Article

Neuron

Coupling between One-Dimensional Networks Reconciles Conflicting Dynamics in LIP and Reveals Its Recurrent Circuitry

Highlights

- Slow dynamics of LIP local population can appear one-dimensional or high-dimensional
- Model of coupling between local networks reconciles conflicting data
- Reveals LIP's internal, recurrent circuitry underlying surround suppression
- Data show two-dimensional slow dynamics as predicted by model

Authors

Wujie Zhang, Annegret L. Falkner, B. Suresh Krishna, Michael E. Goldberg, Kenneth D. Miller

Correspondence

zhang_wujie@hotmail.com

In Brief

Zhang et al. found that slow LIP network dynamics can appear inconsistent with confinement to a single multi-neuronal dimension, unlike previous observations. Their model reconciles the conflict and reveals the circuitry underlying surround suppression.

Zhang et al., 2017, Neuron *93*, 1–14
January 4, 2017
http://dx.doi.org/10.1016/j.neuron.2016.11.023

CellPress

Please cite this article in press as: Zhang et al., Coupling between One-Dimensional Networks Reconciles Conflicting Dynamics in LIP and Reveals Its Recurrent Circuitry, Neuron (2016), http://dx.doi.org/10.1016/j.neuron.2016.11.023

Neuron
Article

Coupling between One-Dimensional Networks Reconciles Conflicting Dynamics in LIP and Reveals Its Recurrent Circuitry

Wujie Zhang,[1,2,3,11,12,*] Annegret L. Falkner,[1,3,6,8,9] B. Suresh Krishna,[3,6,8,10] Michael E. Goldberg,[3,5,6,7,8] and Kenneth D. Miller[2,3,4,5]
[1]Doctoral Program in Neurobiology and Behavior
[2]Center for Theoretical Neuroscience
[3]Department of Neuroscience
[4]Swartz Program in Theoretical Neuroscience
[5]Kavli Institute for Brain Science
[6]Mahoney Center for Brain and Behavior
[7]Departments of Neurology, Psychiatry, and Ophthalmology
College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA
[8]New York State Psychiatric Institute, New York, NY 10032, USA
[9]Institute of Neuroscience, New York University School of Medicine, New York, NY 10016, USA
[10]Cognitive Neuroscience Laboratory, German Primate Center — Leibniz Institute for Primate Research, 37077 Goettingen, Germany
[11]Present address: University of California, Berkley, 1951 Oxford Street, Berkeley, CA 94720, USA
[12]Lead Contact
*Correspondence: zhang_wujie@hotmail.com
http://dx.doi.org/10.1016/j.neuron.2016.11.023

SUMMARY

Little is known about the internal circuitry of the primate lateral intraparietal area (LIP). During two versions of a delayed-saccade task, we found radically different network dynamics beneath similar population average firing patterns. When neurons are not influenced by stimuli outside their receptive fields (RFs), dynamics of the high-dimensional LIP network during slowly varying activity lie predominantly in one multi-neuronal dimension, as described previously. However, when activity is suppressed by stimuli outside the RF, slow LIP dynamics markedly deviate from a single dimension. The conflicting results can be reconciled if two LIP local networks, each underlying an RF location and dominated by a single multi-neuronal activity pattern, are suppressively coupled to each other. These results demonstrate the low dimensionality of slow LIP local dynamics, and suggest that LIP local networks encoding the attentional and movement priority of competing visual locations actively suppress one another.

INTRODUCTION

It has become increasingly appreciated that neural functions need to be understood in terms of neuronal populations and the dynamics of the circuits to which they belong (Miller and Wilson, 2008; Shenoy et al., 2013). However, the field of systems neuroscience in nonhuman primates has traditionally been dominated by studies of the properties of single neurons. While we have a wealth of knowledge of single-neuron behaviors in many areas of the primate brain, this knowledge remains largely phenomenological—we know *what* neurons do, but not *how* they do it. Especially on the circuit level, the mechanisms and connectivity underlying neuronal behaviors are often obscure.

Such is the case in the lateral intraparietal area (LIP), where a large body of literature has revealed that the activity of single neurons encodes visual attention and saccadic eye movements, as well as decision making variables, abstract categories, and other cognitive variables (Bisley and Goldberg, 2010; Freedman and Assad, 2011; Gold and Shadlen, 2007; Kable and Glimcher, 2009). However, little is known about the circuitry inside or outside the LIP network that produces such activity, and therefore the role of LIP in many of these functions is controversial. A step in understanding this circuitry was taken by Ganguli et al. (2008), who analyzed LIP network dynamics during two different tasks: a delayed saccade task (Bisley and Goldberg, 2003, 2006) and a random-dot motion discrimination task (Roitman and Shadlen, 2002). They found that the dynamics of the high-dimensional LIP network are dominated by one multi-neuronal dimension on slow timescales, which could be explained by a simple circuit model. This one-dimensionality was key to explaining an unexpected correspondence between LIP single-neuron responses and the timing of attentional shifts (examined in more detail below). More recently, Fitzgerald et al. (2013) found further evidence for one-dimensional dynamics in three experiments in which LIP encoded learned associations between visual stimuli.

Using a delayed-saccade task similar to the task of Bisley and Goldberg (2003, 2006) (hereafter BG), Falkner, Krishna et al., 2010 (hereafter FK) reported "surround suppression" in LIP (see also Louie et al., 2011), i.e., stimuli outside the receptive field

(RF) of a cell suppress the cell's activity. In the FK study, the population-averaged activity over time is very similar to that in the BG study, as expected given the very similar tasks. However, we find that the pattern of activity across neurons changes over time in a very different way in the FK study. In particular, the network dynamics in the FK dataset markedly deviate from the one-dimensional dynamics observed in the BG dataset, calling into question the validity of the one-dimensional LIP model of Ganguli et al. We show that the two sets of conflicting results can be reconciled and well characterized by a more general low-dimensional model: each of two local LIP networks in isolation has its own single dominant dimension, and suppressive coupling between them gives rise to two dominant dimensions. We further show that the FK data directly confirm the two-dimensional dynamics predicted by the model. Our study thus represents a step forward in discovering circuit mechanisms and connectivity from single-neuron recordings, and in understanding mechanisms behind LIP functions.

RESULTS

One-Dimensional Dynamics in LIP

We begin by describing the first of the two conflicting datasets (Bisley and Goldberg, 2003), along with the one-dimensional model (Ganguli et al., 2008) to which it gave rise.

The delayed-saccade task of BG is illustrated in Figure 1A (details in Supplemental Information [SI] section 1, available online). During this task, LIP neurons exhibit a large transient visual response to the onset of a saccade target or distractor in the RF, and sustained delay period activity (delay activity) when a saccade is planned to the RF (Figure 1C). When a distractor is flashed away from the target location during the delay period, attention is transiently attracted away from the target location to the distractor location. At the same time, the average visual response level of LIP neurons whose RFs contain the distractor location (the distractor population) rises above the average delay activity level of neurons whose RFs contain the target location (the target population). As the visual activity of the distractor population decays back to baseline, the locus of attention shifts back to the target location. This shift in attention coincides with the shift in the peak of LIP activity from the distractor population to the target population: when the decaying visual activity of the distractor population drops to a level statistically indistinguishable from the sustained delay activity of the target population (the "crossing time"—when the decaying red trace crosses the blue trace in Figure 1C), neither the target nor the distractor location has attentional advantage, whereas 100–250 ms before or after this crossing time, the distractor or target location, respectively, is the clear locus of attention.

Further analyses of these results (Bisley and Goldberg, 2006) revealed that this correspondence between activity crossing and attentional switching also held at the level of single LIP neurons. The crossing time of a single neuron is defined as the time at which the neuron's decaying response to a distractor, on trials in which a distractor is in its RF (distractor trials), crosses its own level of delay activity on trials in which a target is in its RF (target trials). These single-neuron crossing times are surprisingly invariant across neurons and closely aligned with the monkey's attentional switching time, despite high variability across neurons in their peak visual responses, time constants of visual response decay, and delay period responses.

Ganguli et al. (2008) explained this observation with the proposal that the dynamics of a local network (LN) of LIP neurons are dominated on slow timescales by one multi-neuronal activity pattern (i.e., a pattern, or vector, of relative firing rates across the cells of the network). Throughout this paper, we use the term "local network" (or LN) to mean a network of LIP neurons that share the same RF (explained more fully in the section "Simple model of coupled local networks reconciles the results"). Ganguli et al. proposed that the recurrent connectivity of an LN causes certain multi-neuronal activity patterns to persist longer in the absence of input; given steady input, these slowly decaying patterns also build up to be strongly amplified. If the network has only a single pattern that decays slowly, we refer to it as the network's "slow mode," where "mode" is a term borrowed from physics that describes a characteristic pattern of a system's response. As the visual response to a distractor decays, it becomes dominated by this slow mode after all other patterns decay away. Because the slow mode is more strongly amplified than other patterns, it also dominates steady-state responses, such as delay activity and activity during the initial fixation before target onset (fixation activity). Thus, after the other patterns in the distractor response decay away, the decaying distractor activity and the ongoing delay activity are both dominated by the slow mode, meaning that the pattern of distractor activity across neurons is very nearly a scaled-up version of the delay activity pattern. Then as the distractor activity decays further, it becomes very nearly identical to the delay activity pattern, which happens at the crossing time. Thus, each individual neuron has roughly the same activity in its delay response as in its distractor response at the crossing time, so that all neurons have about the same single-neuron crossing time.

