neurosci* 41: 8065–8074, 2021. doi: 10.1523/JNEUROSCI.2883-20.2021.

96. **Tye KM**, **Miller EK**, **Taschbach FH**, **Benna MK**, **Rigotti M**, **Fusi S**. Mixed selectivity: Cellular computations for complexity. *Neuron* 112: 2289–2303, 2024. doi: 10.1016/j.neuron.2024.04.017.

97. **Bruce CJ**, **Goldberg ME**, **Bushnell MC**, **Stanton GB**. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J Neurophysiol* 54: 714–734, 1985. doi: 10.1152/jn.1985.54.3.714.

98. **Zipursky SL**, **Sanes JR**. Chemoaffinity Revisited: Dscams, Protocadherins, and Neural Circuit Assembly. *Cell* 143: 343–353, 2010. doi: 10.1016/j.cell.2010.10.009.

99. **Buschman TJ**. Balancing Flexibility and Interference in Working Memory. *Annu Rev Vis Sci* 7: 367–388, 2021. doi: 10.1146/annurev-vision-100419-104831.

100. **Reynolds JH**, **Heeger DJ**. The Normalization Model of Attention. *Neuron* 61: 168–185, 2009. doi: 10.1016/j.neuron.2009.01.002.

101. **Bouchacourt F**, **Buschman TJ**. A Flexible Model of Working Memory. *Neuron* 103: 147-160.e8, 2019. doi: 10.1016/j.neuron.2019.04.020.

102. **Li X**, **Basso MA**. Preparing to Move Increases the Sensitivity of Superior Colliculus Neurons. *J Neurosci* 28: 4561–4577, 2008. doi: 10.1523/JNEUROSCI.5683-07.2008.

103. **Martínez-Trujillo JC**, **Treue S**. Attentional Modulation Strength in Cortical Area MT Depends on Stimulus Contrast. *Neuron* 35: 365–370, 2002. doi: 10.1016/S0896-6273(02)00778-X.

104. **Mather M**, **Clewett D**, **Sakaki M**, **Harley CW**. Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behav Brain Sci* 39: e200, 2016. doi: 10.1017/S0140525X15000667.

105. **Bichot NP**, **Rossi AF**, **Desimone R**. Parallel and Serial Neural Mechanisms for Visual Search in Macaque Area V4. *Science* 308: 529–534, 2005. doi: 10.1126/science.1109676.

106. **Saenz M**, **Buracas GT**, **Boynton GM**. Global effects of feature-based attention in human visual cortex. *Nat Neurosci* 5: 631–632, 2002. doi: 10.1038/nn876.

107. **Burak Y**, **Fiete IR**. Fundamental limits on persistent activity in networks of noisy neurons. *Proc Natl Acad Sci* 109: 17645–17650, 2012. doi: 10.1073/pnas.1117386109.

108. **Burak Y**, **Fiete IR**. Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. *PLOS Comput Biol* 5: e1000291, 2009. doi: 10.1371/journal.pcbi.1000291.

109. **Geisler WS**, **Albrecht DG**. Visual cortex neurons in monkeys and cats: Detection, discrimination, and identification. *Vis Neurosci* 14: 897–919, 1997. doi: 10.1017/S0952523800011627.

Chapter 3 The role of random projections in shaping feature-specific attentional modulations

Introduction

Prior empirical studies have reported the spreading of the feature-based attentional signals across the visual field. In these studies, the response of a neuron is enhanced according to the similarity between the feature it is tuned to and the attended feature, irrespective of the location of the neuron's spatial receptive field (1–5). In contrast, our network model in Chapter 2 showed that feedback signals propagated by neurons with high-dimensional tuning functions lead to stable activity patterns that can support decoding, but these activity patterns do not resemble the 'coherent' stimulus representations in the stimulated sub-network or the activity patterns of other unstimulated sub-networks.

We hypothesized that the lack of coherent representations in the unstimulated sub-networks was due to the randomness of connections between neurons in the first and the second layer. When top-down gain is provided to a second layer neuron, it sends excitatory signals to the currently attended feature but also sends other signals, excitatory and/or inhibitory, to other first layer neurons in any sub-network that it has connections to. Because this connection is random with regards to the tuned feature of these first layer neurons that get feedback signals, these incidental signals are likely canceled out through destructive interference, leaving intact idiosyncratic modulatory patterns that would rarely develop into coherent representations.

To directly test this account, we modulated how likely a given 'control' neuron will have connections to similarly tuned 'sensory' neurons, introducing more structure to a network with fully random between-layer connections. As we increased the structuredness in our model, spurious signals appeared in a feature-specific manner with stronger top-down modulation. Notably, at an intermediate level of structuredness in the between-layer connections, our model

showed increased activity for the attended feature regardless of presented location, similar to the global spread of feature-based attention (1, 3, 5). In summary, our findings highlight the role of random connections in achieving highly selective top-down attentional modulations relayed by neurons with high-dimensional tuning.

Methods

We used the same setup of the spiking neural network model as in Chapter 2, except the between-layer connections were varied in the degree of randomness as described below.

Modulating randomness of the between-layer connections in the feature space

To test the effect of network randomness, we parametrically modulated the connectivity randomness between the first and second layer. First, for every second layer neuron, we randomly decided how many connections the second layer neuron is going to have to neurons in each sub-network, following the same likelihood of excitatory connection at 0.35. Then, looping through each sub-network, we determined the connections between each second layer neuron and neurons in a given sub-network using a circular normal probability distribution function that was centered on a randomly chosen stimulus feature (μ):

$$p(\theta) = \frac{exp(\kappa\, cos(\theta\ -\ \mu))}{2\pi I_0(\kappa)}$$

The width of the probability distribution was determined by κ, which we modulated between 0 and 0.4 in increments of 0.1 (Figure 3.1). When $\kappa = 0$, the probability distribution was uniform and every neuron in all sub-network had an equal chance of being connected to a given second layer neuron. When $\kappa > 0$, the probability distribution was a bell-shaped curve and the first layer neurons that prefer features closer to the center of the probability distribution had a higher chance of being connected to the given second layer neuron. Note that this process only changed whether a certain between-layer connection existed and not the strength of the connection. As a result, when κ is higher, the second layer neurons will be connected to first

layer neurons that have similar stimulus preferences, which in turn will make the second layer neurons develop a lower-dimensional selectivity for stimulus features.