This one-dimensional model predicts that multi-neuronal activity patterns that change on slow timescales are all highly correlated with one another because all are dominated by the same strongly amplified pattern. These include fixation and delay activity patterns and, to a lesser extent, slowly decaying visual activity patterns and slowly increasing activity patterns during decision-making tasks. On the other hand, during the initial transient visual response, many other activity patterns are activated, so the transient visual activity pattern is not highly correlated with the steady-state activity patterns. Ganguli et al. (2008) confirmed these predictions using the following analysis, which reveals network dynamics from the activity of a population of singly recorded neurons. At any millisecond time point t, we represent the trial-averaged activity of a population of N neurons as an N-dimensional vector, $\vec{r}(t)$, in an N-dimensional multi-neuronal firing rate space; each of the N elements of $\vec{r}(t)$ is the activity of one neuron at time t, averaged over trials. We also compute the N-dimensional fixation activity vector, $\vec{F}$, where each element is the activity of one neuron averaged over the fixation period before target onset and over target trials. Then, at each time point t over the course of the trial, a correlation coefficient is computed between $\vec{F}$ and $\vec{r}(t)$. Figure 1E shows that the correlation to fixation activity is indeed high for

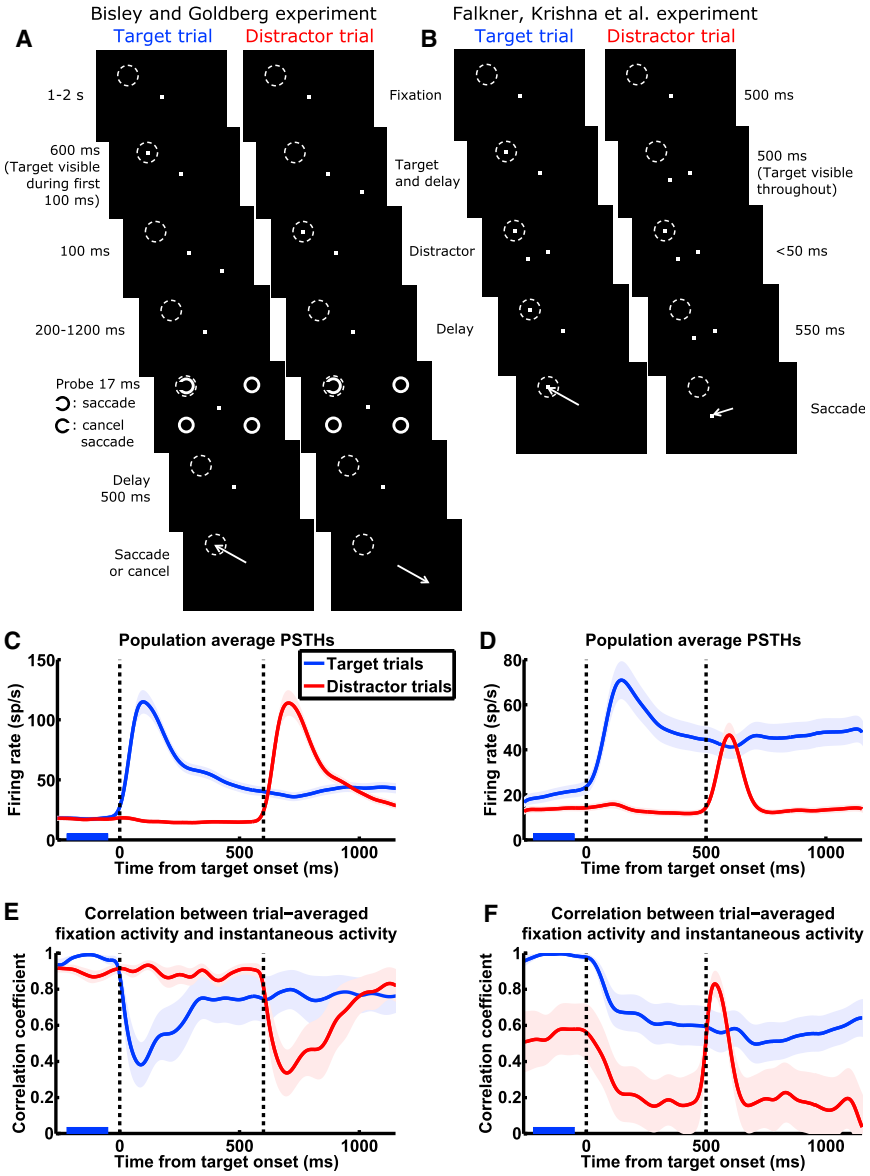

Figure 1. The Conflicting Population Dynamics Observed by Bisley and Goldberg and Falkner, Krishna et al.

(A, C, and E) Bisley and Goldberg (BG).

(B, D, and F) Falkner, Krishna et al. (FK).

(A and B) Task schematics. While the monkey fixates on a central spot, a target appears. The monkey holds fixation until the disappearance of the fixation spot, at which time it makes a saccade to the location of the target (except on "no-go" trials in BG; see below). During the delay between target onset and fixation spot disappearance, a task-irrelevant distractor stimulus is flashed. We call a given trial a target trial or distractor trial when the target or distractor, respectively, is in the RF (dashed circles) of the neuron being recorded. In the BG task, the target and distractor are in opposite visual quadrants and equidistant from the fixation spot; in the FK task, either the target or the distractor is in the RF, and the other stimulus is at the location that elicits maximum surround suppression for the recorded neuron. In the BG task, between 200 and 1,200 ms after the distractor disappears, a probe (a Landolt ring) is flashed at either the target or the distractor location, along with three complete rings elsewhere; a left-facing or right-facing ring instructs the monkey to proceed with the planned saccade ("go" trial) or cancel it and maintain fixation ("no-go" trial), respectively. In (C) and (E), we only include trials in which the probe appeared at least 700 ms after distractor onset. For task details, see SI section 1.

(C and D) Population average peristimulus time histograms (PSTHs) in the BG (C; n = 41 cells) and FK (D; n = 27) studies. Blue/red traces denote trials in which the target/distractor appears in the RF of the neuron being recorded; every neuron was recorded during both target and distractor trials and contributes to both traces. The first and second vertical dashed lines denote the onset of the target and the distractor, respectively. Shading around traces indicates SEM. PSTHs have been smoothed by convolution with a Gaussian kernel (σ = 30 ms; firing rates and correlations appearing to change before stimuli onset in C–F are artifacts of this smoothing).

(E and F) Correlation analysis for the BG (E) and FK (F) datasets. We define a trial-averaged population fixation activity vector $\vec{F}$, each element of which is the activity of one cell on target trials, averaged over trials and over the period from 220 to 50 ms before target onset (marked by blue bars in C–F). At each millisecond time point over the course of the target trial (blue traces) or distractor trial (red traces), the correlation coefficient was computed between the trial-averaged population instantaneous activity vector at that point in time and $\vec{F}$. The BG correlation patterns (E; presented in similar format in Ganguli et al., 2008) exhibit one-dimensional dynamics on slow timescales (high correlations during stable fixation activity and delay activity), while the FK correlation patterns (F) markedly deviate from one dimension (on distractor trials, low correlations during stable activity, and transient increase in correlation during distractor visual response). Vertical dashed lines are as in (C) and (D). Shading around traces indicates SE estimated from 1,000 bootstrap samples.

See Figure S1A for correlations between distractor trial fixation activity and instantaneous activity for the FK data, and Figures S2A–S2D for the FK data plotted separately for different reward conditions.

delay activity or distractor activity after the transient visual response decays away, indicating that fixation, delay, and post-transient distractor activity patterns all lie roughly in a single dimension, corresponding to the dominant activity pattern. The drop in correlation coefficient during the visual response indicates the transient deviation of activity from this one dimension caused by the transient activation of other non-dominant patterns.

Surround Suppression and Violations of One-Dimensional Dynamics

We continue by describing the second of the two conflicting datasets (Falkner, Krishna et al., 2010) and how it exhibits large deviations on both fast and slow timescales from the predictions of the one-dimensional model.

The task of FK (Figure 1B) is very similar to that of BG. For both tasks, we analyze data in each trial during time windows ending

shortly after distractor onset (i.e., before the onset of the probe in the BG task; see Figure 1A), up to which point the two tasks are virtually identical aside from three differences. First, BG used a flashed target while FK presented a target that stayed visible during the delay. This does not result in qualitatively different delay activity levels (compare delay activity between Figures 1C and 1D), consistent with LIP encoding the attentional and saccadic priority of the target location regardless of the visibility of the target. Second, BG randomly interleaved target trials and distractor trials, while FK presented target and distractor trials in blocks. Thus, in the FK experiment, on almost every trial the monkey had an expectation of where the target and distractor would be. This is reflected in higher anticipatory firing on target trials compared to distractor trials during the fixation period before target onset. The third difference is likely to be the key difference that led to different neural responses observed during the two tasks. In the BG task, the target and distractor are in opposite visual quadrants and equidistant from the fixation spot. In the FK task, in contrast, either the target or the distractor is in the RF of the cell being recorded in a given session, and the other stimulus is at the location eliciting maximum surround suppression of the recorded neuron. With this placement of stimuli, a saccade plan to the surround significantly suppressed the visual response to the distractor, while distractor appearance in the surround transiently and weakly, but significantly, suppressed delay activity during saccade planning (Figure 1D; quantified in Falkner, Krishna et al., 2010). Surround suppression was not observed in the BG dataset (quantified in Bisley and Goldberg, 2006), in which the stimulus locations were not selected for suppression. Other than the surround suppression of response amplitudes, the FK dataset displays the same overall pattern of fixation, visual, and delay activity as the BG dataset (compare Figures 1C and 1D).