To test the effect of κ modulation, we simulated feature-based attention and concurrent spatial and feature-based attention as before using networks with varying κ levels. Critically, we compared feature-related signals for the attentional target from the stimulated and unstimulated sub-networks, to test if there is an increase in the signal for the attended feature from even in the sub-network where no stimulus was presented, when there is more structure in the between-layer connectivity. See below for details on the quantification of feature-related signals.

Evaluating the effects of modulation in stimulus strength and top-down attention

Using a similar procedure as in Chapter 2, we averaged the firing rates for each neuron in each of the eight sub-networks across the last 250ms of each trial, to quantify and compare the stimulus- or attention-related signals. Contrast response functions (CRFs) for the Stimulated condition were calculated from averaging firing rates of the two neurons that have the highest preference for the attended stimulus, in the sub-network that the stimuli were presented to. Unstimulated condition CRFs were calculated from averaging firing rates of the two neurons that have the highest preference for the attended stimulus, in the sub-networks the stimuli were not presented to. This was to test whether there is heightened activity in the unstimulated sub-networks that corresponds to the attended stimulus feature with the changes in the randomness of the between-layer connectivity.

Decoding analysis

The same cross-generalization procedure from Chapter 2 (training support vector machines on the *sensory* task to predict the attended stimulus in the *attention* task; see *Decoding*

analysis in *Chapter 2 Methods*) was also applied to assess feature-related representational changes with the randomness of between-layer connections for feature-based attention simulations and concurrent spatial and feature-based attention simulations (Figure 3.2, bottom row; Figure 3.4).

Network Initializations

To ensure that the effects we observe in the simulations are not limited to a specific network initialization, we repeated the feature-based attention simulations with different connectivity randomness levels across ten iterations using different random seeds to initialize each network. This gave rise to different connectivity patterns and weights between the first and the second layer. All plots show averaged data across these ten iterations with the shaded region as the standard error of mean, except for concurrent spatial and feature-based attention simulations, for which we ran one network initialization per each level of connectivity randomness.

Results

Structured connections give rise to feature-specific spurious signals while enhancing the response to the attended stimulus

To test the importance of random connections, we modulated the probability that a given second layer neuron is connected to neurons in different first layer sub-networks that are tuned to a similar feature value. For example, in a highly structured network, if a second layer neuron has a connection with a neuron from one sub-network that was tuned to 180°, that second layer neuron would be more likely to have a connection with a neuron in another sub-network that is also tuned to 180°. In contrast, in a fully random network, as was used in the simulations in Chapter 2, no constraints are placed on the formation of connections between the two layers (except those governing the overall probability of forming connections, see *Chapter 2 Methods*). To implement this procedure, we modulated the connections between the sub-networks and the second layer neurons using a circular normal probability distribution. We adjusted the dispersion parameter κ that determines the spread of the probability distribution along the feature space, that in turn modulates how likely a given second layer neuron will be connected to similarly tuned first layer neurons across the sub-networks (Figure 3.1). This modulation of randomness is expected to change the network properties in two main ways: With higher κ, 1) second layer neurons will have lower-dimensional tuning, as the first layer neurons they get converging inputs from will have similar feature tuning, and 2) stimulus input to one sub-network will likely modulate activity in other sub-networks through the connections with the second layer, in a feature-specific way and not idiosyncratically as before, since the second layer neurons will relay signals to neurons in other sub-networks that are tuned to the similar feature as the stimulus input.

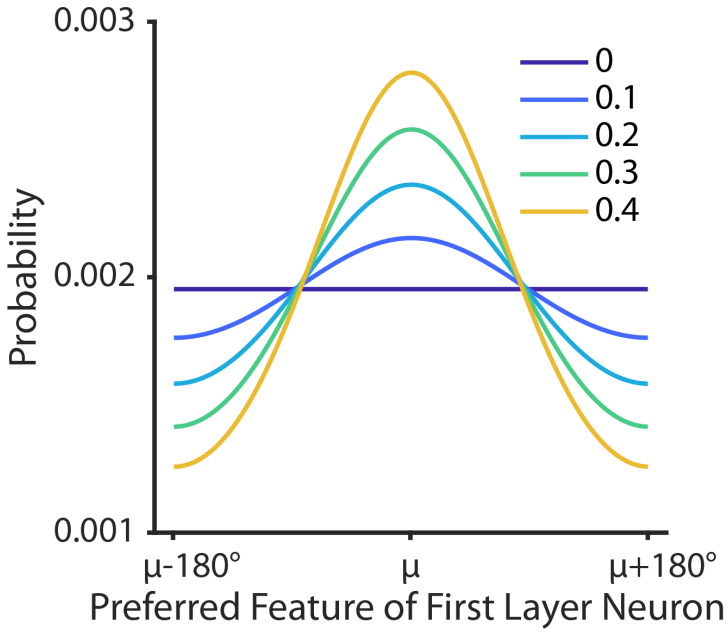


Figure 3.1 Probability distribution of connections between a second layer neuron and first layer neurons as a function of preferred feature of the first layer neurons. For every second layer neuron, the center of the probability distribution (μ) was randomly chosen from the stimulus space. Each line represents a circular normal distribution with different levels of κ, as shown in the legend.

With varying degrees of structure in the connections between the first layer sub-networks and the second layer, we performed feature-based attention simulations using the same *sensory* task and the *attention* task from Chapter 2 (Figure 2.3A, 2.3B). We measured the average firing rates of neurons that maximally prefer the attended stimulus in the stimulated sub-network (Figure 3.2, top row) and in the unstimulated sub-networks (Figure 3.2, middle row) as a function of stimulus strength. We found that the response functions for the stimulated sub-network generally shifted leftward when there was more structure in the between-layer connections (i.e. higher κ). Importantly, in the unstimulated sub-networks, even though no external stimulus input was provided to these networks, the neurons that are tuned to the attentional target started to show an increase in structured activity (i.e. κ = 0.3, 0.4). This effect