However, beneath this apparent similarity in population average activity, the network dynamics are radically different; moreover, the FK dynamics appear to clearly violate the predictions of the one-dimensional model. Figure 1F shows the result of the correlation analysis on the FK data. Most strikingly, on distractor trials (red trace), even though the appearance of the target in the surround only minimally affects the mean firing rate of the population, target appearance causes a large, sustained drop in correlation, when the one-dimensional model would predict an unchanging and high level of correlation, as in Figure 1E. This indicates that the activity pattern of the population has changed dramatically while its mean firing rate has remained about the same. Furthermore, the later appearance of the distractor in the RF causes a large, transient rise in correlation that subsequently returns to the steady low level present before distractor onset, when the one-dimensional model would predict the opposite change—a large and transient drop in correlation upon distractor onset, as in Figure 1E. In target trials (blue traces), the difference is more subtle, with target onset evoking a small, sustained drop in correlation, similar to the sustained drop in the BG case, but without the initially larger transient decrease.

Note that in the BG dataset, the two trial types are randomly interleaved; thus, the monkey does not know the trial type during the initial fixation, and fixation activities are the same in the two trial types. In the FK dataset, however, fixation activities are different on the two trial types due to the block design. We chose to use the fixation activity on target trials as opposed to distractor trials to calculate correlations because it reveals salient patterns in the network dynamics. Using distractor trial fixation activity is another angle from which to examine the network dynamics that give less informative results, i.e., correlations do not rise and drop saliently over time (Figure S1A).

Thus, the results of BG and of FK seem incompatible. The robust one-dimensional dynamics observed in the BG data require that the local anatomical connectivity of LIP selectively amplify only one multi-neuronal activity pattern. How can this same anatomical connectivity realize dynamics that deviate so far from the one activity pattern that it so strongly amplifies?

Simple Model of Coupled Local Networks Reconciles the Results

We found the answer in a simple model of the interactions between two coupled LIP LNs. This model replicates the FK findings and yet reduces to the one-dimensional dynamics that characterize the BG findings when the two LNs are not coupled.

We model two LNs in LIP, each composed of excitatory (E) and inhibitory (I) neurons that share an RF, with randomly distributed neuronal time constants (Figures 2A and 2B; see SI section 2.2 for details of the model). Connections between the neurons are sparse, and their strengths are randomly distributed. Within each LN, excitatory connections are, on average, stronger than inhibitory connections. This dominance of excitation is consistent with evidence based on dendritic structure of increased connectivity between excitatory cells in LIP compared to primary sensory cortices (Elston and Rosa, 1997). Such connectivity within an LN, when it's not connected to another LN, amplifies a single pattern, one of increased activity across most cells, more strongly than all other patterns.

The LIP cortical surface contains rough topological maps of visual space (Blatt et al., 1990; Patel et al., 2010). Neurons sharing an RF, which are more likely to be located close to each other on the cortical surface, make up an LN in our model. We model the connections of I cells to be restricted to the LN to which they belong, for inhibitory interneurons generally only make short-range projections, whereas E cells can potentially make long-range projections to the other LN. Since no significant interaction between RFs was observed in the BG dataset (quantified in Bisley and Goldberg, 2006), we infer that for these RFs, the corresponding LNs are not directly connected (Figure 2A). In contrast, by maximizing surround suppression, FK selected for RFs that did interact. Since the interaction observed was predominantly suppressive, it's likely that the excitatory connections from each LN are stronger to the I cells than to the E cells of the other LN. For simplicity, we model the across-network connections as being from the E cells of each LN to the other LN's I cells only, with sparse and random connectivity (Figure 2B). Our results do not change if we include weaker across-network E-to-E connections (data not shown).

We use a standard linear firing rate model (SI section 2.2; Dayan and Abbott, 2005) to simulate the trial-averaged activity in the experiments. We do not explicitly simulate single trials,for we have no knowledge of the single-trial population dynamics

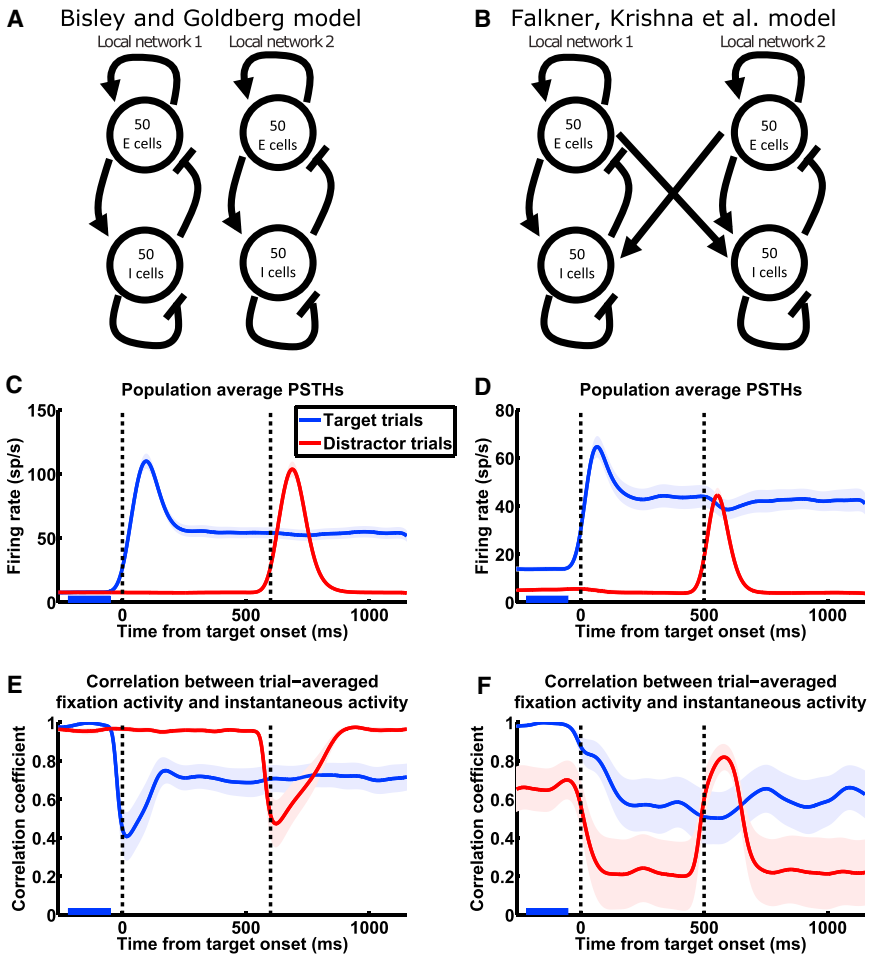

Figure 2. Model Reproduces the Response and Network Dynamics of Bisley and Goldberg and Falkner, Krishna et al.

(A and B) Schematics of the model network connectivity for the BG (A) and the FK (B) scenarios. In both cases, we model two recurrently connected local LIP networks (together referred to as a global network) corresponding to two RF locations, with each LN consisting of E and I cells. Connectivity within each LN is such that each LN by itself amplifies a single multi-neuronal activity pattern much more strongly than other patterns (see Results for details). The FK model network (B) differs from the BG model network (A) in the addition of coupling between the LNs that mediates interactions between them.

(C and D) Model reproduces LIP activity observed by BG (C; n = 41 cells; in all simulation results except where noted, each neuron was "recorded" from a different simulated global network) and FK (D; n = 27). Population average PSTHs with same conventions as Figures 1C and 1D. PSTHs have been smoothed by convolution with a Gaussian kernel (σ = 30 ms; firing rates and correlations appearing to change before stimuli onset in C–F are artifacts of this smoothing).

(E and F) Model reproduces LIP network dynamics observed by BG (E) and FK (F). Correlation analysis with same conventions as Figures 1E and 1F. See Figure S1B for correlations between distractor trial fixation activity and instantaneous activity for the FK simulation, and Figures S2E–S2H for separate simulations of the different reward conditions of the FK experiment.

during the tasks. The experiments involve a variety of sensory, motor, and cognitive processes that likely give rise to a variety of external inputs to LIP during a trial, which we model as the following four types. (1) Fixation input, which is spontaneous firing from the external input sources when there is no stimulus in or saccade plan to the RF, such as during the fixation period. (2) Visual input, which is bottom-up input to an LN when a visual stimulus is in the RF, which is strong upon stimulus onset and becomes weak as the stimulus is sustained. Visual input arrives from areas that could include V2, V3, V3A, V4, middle temporal area (MT), and inferotemporal cortex (Baizer et al., 1991; Blatt et al., 1990; Lewis and Van Essen, 2000). (3) Delay input, which is persistent top-down input to an LN when a saccade is being planned to the RF, arriving from frontal areas such as the frontal eye field (FEF) or dorsolateral prefrontal cortex (dlPFC; Blatt et al., 1990; Stanton et al., 1995; in SI section 3, we discuss other possible mechanisms underlying delay activity and their implications for our model). (4) Expectation input, which is top-down input to one LN during the fixation period before target onset, when the animal is in a block of trials during which the target always appears in the RF of that LN (as in the blocked experiment of FK). Expectation input likely also arrives from frontal areas such as FEF or dlPFC (Coe et al., 2002; Roesch and Olson, 2003). The total external input to the neurons at any time is the sum of one or more of these four types of input. For each of the four types of input, input to each cell is independently drawn from a uniform distribution, with ranges of the distributions chosen to fit experimentally observed neural responses. Thus, it is important that the inputs from different sources are uncorrelated. In addition, the external input contains weak, temporally correlated noise that is independent for different neurons, simply to produce small firing rate fluctuations similar to those seen in the experiments.