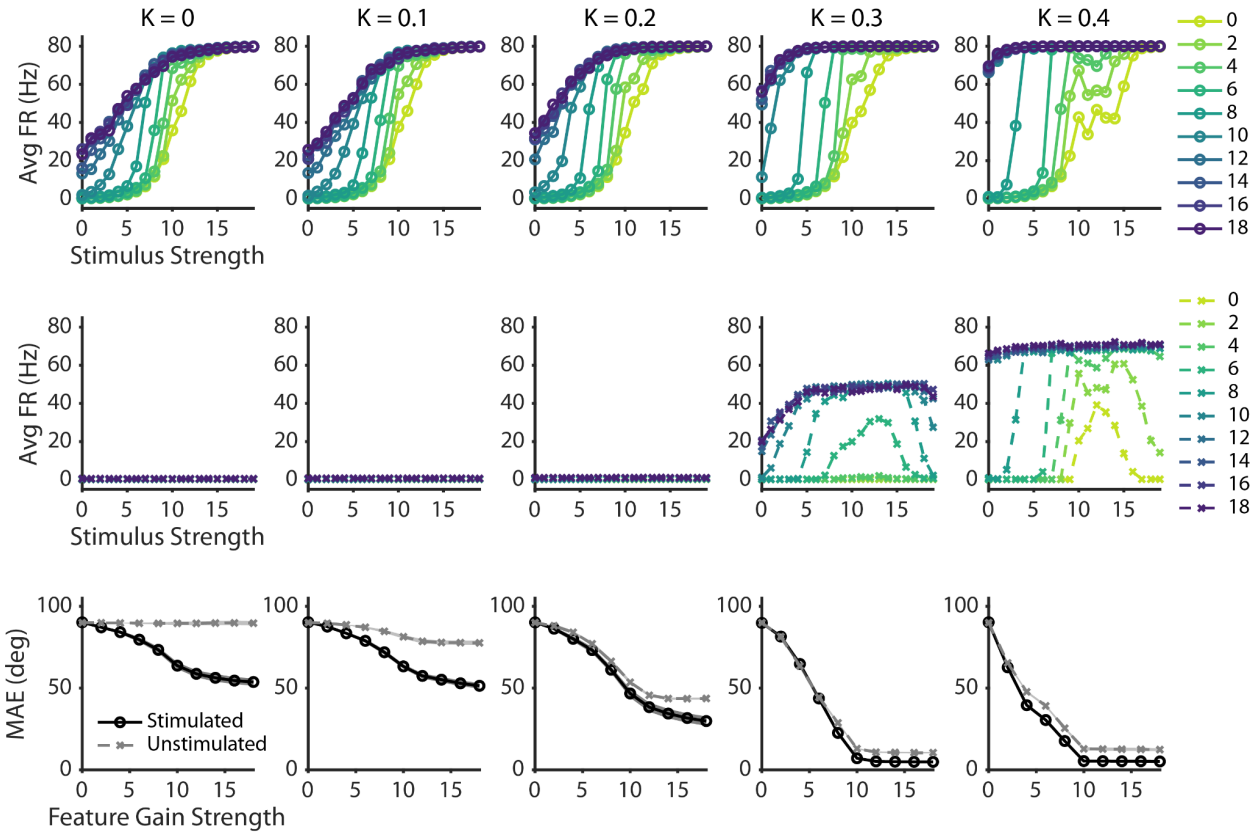


Figure 3.2 Randomness modulations in feature-based attention simulations. Each column represents results from different level of connectivity randomness (κ), as labeled at the top. Top row shows contrast response functions (CRFs) for the attended stimulus in the stimulated sub-network. Middle row shows CRFs in the unstimulated sub-network, based on firing rates of the neurons that prefer the attended feature. Colors of the lines represent different levels of feature-based attention gain, as shown in the legend. Bottom row shows average mean absolute error (MAE) between the predicted and actual stimulus input for Stimulated (solid black line) and Unstimulated (dotted gray line) conditions based on regression model predictions. Shaded areas represent standard error of mean across network initializations.

was most pronounced for higher feature-based attention gain conditions, for example, when κ = 0.3 and the attention gain was 8 or higher. Interestingly, when κ = 0.4, we observed increased activity in the unstimulated sub-networks even when no top-down feature gain was applied. This suggests that due to the high structuredness in the between-layer connections, strong stimulus-driven activity in one sub-network can indirectly increase activity in other sub-networks through

that structured connection without additional top-down amplification of that signal. Moreover, we also found a drop in the response function when stimulus strength was at a higher level and top-down gain at a relatively lower level. This is likely due to the strong input of the competing stimulus having a suppressive effect on the propagation of signals of the attentional target. In other words, the top-down attentional modulation is not strong enough to dominate the propagated signals from the competing stimulus.

Structured between-layer connections lead to stimulus-like activity patterns in unstimulated sub-networks and capture global spread of feature-based attention

To better understand the representational changes in each sub-network and to quantify the resemblance between the attentional modulations and stimulus-driven response patterns, we performed cross-generalization decoding analysis by training the circular ridge regression model on the *sensory* task to predict the attended stimulus in the *attention* task. When the between-layer connections were slightly more structured (i.e. $\kappa = 0.1$), mean absolute error (MAE) for the stimulated sub-network showed a similar pattern to when the connections were fully random, decreasing as a function of top-down attentional gain at a similar magnitude (Figure 3.2, bottom row). Importantly, in the unstimulated sub-networks, there was a slight decrease in MAE with stronger top-down gain, although it was smaller in magnitude compared to the stimulated sub-network. This suggests that when the network has more structure in the connections, the feedback signals can modulate activity in unstimulated sub-networks in a way that resembles bottom-up feature-related signal changes, similar to the empirically reported global spread of feature-based attention (1, 2). As we increased the structuredness further, with $\kappa = 0.2$ and higher, the MAE of stimulated sub-network decreased even more with higher feature gain, suggesting that the activity in the sub-network closely resembles the stimulus-driven activity

pattern, with the MAE value dropping close to 0 when κ = 0.3, 0.4. However, the magnitude of decrease in MAE in the unstimulated sub-networks became more similar to that of the stimulated sub-network. This implies that when the connections were more structured, the feedback-modulated activity patterns in the unstimulated sub-networks resembles the stimulus-driven signals, enough to support decoding of the attended stimulus.

Overall, having more structure in the connections amplified the effect of top-down attention, enhancing representations of the attended stimulus. This is likely due to top-down gain signals that spread to other sub-networks converging on similar feature-tuned neurons, which in turn send signals back to the second layer neurons that receive the attentional gain, and therefore to the stimulated sub-network, strengthening the representation of the attended target. This comes at a cost, because as the feedback signals in unstimulated sub-networks converge rather than cancel each other out, they can form a coherent representation similar to a stimulus-driven activity, which might lead to incorrect readout of stimuli being present in unstimulated sub-networks (e.g. illusory percept). Therefore, in our simulations, randomness in the between-layer connections play a role in canceling out these spurious signals from forming a stable feature-related activity pattern and ensures that only the sub-network presented with the stimulus benefits from the attentional gain, while adding a limited amount of structure enables our network model to capture the global feature-based attention effect as well.

Applying concurrent spatial and feature-based attentional modulation amplifies the effect of randomness modulation

In addition to the feature-based attention simulations, we also investigated the effect of structuredness in connections in simulations of concurrent spatial and feature-based attention,

using the same task as previously shown in Chapter 2 (Figure 2.6A). We measured the average firing rates of neurons that maximally prefer the attended stimulus in the stimulated sub-network and in the unstimulated sub-networks as a function of stimulus strength. We chose $\kappa = 0.1$ and $\kappa = 0.3$ conditions and spatial attention gain of 2 and 6 as examples to plot the response functions in Figure 3.3 (all results shown in Supplementary Figure 3.1). Replicating previous results, the response functions for the stimulated sub-network generally shifted leftward when there was 1) higher feature-based attention gain, 2) higher spatial attention gain, and 3) more structure in the between-layer connections (top rows in Figure 3.3A and 3.3B). In the unstimulated sub-networks, when $\kappa = 0.3$, the neurons tuned to the attended feature showed increased activity when feature-based attention gain was high. This effect was more pronounced when stronger spatial attention gain was applied, showing increased activity for the attended feature even at lower feature-based attention gain levels.

We also conducted decoding analyses for the concurrent spatial and feature-based attention simulations. We used the circular ridge regression model that was trained on the *sensory* task and cross-generalized to the *attention* task to predict the attended feature, separately for each level of κ. As before, we quantified the decoding performance as MAE between the attended feature and the feature predicted by the model and plotted as a function of feature-based attention gain strength, for each spatial attention gain strength and κ value (Figure 3.4). When the between-layer connections were fully random ($\kappa = 0$), we observed a decrease in MAE for the stimulated sub-network with stronger feature-based attention gain while MAE for the unstimulated sub-networks stayed at chance, similar to the previous results. This decrease for stimulated sub-network MAE was more pronounced with stronger spatial attention gain. Next, when the connections were more structured ($\kappa = 0.1$), the stimulated sub-network MAE

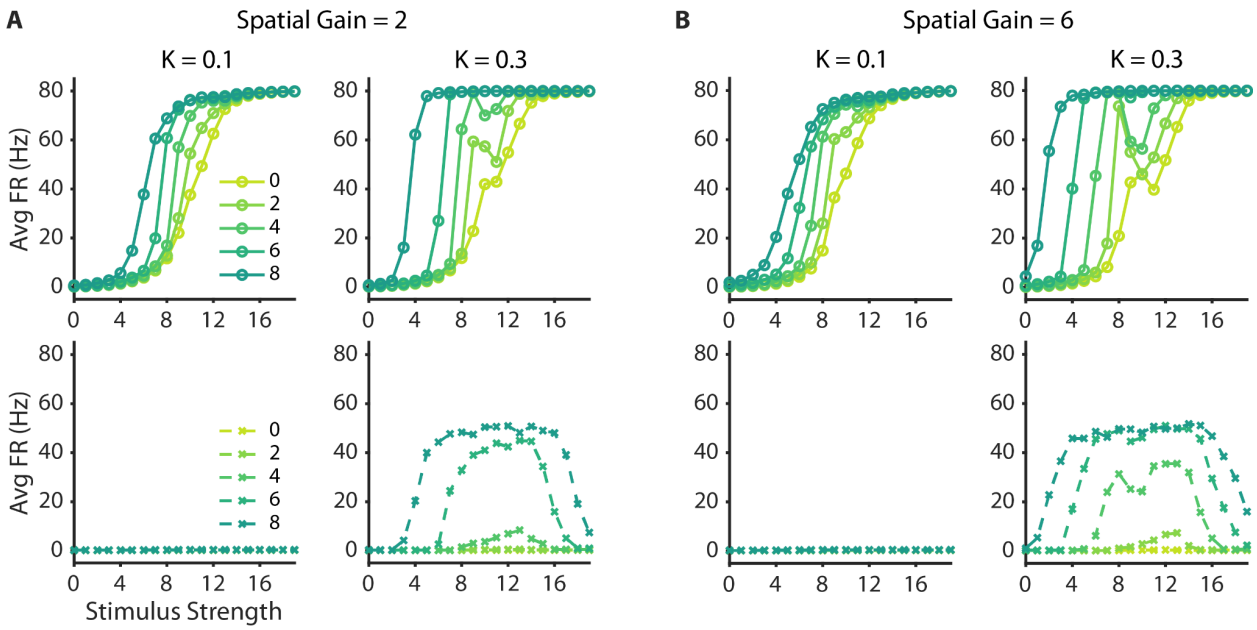


Figure 3.3 Randomness modulations in concurrent spatial and feature-based attention simulations. A, B. CRFs for the attended stimulus in stimulated (top row) and unstimulated (bottom row) sub-networks. Colors of the lines represent different levels of feature-based attention gain, as shown in the legend. Left panels show results for higher randomness condition ($\kappa = 0.1$), and right panels show results for lower randomness condition ($\kappa = 0.3$). Panels in A show results for lower spatial attention gain, and panels in B show results for higher spatial attention gain.

decreased by a larger magnitude with stronger feature-based attention gain, while the unstimulated sub-network MAE still stayed close to chance levels, decreasing slightly towards the strongest feature-based attention gain. This big gap in the MAE between the stimulated and unstimulated sub-networks suggests that while the stimulated sub-network is benefiting from the attentional gain, with its response pattern more closely resembling the sensory-driven responses for the attentional target, the unstimulated sub-networks showed marginal shift in their responses getting similar to coherent feature representations. When the network was even more structured ($\kappa = 0.2$ and higher), MAE for the unstimulated sub-networks decreased at a similar rate as the stimulated sub-network. This suggests that the feedback modulations in the unstimulated