In the experiments, different neurons are recorded from different LIP locations and have different RF positions. As in Ganguli et al. (2008), we interpret this to mean that these neurons are situated in different LNs, which share the same set of connectivity, neuronal, and input statistics. To model this, we run the simulation multiple times, each time with a different random instantiation of network connectivity, neuronal time constants, and input patterns, and "record" from a single randomly chosen cell during each simulation. Each simulation includes target and distractor trials for the recorded cell. Figures 2C and 2D show the population peristimulus time histograms (PSTHs) from such simulations of the BG (Figure 2C) and FK (Figure 2D) experiments, which reproduce the experimentally observed firing patterns,

including the observed absence or presence of surround interactions. More significantly, our model reproduces the apparently conflicting network dynamics of the two experiments, as revealed from the correlation analysis: the BG model shows one-dimensional dynamics on slow timescales (Figure 2E), and the FK model shows the same higher-dimensional dynamics as experimentally observed (Figure 2F).

If we compute correlations of instantaneous activity to distractor trial fixation activity, rather than to target trial fixation activity, the model also qualitatively reproduces the experimental results (Figure S1B). Furthermore, modeling higher reward levels as resulting in higher levels of delay input (Leon and Shadlen, 1999; Kennerley and Wallis, 2009), we reproduce the results found when the data of FK, which consist of trials with large or small reward, are analyzed separately by reward level (Figure S2). Because the activity and correlation patterns are qualitatively similar across reward levels in the data (Figures S2A–S2D), in all other simulations we simply modeled the average reward level.

In addition to the correlation analysis, another way to examine network dynamics is to calculate, for each instantaneous activity vector, the norms of its component parallel to $\vec{F}$ and its component orthogonal to $\vec{F}$. This reveals a similar picture for the FK data to the correlation analysis, which our model reproduces (Figure S3).

Conceptual Picture: Coupling of Local Slow Modes Explains LIP Dynamics

We fully analyze the behaviors of the model in the next sections, but first, in this section, we presage those results by presenting a simplified conceptual understanding.

Each LN has its own single dominant activity pattern (its slow mode), and therefore each on its own would follow one-dimensional dynamics. The circuitry that creates surround suppression causes these two patterns to suppress one another, and this mutual suppression in turn qualitatively explains the FK correlation patterns, as follows.

Suppose we are recording in one of the LNs, call it LN1, and let the other LN be LN2. $\vec{F}$, the fixation activity of LN1 on target trials, is dominated by LN1's slow mode because it is driven by both fixation input and expectation input. At any given time, the correlation of LN1's instantaneous activity with $\vec{F}$ is high or low according to whether or not that instantaneous activity is dominated by the slow mode. Now consider LN1 on distractor trials. During the initial fixation period, LN1 receives fixation input, but not expectation input. Thus, its slow mode is activated less than on target trials; in addition, its slow mode is suppressed by the more activated slow mode of LN2, which is receiving both fixation and expectation inputs. As a result, the relative contribution of activity patterns other than the slow mode to LN1's activity is larger than on target trials, resulting in reduced correlation between distractor trial fixation activity and $\vec{F}$. After the target appears in LN2's RF, LN1's slow mode continues to be driven only by fixation input; in addition, it is strongly suppressed by the slow mode of LN2, which is strongly driven by both visual stimulation and the subsequent top-down delay input. This greatly reduces the correlation. Finally, when the distractor appears, strong visual stimulation transiently drives up LN1's slow mode as well as other patterns. In BG, this would lower the correlation—LN1 was dominated by the slow mode before distractor onset, and is now less so. In FK, LN1's slow mode was strongly suppressed before distractor onset, and is now driven up to dominate LN1's activity, resulting in the transient rise in correlation.

The conceptual picture just given is simplified in that it describes each LN as having only one dimension of activity that is strongly amplified. In reality, a second strongly amplified dimension is created in each LN by the coupling between the two LNs. When LN2 is the more strongly driven LN, it strongly drives the I cells of LN1. In addition to suppressing LN1's slow mode, this amplifies a pattern of differential firing between the E and I cells in LN1, making the slow mode an even less dominant part of LN1's activity. This will become clear with the detailed analysis below.

Detailed Analysis: Two-Dimensional Dynamics Result from the Coupling of Local Slow Modes

We now take a closer look at the mechanisms of the model. We modeled the BG scenario with two unconnected LNs, each with a slow mode. The model simply behaves like two copies of the one-dimensional model of Ganguli et al., reproducing one-dimensional dynamics and the absence of surround interaction.

The only difference in network architecture in our model of the FK scenario is the presence of connections between the two LNs. Thus, the dominant activity patterns of the two LNs influence each other and are no longer independent. To understand the activity patterns of the global network consisting of the two coupled LNs, we examine the global connectivity matrix, which describes all connections, both within and between the LNs. This connectivity between neurons determines how strongly each neuron excites or inhibits other neurons. The connectivity can equivalently be described as connections between sets of activity patterns, determining how strongly activity in one pattern excites or inhibits activity in itself and in other patterns. For a network composed of separate excitatory and inhibitory neurons, it is often informative to analyze its connectivity as the connections between its Schur activity patterns (Murphy and Miller, 2009; Goldman, 2009; described in more detail in SI section 4). These are an ordered set of orthogonal activity patterns whose connections with each other are as simple as possible for a set of orthogonal patterns: each Schur pattern has a self-connection, and in addition, there is a set of purely feedforward connections between the patterns. We choose to order the patterns by their self-connection strength. Then, activity in pattern 1 (the pattern with the strongest self-excitation) can only influence itself by its self-connection; activity in pattern 2 can excite or inhibit activity in pattern 1, in addition to influencing itself; pattern 3 can excite or inhibit pattern 1, pattern 2, and itself, etc. Thus, given similar external inputs to the patterns, the dominance of any pattern in the network's dynamics can be predicted by the strengths of its self-connection and the feedforward connections it receives. The activity of the network at any moment can be uniquely decomposed as a weighted sum of all Schur patterns. The dominant patterns will generally have weights with the largest absolute values (the weights can be positive or negative).

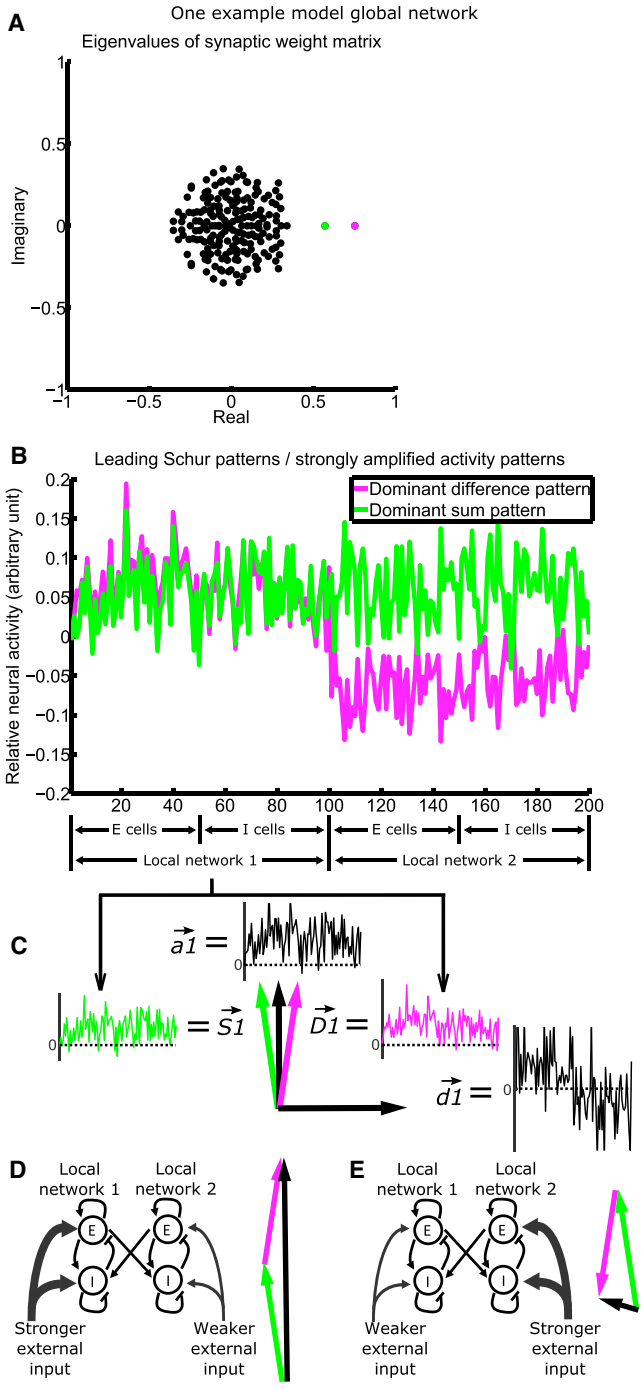

Figure 3. Recurrent Connectivity Strongly Amplifies Two Activity Patterns

(A) The eigenvalue spectrum of the connectivity matrix (matrix plotted in Figure S4D) for a model network composed of two interconnected LNs of 100 neurons each. Each eigenvalue is associated with a Schur vector, representing a pattern of relative activation across neurons (see Results for details). The more positive the real part of an eigenvalue, the more strongly the network amplifies the corresponding Schur activity pattern. Two patterns (magenta and green) are more strongly amplified than others and are plotted in (B).

(B) Relative activation across neurons in the dominant difference pattern (differential activation of the two LNs; magenta) and the dominant sum pattern (equal activation of the two LNs; green), or, equivalently, the two leading Schur vectors of the connectivity matrix. The difference/sum pattern is driven by the difference/sum of the mean inputs to the two LNs. Note the similarity of the two patterns across cells of the same LN.

(C) The LN1 portion of the sum ($\overrightarrow{S1}$) and difference ($\overrightarrow{D1}$) patterns can be represented as vectors in the two-dimensional space they define. We can take the axes of the 2D space to be $\overrightarrow{a1}$, a vector proportional to the average of $\overrightarrow{S1}$ and $\overrightarrow{D1}$, and $\overrightarrow{d1}$, a vector proportional to their difference.

(D and E) When LN1 receives stronger (D)/weaker (E) mean external input than LN2, $\overrightarrow{S1}$ is activated positively, and $\overrightarrow{D1}$ is activated positively (D)/negatively (E). Thus, the $\overrightarrow{a1}$ components of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ add (D)/cancel (E), while the $\overrightarrow{d1}$ components of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ cancel (D)/add (E). The actual activity vectors (black) thus point in very different directions in (D) and (E).

See Figure S4 for analysis of the connectivity and the Schur patterns, and Figure S5A for comparisons of the directions of dominant activity patterns.

A set of numbers called eigenvalues can be calculated from the connectivity matrix; each eigenvalue is associated with a Schur pattern, and the real part of the eigenvalue corresponds to the strength of that pattern's self-connection. Plotting the eigenvalues of the global connectivity matrix from one representative simulation, we see that two eigenvalues have real parts more positive than the rest, indicating that there are two strongly self-excitatory activity patterns (Figure 3A; SI section 5). Analysis of the feedforward connections between patterns (SI section 4; Figure S4) shows that (1) these two patterns are nearly independent, with only a very weak feedforward connection from one to the other (Figure S4G); (2) these two patterns receive approximately orthogonal sets of feedforward inputs from the less self-excitatory patterns (Figures S4B, S4C, and S4F); and (3) the less self-excitatory patterns form much stronger feedforward connections to the two strongly self-excitatory patterns than to each other (Figure S4G). Thus, given similar external input to all patterns, these two patterns will dominate the network's activity.

To understand the structure of these two potentially dominant activity patterns, in Figure 3B we plot the relative activation of different neurons in these two Schur patterns. We have arbitrarily chosen the overall sign of each pattern in Figure 3B such that both have mostly positive elements in LN1, and we have arbitrarily set the amplitude (i.e., the vector norm) of each pattern to 1. We note two key points about these two global patterns, which together show that they represent the coupled activation of the two local slow modes. (1) The two patterns represent two different forms of coupled activation of the two LNs: one is a "sum pattern," representing roughly equal activation of the two LNs; the other is a "difference pattern," representing differential activation of the two LNs, i.e., this pattern increases the activity of one LN and decreases the activity of the other. (2) The portions of the sum and difference patterns within a given LN are very similar to each other (e.g., in Figure 3B, compare the two patterns restricted to neurons 1–100; for this comparison, the overall sign of activation within an LN is arbitrary; also see Figure S5A), as well as to the slow mode of that LN if it were not connected with the other LN (Figure S5A), which reflects the connectivity within that LN.

The sum and difference patterns, in addition to being strongly amplified by recurrent connectivity, also typically receive stronger external input than the other patterns. The intuition for this is the following: The external input is non-negative, since we

assume it's carried by purely excitatory projection neurons. The purely excitatory input to an LN most strongly activates patterns that represent concerted activation of cells in the LN, and the sum and difference patterns are the only such patterns (Figure S6). We now present the math behind this intuition.

For a given global network, consider the vector of external inputs $\vec{I}$, each of whose elements is the input to one neuron of the network. Let's decompose it as $\vec{I} = \vec{I}_{mean} + \vec{I}_{res}$, where the LN1 elements of $\vec{I}_{mean}$ all equal the mean input to LN1, which we call I_1, and similarly the LN2 elements all equal the mean input to LN2, I_2. $\vec{I}_{res}$ contains the residuals, which sum to zero over each LN. Similarly, we can decompose a Schur pattern $\vec{P}$ of the given global network as $\vec{P} = \vec{P}_{mean} + \vec{P}_{res}$, where the elements of $\vec{P}_{mean}$ are P_1 and P_2, the LN means of $\vec{P}$. Given the orthonormality of the Schur patterns, the external input to $\vec{P}$ is $\vec{I} \cdot \vec{P}$. Over each LN, the residuals sum to zero while the mean vectors are constant, so the dot product of any residual vector with any mean vector is 0. Thus, $\vec{I} \cdot \vec{P} = \vec{I}_{mean} \cdot \vec{P}_{mean} + \vec{I}_{res} \cdot \vec{P}_{res}$. The first term $\vec{I}_{mean} \cdot \vec{P}_{mean} = N(I_1 P_1 + I_2 P_2)$, where N is the number of neurons in an LN. The second term is a dot product of random vectors drawn independently for each global network and input pattern. By the central limit theorem, for large N, $\vec{I}_{res} \cdot \vec{P}_{res}$ across different random instantiations of networks and inputs approaches a Gaussian distribution with mean zero and SD $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$, where σ_I and σ_P are the SDs across the elements of $\vec{I}_{res}$ and $\vec{P}_{res}$, respectively. Thus, the typical order of magnitude of $\vec{I}_{res} \cdot \vec{P}_{res}$ in any given global network will be $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$. To compare the magnitude of $N(I_1 P_1 + I_2 P_2)$ and $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$, we note that N is much greater than $\sqrt{N}$, and I_1 and I_2 are larger than or comparable to σ_I (since external inputs are positive). For the sum and difference patterns, the absolute values of P_1 and P_2 are comparable to σ_P (Figure S6). Thus, $N(I_1 P_1 + I_2 P_2) >> \sqrt{N}\sqrt{2}\sigma_I\sigma_P$ for these two patterns, so their inputs are approximately $N(I_1 P_1 + I_2 P_2)$. For the other patterns, P_1 and P_2 are close to zero, much smaller than σ_P (specifically, for our model with N = 100, P_1, $P_2 < \sigma_P / \sqrt{N}$; Figure S6) and much smaller than the LN means of the sum and difference patterns (Figure S6). Thus, their mean-driven inputs are small, and their input is dominated by the relatively small input from the residual terms—intuitively, they represent random activations of cells and are weakly driven by the random fluctuations across cells of inputs about their mean. In our model, the range of the uniform distribution from which visual inputs for different cells are drawn is larger than that for delay inputs, which is larger than those for fixation and expectation inputs (SI section 2.2). A larger input variance (i.e., σ_I) means larger random fluctuations of inputs across cells; thus, patterns other than the sum and difference patterns are activated more strongly by visual and delay inputs than by fixation and expectation inputs. This is consistent with the large variance of activity across cells during visual and delay responses (Figures 1C and 1D), and the strong activation of weak patterns by the visual input (Ganguli et al., 2008).

When inputs are randomly redrawn for simulations of different global networks, the means of the inputs will be consistent across simulations. The large means of the sum and difference patterns within each LN will also be consistent across simulations (Figure S6) because these are primarily determined by the statistically consistent strength of the overall excess of excitation in each LN's connectivity. The means of the other patterns will be consistently small (Figure S6) because they are determined by random factors in the connectivity. Thus, across different global networks, the same analysis will apply and the sum and difference patterns will consistently receive strong input, while the other patterns will receive weaker inputs that vary from simulation to simulation.

We note that for a small proportion of random instantiations of connectivity matrices, a pair of complex patterns (which are complex conjugates in the eigenvector basis) take the place of the single real sum pattern described above. We show in SI section 6 and Figure S5 that in these cases, the complex pattern pair behaves effectively like the single real sum pattern.

We can use our understanding of the two dominant global activity patterns to understand the activity within a single LN, which we take to be LN1. We will call the LN1 portions of the sum and difference patterns $\overrightarrow{S1}$ and $\overrightarrow{D1}$, respectively, and take them to be normalized to unit vector length. Because these two patterns are not exactly equal to one another, they define a two-dimensional space of strongly amplified activity patterns in LN1. A convenient orthogonal pair of vectors to serve as a basis for this space is a vector $\overrightarrow{a1}$ proportional to the average of $\overrightarrow{S1}$ and $\overrightarrow{D1}$, and a vector $\overrightarrow{d1}$ proportional to their difference (again, both normalized to unit vector length; Figure 3C). $\overrightarrow{a1}$ represents concerted firing of the cells in LN1, while $\overrightarrow{d1}$ represents differential firing of the E and I cells in LN1. $\overrightarrow{a1}$ is almost precisely the slow mode of LN1 if it were isolated, while $\overrightarrow{d1}$ is very nearly orthogonal to that slow mode (Figure S5A). From the analysis above, the activation of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ is largely determined by the mean inputs to the two LNs, as illustrated in Figures 3D and 3E.

In Figure S7 and SI section 7, we show that because there are two strongly amplified activity patterns, single neurons in both the FK data and model no longer have a common "crossing time," as observed in BG by Ganguli et al. (2008), and we discuss possible consequences of this for attentional switching.

Detailed Analysis: Two-Dimensional Dynamics Explain Correlation Patterns

We are now in a position to understand the behavior of correlations between fixation and instantaneous activities in the FK model. Here we consider a population of neurons simultaneously recorded from a single LN, part of a single global network (Figure 4). In SI section 8, we explain why the conclusions we reach remain valid for a population in which each neuron is recorded from a different global network (the case of our main simulations in Figure 2 and likely of the experiments in Figure 1).