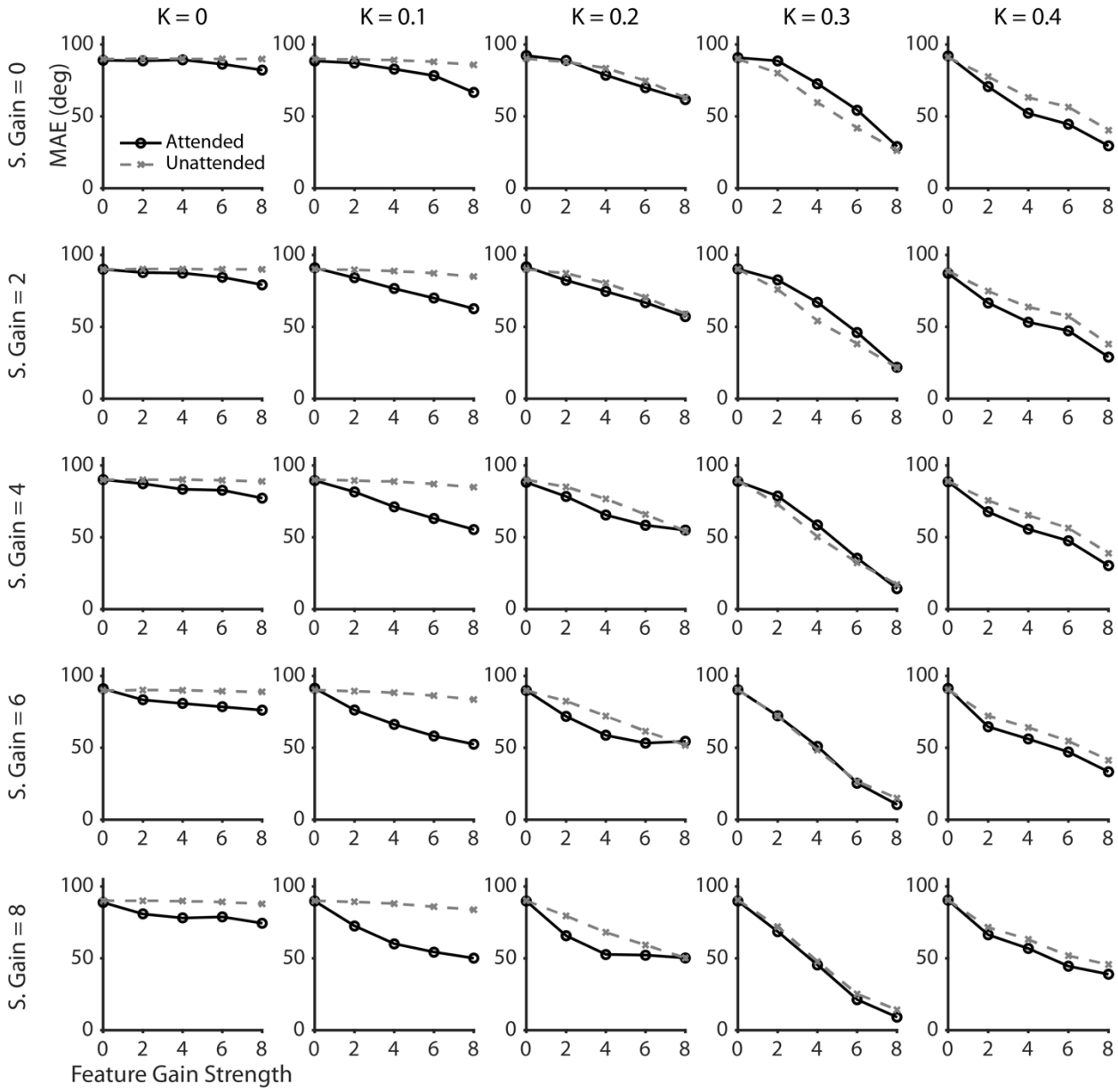


Figure 3.4 MAEs for concurrent spatial and feature-based attention simulations with randomness modulation. MAEs were calculated by training the regression model on the *sensory* task from feature-based attention simulations and testing on the *attention* task, separately for each randomness level (κ). Each column shows results for different level of connectivity randomness, as labeled at the top, and each row shows results for different level of spatial attention gain, as labeled on the left. Solid black lines represent stimulated sub-network and dotted gray lines represent unstimulated sub-network in each panel.

sub-networks were able to form a coherent representation that closely resembled sensory representation just as much as the stimulated sub-network, even in the absence of any stimulus input.

Although not central to our hypothesis, we noticed that MAE values drop to a lower point with higher feature-based attention gain when $\kappa = 0.3$ compared to when $\kappa = 0.4$, contrary to the general trend of higher κ leading to lower MAE. This seems to be due to more structure in the network making it easier to fall into a self-exciting loop that is sustained by the sub-networks indirectly reinforcing activities in each other through the second layer neurons that they converge on for one of the presented features that is hard to break out of unless the feature-based attention gain is high enough. The example trials in Supplementary Figure 3.2 illustrate cases when the to-be-unattended stimulus ends up dominating the activity over the to-be-attended stimulus in the stimulated sub-network shortly after the two stimuli are presented. When $\kappa = 0.3$ (left panel), this is reversed when the feature-based attention gain is applied for the attended stimulus (lower cluster of spikes in sub-network 1), leading to the stimulated sub-network correctly representing the attended feature and the unstimulated sub-networks following that shift in representation. However, when $\kappa = 0.4$ (right panel), the activities for the to-be-unattended stimulus quickly gain strength in both the stimulated sub-network and the unstimulated sub-networks shortly after the stimuli were presented, potentially through indirect interactions that send excitatory signals for the to-be-unattended stimulus and inhibitory signals for the to-be-attended stimulus. Thus, this robust attractor state maintained through structured connections between the first and the second layer neurons makes it harder for the feature-based attention to bias the activity in favor of the attentional target.

In summary, a network with fully random connections is most efficient at relaying targeted space and feature-based feedback signals to sub-networks that code sensory information. However, if some minimal structure is introduced, other empirical phenomena, such as the global spread of feature-based attention can also be recapitulated while a reasonable level of specificity in the feedback modulation is maintained.

Discussion

As pointed out in previous studies, the randomness in connections between the early and later areas in the brain is important for flexibly encoding information in working memory (6). Even though the early sensory areas are more structured and lower-dimensional in their selectivity, random connections allow convergence from these early sensory areas to result in high-dimensional selectivity in the later areas, in turn endowing the ability to encode and integrate diverse information. Building on the idea that randomness supports the flexible encoding and storage of sensory information, our findings highlight the importance of the random projections in the context of top-down attentional modulations. Given the high degree of convergence from early sensory areas to high-order control areas, feedback signals propagated from the control neurons will inevitably spread to many early sensory neurons (7, 8). When the connections are structured, feedback signals can modulate the unstimulated neurons in a way that resembles the response patterns to a sensory input, as well as increasing the response of the attentional target. In contrast, we demonstrate that the randomness of these projections is crucial to selectively enhancing relevant information while canceling out idiosyncratic signals.