First, we see in simulations of a single global network that, indeed, the two dominant activity patterns, $\overrightarrow{S1}$ and $\overrightarrow{D1}$, largely explain the population-averaged activity of LN1 (Figures 4A and 4E; the results and analysis are identical for LN2). Moreover, we can see the contributions of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activity to the correlation patterns by breaking up the correlations into two

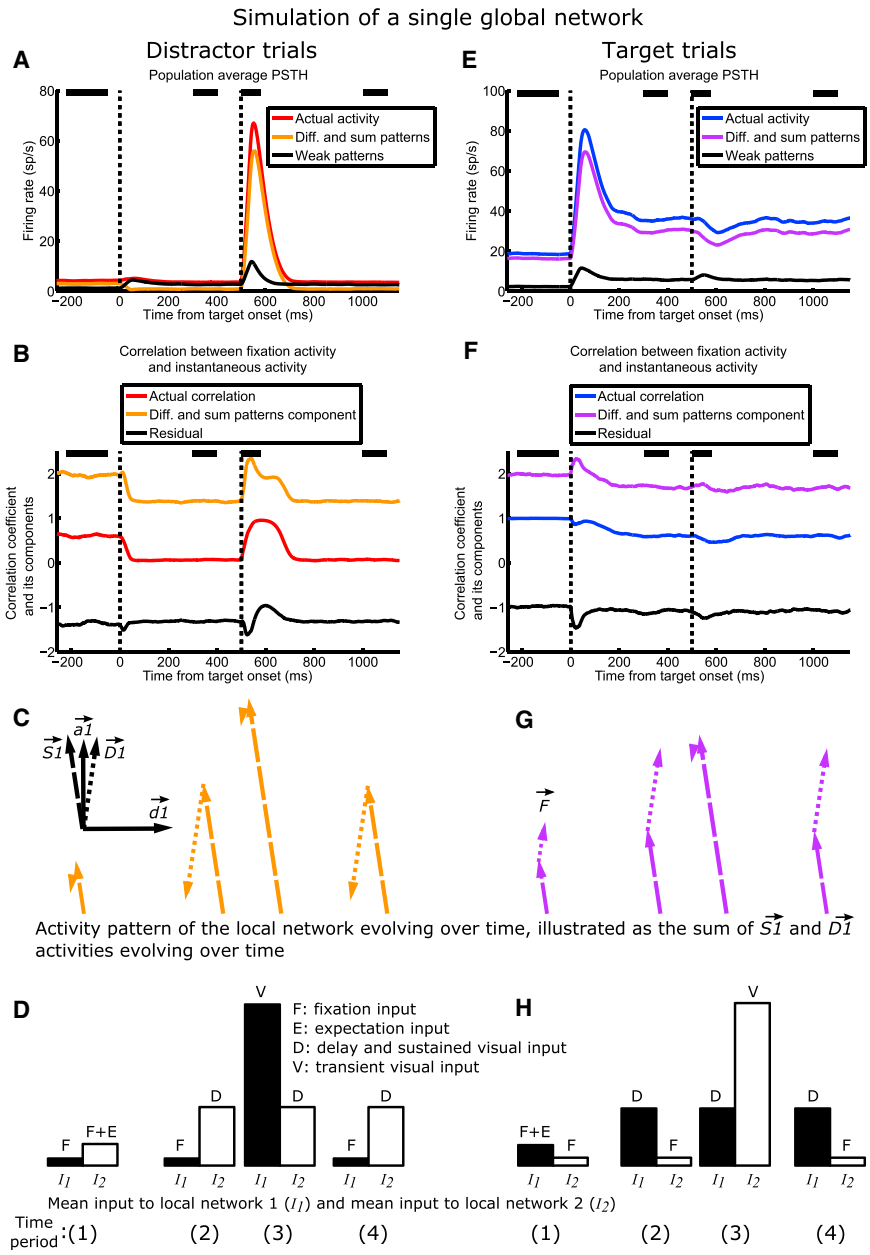

Figure 4. Two Multi-neuronal Activity Patterns Explain LIP Dynamics

One global network composed of LNs 1 and 2 is simulated, and the dynamics in LN1 on distractor trials (A–D) and target trials (E–H) are analyzed.

(A and E) $\overrightarrow{S1}$ and $\overrightarrow{D1}$ patterns dominate activity. Population average activity (red/blue), its component in the space of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ (orange/purple), and its component in the space of all other patterns (black) on distractor/target (A/E) trials (n = 100 cells). In (A), the orange and black traces add up to the red trace, and in (E), the purple and black traces add up to the blue trace.

(B and F) Actual correlation (red/blue), $Corr_{sum,diff}$ (orange/purple, the component of correlation due to the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ patterns alone), and $Corr_{residual}$ (black, the residual component) on distractor/target (B/F) trials. In (B), the orange and black traces add up to the red trace, and in (F), the purple and black traces add up to the blue trace. $Corr_{sum,diff}$ mirrors the salient ups and downs in the actual correlation, while $Corr_{residual}$ largely does not change with time—thus, the changes in actual correlation over a trial are largely due to the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ patterns. See Results for how the correlation was broken down into two components. Note that only the actual correlation, but not $Corr_{sum,diff}$ or $Corr_{residual}$, is restricted to lie within −1 and 1.

(C) Top left inset: the two-dimensional space spanned by the two 100-dimensional dominant activity patterns of LN1, $\overrightarrow{S1}$ (dashed vector) and $\overrightarrow{D1}$ (dotted vector), with $\overrightarrow{a1}$ and $\overrightarrow{d1}$ as axes.

(C and G) The evolution of $\overrightarrow{S1}$ (dashed vectors) and $\overrightarrow{D1}$ (dotted vectors) activity during distractor (C; orange vectors) and target (G; purple vectors) trials, as a result of the evolving inputs illustrated in (D) and (H). For each trial type, activities in the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ directions are each averaged over each of four time periods (spanned by black bars in A, B, E, and F), and are illustrated in their two-dimensional space, with the angle between $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities and their relative lengths accurately rendered. In this 2D space, at a given time, the activity pattern across the cells of LN1 is the vector sum of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities at that time. Thus, $\overrightarrow{F}$, the vector of target trial fixation activities, is the vector sum of the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ vectors at time (1) in (G). The angle between $\overrightarrow{F}$ and the vector sum of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities at a given time period generally determines the actual correlation at that time: the larger the angle, the lower the correlation, and vice versa (see Figure S8 for the precise relationship between the vectors and correlation). For example, the angle between the vector sum during the delay on distractor trials (vector sum of the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities at time (2) in C) and $\overrightarrow{F}$ is large, so the correlation during that time period is low; the vector sum following distractor onset on distractor trials (vector sum of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activity at time (3) in C) points in similar directions as $\overrightarrow{F}$, so the correlation during that time period is high.

(D and H) The relative inputs to LNs 1 and 2 during the four time periods on distractor (D) and target (H) trials. Black and white bars denote the mean input to LN1 (I_1) and mean input to LN2 (I_2), respectively. As illustrated in Figures 3D and 3E, during each time period, the sum of (difference between) the black and white bars largely determines the magnitude and direction of $\overrightarrow{S1}$ ($\overrightarrow{D1}$) activity, which is plotted directly above the bars in (C) and (G) (see SI section 9 for explanations of why the inputs here do not perfectly predict the activations in C and G).

components, the component due to activity in the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ patterns alone and the residual component, as follows. At any given time point, the correlation between instantaneous activity and fixation activity is $\hat{r}\cdot\hat{F}/(|\hat{r}|\,|\hat{F}|)$, where $\hat{r}$ and $\hat{F}$ are the vectors of mean-subtracted instantaneous activities and mean-subtracted fixation activities, respectively (each element of $\hat{r}$ is the instantaneous activity of one neuron minus the population mean instantaneous activity, and similarly for $\hat{F}$), and $|\cdot|$ denotes vector norm. We break $\hat{r}$ into components $\hat{r}_{sum}$, $\hat{r}_{diff}$, and $\hat{r}_{weak}$, the mean-subtracted instantaneous activity in the $\overrightarrow{S1}$ pattern,

the $\overrightarrow{D1}$ pattern, and all other patterns, respectively, and do likewise for $\hat{F}$:

$$\begin{aligned}\frac{\hat{r}\cdot\hat{F}}{|\hat{r}||\hat{F}|} &= \frac{(\hat{r}_{sum}+\hat{r}_{diff}+\hat{r}_{weak})\cdot\left(\hat{F}_{sum}+\hat{F}_{diff}+\hat{F}_{weak}\right)}{|\hat{r}||\hat{F}|}\\ &= \frac{(\hat{r}_{sum}+\hat{r}_{diff})\cdot\left(\hat{F}_{sum}+\hat{F}_{diff}\right)}{|\hat{r}||\hat{F}|}\\ &\quad + \frac{(\hat{r}_{sum}+\hat{r}_{diff})\cdot\hat{F}_{weak}+\hat{r}_{weak}\cdot\left(\hat{F}_{sum}+\hat{F}_{diff}\right)+\hat{r}_{weak}\cdot\hat{F}_{weak}}{|\hat{r}||\hat{F}|}\\ &= Corr_{sum,diff}+Corr_{residual}.\end{aligned}$$

The two terms $Corr_{sum,diff}$ and $Corr_{residual}$ that sum to the actual correlation are plotted in Figures 4B and 4F—we see that $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities largely explain the qualitative changes in correlations over time.

Thus, the actual activity pattern across cells of the LN, $\vec{r}$, can be approximated as the vector sum of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities, which determines the correlation patterns. Figures 4C, 4G, and S8 illustrate $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities evolving over four time periods during a trial, as well as how their dynamics explain the correlation patterns. In SI section 10, we discuss why correlations at the peaks of visual responses are higher in FK than in BG. In Figures S9 and S10 and SI section 11, we show how the model dynamics change smoothly from the BG to the FK case for varying strengths of surround suppression.

Now we turn to examine how the dynamics of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities are determined by their inputs. We have shown above that the external input to the sum or difference pattern is approximately $N(I_1\ P_1+I_2\ P_2)$. We note that the absolute values of P_1 and P_2 for the sum and difference patterns are all about equal (Figure 3B), which we can call m. That is, for the sum pattern, $P_1 \approx P_2 \approx m$, while for the difference pattern, $P_1 \approx m$ and $P_2 \approx -m$. Thus, the inputs to the sum and difference patterns are approximately $Nm(I_1+I_2)$ and $Nm(I_1-I_2)$, respectively. When the network is in a steady state, these inputs are amplified by the connectivity: $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activations are approximately given by $1/\sqrt{2}\ Nm(I_1+I_2)/(1-\lambda_S)$ and $1/\sqrt{2}\ Nm(I_1-I_2)/(1-\lambda_D)$ respectively, where λ_S and λ_D are the eigenvalues of the sum and difference patterns, respectively (for this approximation, we ignore the feedforward input from the other patterns, which is weak relative to the mean-driven input; the $1/\sqrt{2}$ factor arises because $\overrightarrow{S1}$ is only the LN1 half of the sum pattern, and similarly for $\overrightarrow{D1}$). As N, m, λ_S, and λ_D are all fixed properties of the network, $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities just depend on the dynamics of the mean inputs I_1 and I_2, being simply proportional to I_1+I_2 and I_1-I_2, respectively (Figures 4C, 4D, 4G, and 4H).

Direct Evidence for Two-Dimensional Dynamics in the FK Dataset

Since we propose that the BG data are predominantly one-dimensional and the FK data two-dimensional, we used principal component analysis (PCA) to directly examine the dimensionality of the two datasets. We focus on distractor trials because our correlation analysis revealed that they show the most salient dynamical differences between BG and FK. Furthermore, our model predicts that the activity on FK distractor trials has both large $\overrightarrow{a1}$ and $\overrightarrow{d1}$ activations, and thus is likely to reveal two dynamical dimensions, whereas the activity on FK target trials is more strongly one-dimensional, dominated by the single $\overrightarrow{a1}$ direction (Figures 4C and 4G). We excluded the transient visual responses to the distractor, as they involve activation of weak patterns, and performed PCA on the remaining slowly varying activity patterns of distractor trials. The results indeed confirm the one-dimensionality of the BG data and two-dimensionality of the FK data (Figures 5A and 5B).

Given the 2D space spanned by the top two principal components (PCs) identified from the FK data, we ask further, do activity patterns in this dominant 2D space actually behave as our model predicts? To answer this question, we first estimate the activity directions in the data that correspond to those in our model. We take the direction with the maximum mean firing rate within the 2D space of the two PCs as the putative $\overrightarrow{a1}$ direction, since $\overrightarrow{a1}$ is a direction representing concerted firing of neurons in an LN, and take the direction orthogonal to the putative $\overrightarrow{a1}$ as the putative $\overrightarrow{d1}$ (Figure 5C). In Figures 5D–5G, we plot the activities over time in the $\overrightarrow{a1}$ and $\overrightarrow{d1}$ directions in the data and model (in SI section 12, we discuss differences between data and model). The activities in the putative $\overrightarrow{a1}$ and $\overrightarrow{d1}$ directions in the data match those predicted by the model, providing direct evidence that our proposed two-dimensional dynamics underlie the FK data.

In SI section 13, we discuss the dynamics and dimensionality of E and I subpopulations.

Two-Dimensional Dynamics Suggest a Recurrent Origin for LIP Surround Suppression

Surround suppression is observed in multiple cortical areas (reviewed in Rubin et al., 2015) and has been extensively studied as a model for understanding cortical computations and circuit mechanisms (e.g., in V1; Ozeki et al., 2009; Rubin et al., 2015; and see review by Nurminen and Angelucci, 2014). When considering surround suppression in a given cortical area, a key mechanistic question is the following: to what extent is it inherited from surround suppression in other areas, i.e., resulting from a withdrawal of input from those areas, and to what extent is it due to reciprocal, suppressive coupling within the given area?

Of areas that directly or indirectly project to LIP (Blatt et al., 1990; Clower et al., 2001), surround suppression has been observed in MT (Hunter and Born, 2011; Tsui and Pack, 2011), V4 (Desimone and Schein, 1987; Sundberg et al., 2009), superior colliculus (Dorris et al., 2007), FEF (Schall and Hanes, 1993; Cavanaugh et al., 2012), and dlPFC (Suzuki and Gottlieb, 2013). Thus, it is possible that LIP surround suppression is inherited from one or more of these areas. However, according to our model, the observed pattern of correlation between fixation and instantaneous activity arises from the coupling of local LIP networks. We argue that the experimentally observed correlation pattern is a signature indicating that surround suppression of external input alone cannot account for LIP surround suppression.

This can be demonstrated by simulating the scenario of the null hypothesis—LIP surround suppression being inherited from external inputs. In this version of the model, the two LNs

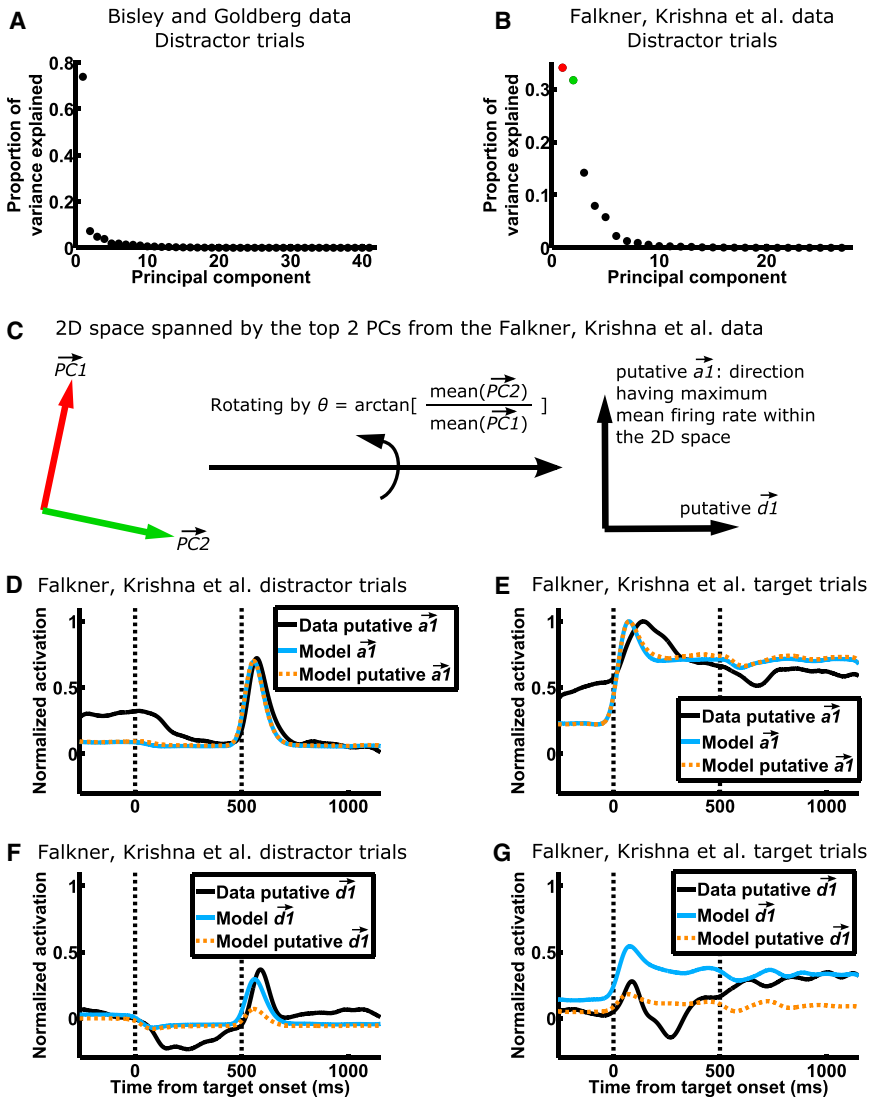

Figure 5. Direct Evidence for Two-Dimensional Dynamics in the Falkner, Krishna et al. Data

(A and B) PCA, where the variables are neurons and the observations are the instantaneous activity vectors during distractor trials, for the BG (A) and FK (B) datasets. Activity vectors during the transient visual responses to the distractor (600–1,100 ms after target onset for BG; 450–750 ms after target onset for FK) were not included for this analysis because they involve activation of weak patterns. The majority of the variance is explained by one PC in BG, while a comparable proportion is explained by two PCs in FK, consistent with one-dimensional dynamics in BG and two-dimensional dynamics in FK.

(C) We hypothesize that the 2D space spanned by the top two PCs (colored as in B) in the FK data is the 2D space of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ (Figures 3C, 4C, and 4G). We further hypothesize that the direction with the maximum mean firing rate within the 2D space of the two PCs is the putative $\overrightarrow{a1}$ direction, since $\overrightarrow{a1}$ represents concerted firing of neurons in an LN. We can thus find the putative $\overrightarrow{a1}$ and $\overrightarrow{d1}$ of the FK data by rotating the two PCs by an angle of arctan[mean $(\overrightarrow{PC1})$ /mean $(\overrightarrow{PC2})$], where mean(•) denotes mean over the elements of a vector.