One possible mechanism to implement random connections at the cellular level is cell adhesion molecules such as protocadherin that mediate self-avoidance in dendritic connections (9–12). These proteins are expressed in a probabilistic way in different combinations that give individual neurons with random, unique molecular identities. Thus, while speculative, matching the neurons in different brain areas with the same molecular identity may support the initialization of random and reciprocal connections (6, 13).

Based on previous literature and our simulations, maximal randomness in a network is desirable to achieve flexible encoding and the most targeted top-down enhancement of the relevant input signal. However, a fully random network cannot account for the global spread of feature-based attention, in which a neuron's activity is amplified relative to the similarity between the neuron's preferred feature and the attended feature regardless of whether the presented feature is in its spatial receptive field or not (1–5, 14). As shown in Figure 3.2, parametrically changing the structuredness/randomness in the between-layer connections revealed a balance point where the unstimulated sub-networks exhibit activity patterns that are similar to the stimulated sub-network, but at a lower level than the stimulated sub-network. This pattern emulates the spread of feature-based attention to locations where no stimulus was presented that would support increased sensitivity for the attended feature across the entire visual field (1, 2), but still not overwhelm the representation of the physically present feature. Alternatively, similar global feature-based attention effects might arise if each neuron in the second layer is connected to like-tuned neurons in different first-layer sub-networks. In addition, the strength of those connections could vary such that feedback would upregulate like-tuned neurons in one sub-network more than the others. However, this architecture would yield unimodal tuning in both sensory and higher-order areas. This rigid coding scheme is contrary to what is observed empirically, and it would not capitalize on the benefits of multiplexing, such as using high-dimensional codes with a smaller number of neurons to encode diverse inputs and solve complex tasks. Finally, global feature-based attention effects might arise due to local excitatory connections between similarly tuned neurons in sensory areas. This scheme would allow full randomness in the connections between layers while still giving rise to global feature-based attention effects.

In the current study, we simulated top-down space- and feature-based attention by manipulating a single "second" layer of neurons conceptualized as the control area neurons. This simplified model was sufficient to test our hypotheses about the importance of randomness when relaying feedback signals via neurons with high-dimensional tuning. However, it is also possible that a network with two "second" layers - one with structured connections for spatial attention and one with less structured connections for feature-based attention - could also account for spatial and feature-based attention effects. For example, frontal eye field (FEF) neurons have spatial tuning and have been implicated in space-based attentional control (15–19). In contrast, the inferior frontal junction (IFJ) has been implicated in feature/object-based control in humans (20–25). In addition, a recent study showed that inactivating ventral prearcuate (VPA) in monkeys impaired the monkey's ability to saccade to a target in a search array while the ability to saccade to a single target was intact, suggesting VPA's involvement in guiding feature-based attention, similar to IFJ in humans (26, 27). In turn, signals in VPA may be relayed to FEF to modulate prioritization and orienting to the selected feature in the spatial domain. Thus, more complex interactions between spatial and feature-based attention may be better addressed with a more comprehensive network model with functionally separate spatial and feature-based control layers.

The idiosyncratic signal we found in unstimulated sub-networks was consistent with regards to which feature was the target of attentional enhancement, enough to support the decoding of the attended feature without necessarily resembling the sensory response to this feature. This may explain some previous findings reporting successful decoding of the attended feature from cortical areas with receptive fields in the unstimulated part of the visual field (28). While feature-specific modulations (i.e. the global spread of feature-based attention) could co-

occur with the idiosyncratic feedback signals, unless direct comparison to feature-specific sensory responses is done, it cannot be ruled out that the decoding in the previous studies could have relied on the consistent, but random, signals instead of global feature-specific modulations. More generally, this suggests that interpretation of decoding results such as this needs to be done carefully. The decoding performance will be higher than chance as long as the signals change systematically with the label, regardless of the nature of the signal.

While our study focused on relaying top-down modulations, it still begs the question about how the modulations are initialized in the first place. For top-down modulations of stimulus evoked responses, norepinephrine released from locus coeruleus can interact with stimulus-driven local glutamatergic signaling in areas like parietal cortex or prefrontal cortex (PFC) to enhance high priority representations (29–31). In this "glutamate amplifies noradrenergic effects" (GANE) model, high levels of local glutamate for the stimulus representation can lead to more release of local norepinephrine to form "hotspots" of high norepinephrine activity, maintained through a positive feedback loop. At the same time, the locus coeruleus inhibits representations that do not form these hotspots, by increasing lateral inhibition where the hotspots form and suppressing activities where there are no hotspots, resulting in a widespread suppression while selectively enhancing high-priority representations. This model provides a viable alternative explanation for how top-down modulations can be initialized without the need to hypothesize a hierarchical series of control areas. However, the mechanism that mediates the initial prioritization of a relevant location or feature to initiate the "GANE" process is still unclear.

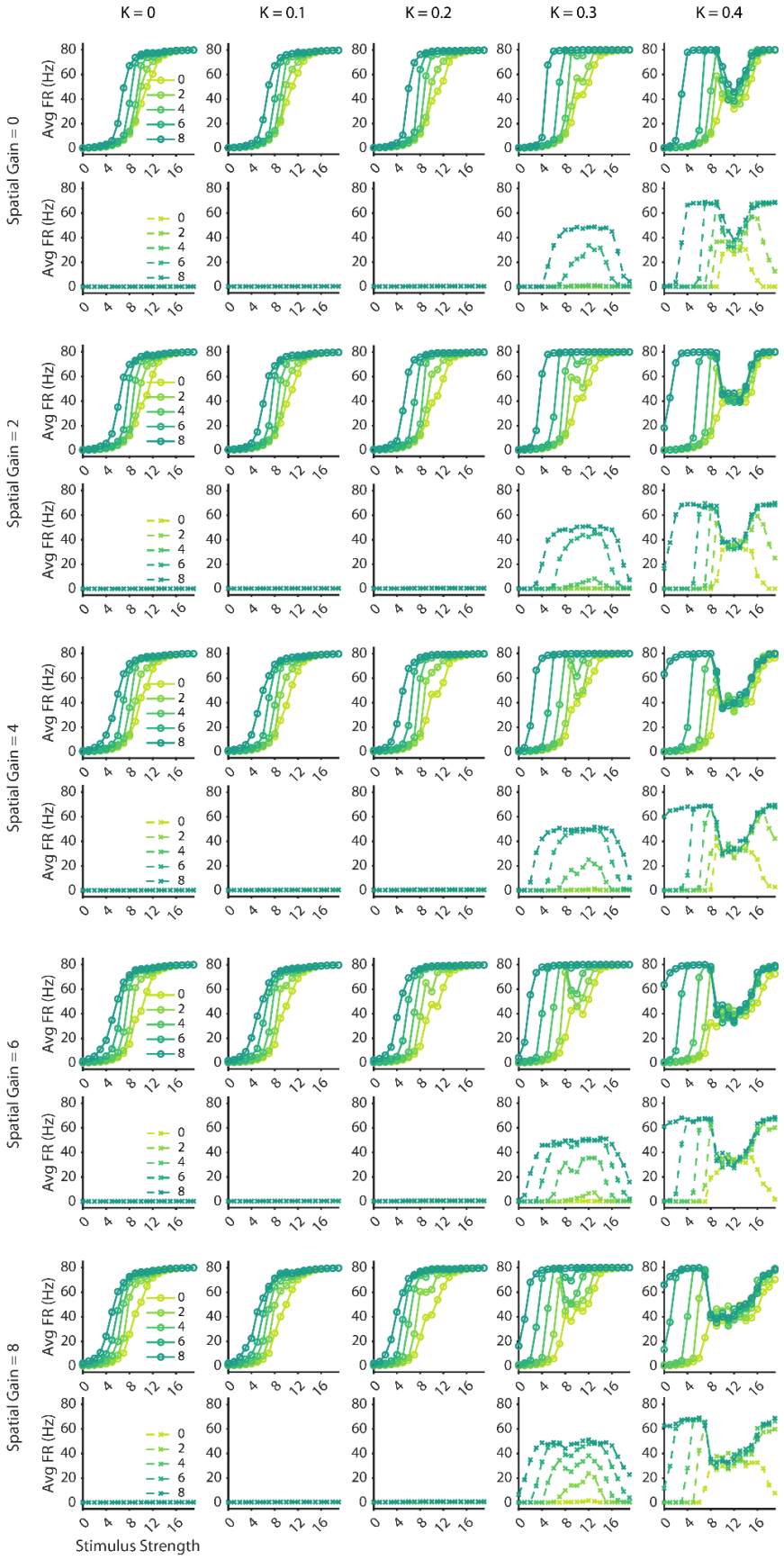
In conclusion, our findings suggest that random connections provide a mechanism to relay highly selective top-down attentional modulations in a system with high-dimensional

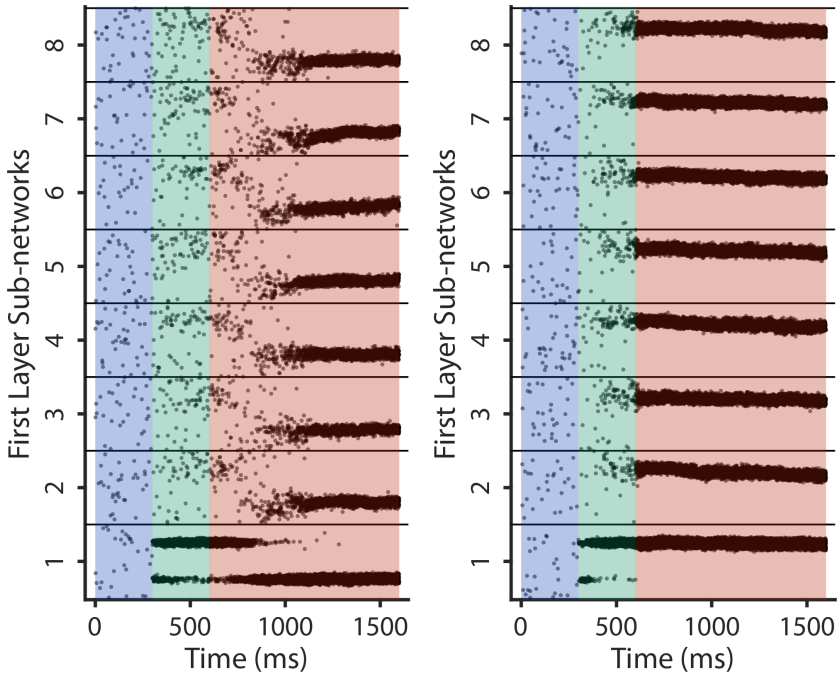
tuning functions. When the connections are sufficiently random, only relevant signals are maintained in a coherent, stimulus-like, pattern. In contrast, feedback signals related to other features that are encoded by higher-order neurons largely cancel out and do not form representations that might be confusable with the attended stimulus. This emphasizes the importance of randomness in the connectivity in a hierarchical network with converging inputs for precise feedback modulations.

Acknowledgements

Chapter 3, in part, is currently being prepared for submission for publication of the material. Park, Sunyoung; Serences, John. T. The dissertation author was the primary researcher and author of this material.

Supplementary Figure 3.1 CRFs for concurrent spatial and feature attention simulations for all levels of κ and spatial attention gain. For each level of spatial attention gain, labeled on the left side, solid lines represent CRFs for the attended stimulus in stimulated sub-networks (top row) and dotted lines represent CRFs for unstimulated (bottom row) sub-networks. Colors of the lines represent different levels of feature-based attention gain, as shown in the legend.


K = 0
K = 0.1
K = 0.2
K = 0.3
K = 0.4
Spatial Gain = 0
Spatial Gain = 2
Spatial Gain = 4
Spatial Gain = 6
Spatial Gain = 8
Avg FR (Hz)
0
2
4
6
8
0 4 8 12 16
Stimulus Strength



Supplementary Figure 3.2 Comparison of raster plots for concurrent spatial and feature attention simulations. Left panel shows an example trial for when $\kappa = 0.3$ and right panel is for when $\kappa = 0.4$. In both trials, the spatial and feature-based attention gain strength was 8, stimulus strength was 11, and the attended stimulus was 90° (the lower cluster of spikes within sub-network 1).

References

1. **Treue S**, **Trujillo JCM**. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399: 575–579, 1999. doi: 10.1038/21176.

2. **Martinez-Trujillo JC**, **Treue S**. Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. *Curr Biol* 14: 744–751, 2004. doi: 10.1016/j.cub.2004.04.028.

3. **Sàenz M**, **Buracas GT**, **Boynton GM**. Global effects of feature-based attention in human visual cortex. *Nat Neurosci* 5: 631–632, 2002. doi: 10.1038/nn876.

4. **Sàenz M**, **Buraĉas GT**, **Boynton GM**. Global feature-based attention for motion and color. *Vision Res* 43: 629–637, 2003. doi: 10.1016/S0042-6989(02)00595-3.

5. **Bichot NP**, **Rossi AF**, **Desimone R**. Parallel and Serial Neural Mechanisms for Visual Search in Macaque Area V4. *Science* 308: 529–534, 2005. doi: 10.1126/science.1109676.

6. **Buschman TJ**. Balancing Flexibility and Interference in Working Memory. *Annu Rev Vis Sci* 7: 367–388, 2021. doi: 10.1146/annurev-vision-100419-104831.

7. **Pandya DN**, **Yeterian EH**. Prefrontal cortex in relation to other cortical areas in rhesus monkey: architecture and connections. *Prog Brain Res* 85: 63–94, 1990. doi: 10.1016/s0079-6123(08)62676-x.

8. **Miller EK**, **Cohen JD**. An Integrative Theory of Prefrontal Cortex Function. *Annu Rev Neurosci* 24: 167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167.

9. **Zipursky SL**, **Sanes JR**. Chemoaffinity Revisited: Dscams, Protocadherins, and Neural Circuit Assembly. *Cell* 143: 343–353, 2010. doi: 10.1016/j.cell.2010.10.009.

10. **Hirayama T**, **Yagi T**. Clustered protocadherins and neuronal diversity. *Prog Mol Biol Transl Sci* 116: 145–167, 2013. doi: 10.1016/B978-0-12-394311-8.00007-8.

11. **Kostadinov D**, **Sanes JR**. Protocadherin-dependent dendritic self-avoidance regulates neural connectivity and circuit function. *eLife* 4: e08964, 2015. doi: 10.7554/eLife.08964.

12. **Sanes JR**, **Zipursky SL**. Synaptic Specificity, Recognition Molecules, and Assembly of Neural Circuits. *Cell* 181: 536–556, 2020. doi: 10.1016/j.cell.2020.04.008.

13. **Bouchacourt F**, **Buschman TJ**. A Flexible Model of Working Memory. *Neuron* 103: 147-160.e8, 2019. doi: 10.1016/j.neuron.2019.04.020.

14. **Liu T**, **Mance I**. Constant spread of feature-based attention across the visual field. *Vision Res* 51: 26–33, 2011. doi: 10.1016/j.visres.2010.09.023.

15. **Moore T**, **Fallah M**. Microstimulation of the frontal eye field and its effects on covert spatial attention. *J Neurophysiol* 91: 152–162, 2004. doi: 10.1152/jn.00741.2002.

16. **Moore T**, **Armstrong KM**. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421: 370–373, 2003. doi: 10.1038/nature01341.

17. **Gregoriou GG**, **Gotts SJ**, **Desimone R**. Cell-type-specific synchronization of neural activity in FEF with V4 during attention. *Neuron* 73: 581–594, 2012. doi: 10.1016/j.neuron.2011.12.019.

18. **Veniero D**, **Gross J**, **Morand S**, **Duecker F**, **Sack AT**, **Thut G**. Top-down control of visual cortex by the frontal eye fields through oscillatory realignment. *Nat Commun* 12: 1757, 2021. doi: 10.1038/s41467-021-21979-7.

19. **Ekstrom LB**, **Roelfsema PR**, **Arsenault JT**, **Bonmassar G**, **Vanduffel W**. Bottom-Up Dependent Gating of Frontal Signals in Early Visual Cortex. *Science* 321: 414–417, 2008. doi: 10.1126/science.1153276.

20. **Zanto TP**, **Rubens MT**, **Bollinger J**, **Gazzaley A**. Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage* 53: 736–745, 2010. doi: 10.1016/j.neuroimage.2010.06.012.

21. **Zanto TP**, **Rubens MT**, **Thangavel A**, **Gazzaley A**. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat Neurosci* 14: 656–661, 2011. doi: 10.1038/nn.2773.

22. **Baldauf D**, **Desimone R**. Neural Mechanisms of Object-Based Attention. *Science* 344: 424–427, 2014. doi: 10.1126/science.1247003.

23. **Gong M**, **Liu T**. Biased Neural Representation of Feature-Based Attention in the Human Frontoparietal Network. *J Neurosci* 40: 8386–8395, 2020. doi: 10.1523/JNEUROSCI.0690-20.2020.

24. **Bedini M**, **Baldauf D**. Structure, function and connectivity fingerprints of the frontal eye field versus the inferior frontal junction: A comprehensive comparison. *Eur J Neurosci* 54: 5462–5506, 2021. doi: 10.1111/ejn.15393.

25. **Meyyappan S**, **Rajan A**, **Mangun GR**, **Ding M**. Role of Inferior Frontal Junction (IFJ) in the Control of Feature versus Spatial Attention. *J Neurosci* 41: 8065–8074, 2021. doi: 10.1523/JNEUROSCI.2883-20.2021.

26. **Bichot NP**, **Heard MT**, **DeGennaro EM**, **Desimone R**. A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron* 88: 832–844, 2015. doi: 10.1016/j.neuron.2015.10.001.

27. **Bichot NP**, **Xu R**, **Ghadooshahy A**, **Williams ML**, **Desimone R**. The role of prefrontal cortex in the control of feature attention in area V4. *Nat Commun* 10: 5727, 2019. doi: 10.1038/s41467-019-13761-7.

28. **Serences JT**, **Boynton GM**. Feature-Based Attentional Modulations in the Absence of Direct Visual Stimulation. *Neuron* 55: 301–312, 2007. doi: 10.1016/j.neuron.2007.06.015.

29. **Mather M**, **Clewett D**, **Sakaki M**, **Harley CW**. Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behav Brain Sci* 39: e200, 2016. doi: 10.1017/S0140525X15000667.

30. **Clewett DV**, **Huang R**, **Velasco R**, **Lee T-H**, **Mather M**. Locus Coeruleus Activity Strengthens Prioritized Memories Under Arousal. *J Neurosci* 38: 1558–1574, 2018. doi: 10.1523/JNEUROSCI.2097-17.2017.

31. **Luccini E**, **Musante V**, **Neri E**, **Brambilla Bas M**, **Severi P**, **Raiteri M**, **Pittaluga A**. Functional interactions between presynaptic NMDA receptors and metabotropic glutamate receptors co-expressed on rat and human noradrenergic terminals. *Br J Pharmacol* 151: 1087–1094, 2007. doi: 10.1038/sj.bjp.0707280.