(D–G) The activation of $\overrightarrow{a1}$ (D and E) and $\overrightarrow{d1}$ (F and G) on FK distractor trials (D and F) and target trials (E and G). In D and E, the data putative $\overrightarrow{a1}$ was derived as in (C). To determine activation in the model, one cell was "recorded" from each of multiple simulated global networks to form the model population. To determine the model $\overrightarrow{a1}$, suppose the *i*th cell of the model population is the *j*th cell from LN1 of the *i*th global network. Then the *i*th element of the model $\overrightarrow{a1}$ is the *j*th element of the actual $\overrightarrow{a1}$ of the *i*th global network. The model putative $\overrightarrow{a1}$ was derived as in (C), but from the model population. The $\overrightarrow{d1}$ directions are determined similarly. Each set of activations (e.g., the four activation traces of data putative $\overrightarrow{a1}$ and $\overrightarrow{d1}$ on target and distractor trials comprise a set) is normalized by its peak $\overrightarrow{a1}$ activation on target trials—thus, (D)–(G) share the same scale. Vertical dashed lines denote the onsets of the target and distractor.

are uncoupled. Whenever a stimulus appears or a saccade is planned, the LN with the corresponding RF is activated by visual or delay input; at the same time, the external input to the other LN is reduced, modeling surround suppression inherent in one or more input sources (see SI section 2.2 for model details). Figure 6A shows the population average PSTHs from a simulation of the FK experiment using this model. On the surface, if we examine only the firing rates, this model of surround suppression reproduces the experimental data. However, if we examine the underlying network dynamics using the correlation analysis (Figure 6B), we find that this model cannot reproduce the experimentally observed correlation pattern. Specifically, the dynamics of each LN here are dominated by its slow mode, as in BG (Figures 1E and 2E).

There are alternative mechanisms that conceivably could account for the FK correlations. We consider these in SI section 14 but conclude, for multiple reasons, that the most likely and parsimonious interpretation is that surround suppression in LIP arises primarily from direct suppressive coupling between LIP LNs.

DISCUSSION

By demonstrating that recurrent interactions between LIP LNs are likely to drive LIP surround suppression, our study suggests the active involvement of LIP in attentional and saccadic selection. LIP is part of a fronto-parietal-collicular network that mediates attentional guidance and eye movements, and the attentional and saccadic priorities associated with different locations (a "priority map") are encoded in the activity of neurons in this network with the corresponding RF locations (Bisley and Goldberg, 2010). It has long been theorized that different

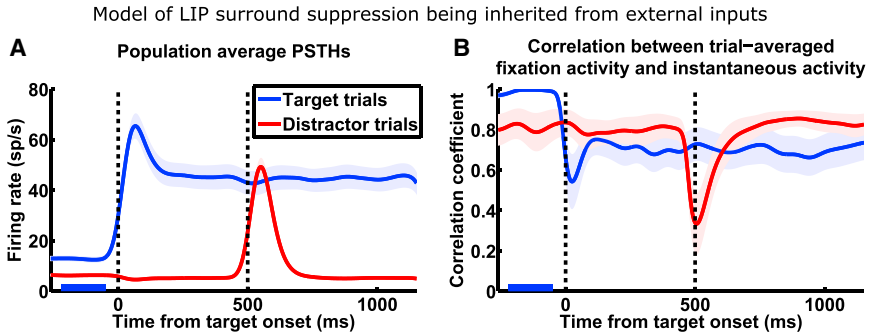

Figure 6. Model of Inherited Surround Suppression Cannot Reproduce Observed Network Dynamics

The model: two LIP LNs are uncoupled; whenever one LN receives visual or delay input, the external input to the other LN is reduced.

(A) Population average PSTHs (n = 27 cells; same conventions as Figure 1D) show that this model reproduces activity observed by FK (Figure 1D) during surround interactions.

(B) This model fails to reproduce network dynamics observed by FK (Figure 1F) during surround interactions. Correlation analysis with same conventions as Figure 1F.

locations on this priority map mutually suppress each other to facilitate attentional and saccadic selection, to allow persistent focus by resisting distraction, and to allow the planning and execution of sequential saccades (Itti and Koch, 2001; Constantinidis and Wang, 2004; Xing and Andersen, 2000). However, the neural substrates and mechanisms of these processes are not clear. Our results suggest that LIP directly participates in these processes and shapes the priority map, instead of merely reflecting computations achieved in other areas.

Ganguli et al. (2008) showed that the slow dynamics of a local LIP network (i.e., neurons sharing the same RF) are 1D, but we can now see that this result is restricted to the case in which a single LN is activated without activation of other LNs to which it is coupled. Nonetheless, we can simply understand the more complex, 2D slow dynamics we have found in the FK data as resulting from coupling between networks that, in isolation, each have 1D slow dynamics. This lays a basis for understanding higher-dimensional local dynamics induced by interactions between larger numbers of simultaneously activated LNs. For example, during visual search, it takes longer to find a target when there are more distractors, and, correspondingly, LIP activity is lower when there are more distractors (Balan et al., 2008). This likely results from increased surround suppression by larger numbers of activated LNs, which would yield correlations corresponding to a higher-dimensional dominant activity space (e.g., number of dimensions equal to the number of mutually interacting networks). Our study thus provides a basis for analyzing activity dynamics when multiple stimuli evoke interaction of multiple local LIP networks, as occurs in natural visual environments.

Our results have implications for the mechanisms underlying certain types of perceptual decision making, where saccadic decisions are made based on noisy sensory evidence (Roitman and Shadlen, 2002; Gold and Shadlen, 2007). This type of decision making has been posited to involve two neuronal pools that integrate opposing sensory evidence, each either accumulating evidence independently and racing toward a decision (Mazurek et al., 2003) or competing with the other by mutual inhibition (Wong and Wang, 2006; Usher and McClelland, 2001). Recent results show that these neuronal pools are not exclusively in LIP (Katz et al., 2016), but if they are distributed across multiple areas that include LIP, which of the two classes of models applies to each instance of decision making would depend on whether the two neuronal pools are recurrently coupled. When they are not coupled (like the neuronal pools studied by BG), the independent accumulator model would apply, and when they are coupled, the mechanisms described here would contribute to the competition that leads to decision making. In the future, it would be interesting to study decision making separately for cases in which the two alternative choices engage LIP LNs that do or do not suppress one another (as BG and FK have done in studying attentional switching), examine the neural correlates on the population level, and compare them with behavior.

The interactions between LNs we describe would cause priority assignments to be in part determined in relative terms, as has been observed in certain forms of value-based decision making. When different saccade targets are associated with different magnitudes of reward or reward probabilities, some LIP neurons encode the expected value of different saccades (Platt and Glimcher, 1999; Dorris and Glimcher, 2004). Importantly, these value representations in LIP are relative, such that the response to one saccade target depends on its value relative to those of other possible saccade targets; this relative value encoding is well described by the phenomenological model of divisive normalization (Louie et al., 2011; Carandini and Heeger, 2011). Surround suppression, computed within LIP in ways similar to those described here, provides a circuit mechanism for divisive normalization of value representations (Louie et al., 2014; LoFaro et al., 2014; Rubin et al., 2015).

Regardless of the cognitive context in which LIP function has been investigated, research has often focused on single-neuron activity or the average activity of LIP populations. Our work adds to other recent work (e.g., see review by Cunningham and Yu, 2014) in suggesting that there is much information in the activity patterns across neurons, which change as a function of external stimuli and internal goals such as saccade plans. As we have seen, even when the mean activity of a population changes only subtly, the pattern of activity across neurons can change drastically (e.g., when a target appears in the suppressive surround of a local LIP population). Thus, beyond the information carried by single neurons or their average activity, downstream areas could potentially read out information from the activity pattern across LIP neurons, although whether or how downstream areas do this remains to be examined. This is potentially important in the natural context, where LIP LNs must interact to process a multitude of changing visual stimuli and internal goals to guide behavior.

EXPERIMENTAL PROCEDURES

A full description of modeling procedures is found in SI section 2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and ten figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.11.023.

AUTHOR CONTRIBUTIONS

W.Z. and K.D.M. performed the analysis and modeling, and wrote the paper with inputs from all authors. A.L.F., B.S.K., and M.E.G. performed the experiments.

ACKNOWLEDGMENTS

We thank S. Morrison for comments on the manuscript; J. Gottlieb and A. Kennedy for helpful suggestions; and M. Osman, M. Shalev, G. Asfaw, Y. Pavlova, J. Caban, G. Duncan, and L. Palmer for assistance with experiments. K.D.M. supported by NIH R01 EY11001 and the Gatsby Charitable Foundation; W.Z. by NIH T32 EY013933 and T32 NS064929; M.E.G. by NIH R24 EY015634, R01 EY01497, and R01 EY017039, and the Kavli, Keck, Dana, and Zegar Foundations; A.L.F. by the NSF Graduate Research Fellowship and the Ruth L. Kirschstein National Research Service Award; and B.S.K. by the German Ministry for Education and Research Grants BMBF 01GQ0433 and 01GQ1005C to the Bernstein Center for Computational Neuroscience and a Gatsby Foundation award.

Received: July 3, 2015
Revised: July 13, 2016
Accepted: November 7, 2016
Published: December 15, 2016

REFERENCES

Baizer, J.S., Ungerleider, L.G., and Desimone, R. (1991). Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. J. Neurosci. *11*, 168–190.

Balan, P.F., Oristaglio, J., Schneider, D.M., and Gottlieb, J. (2008). Neuronal correlates of the set-size effect in monkey lateral intraparietal area. PLoS Biol. *6*, e158.

Bisley, J.W., and Goldberg, M.E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. Science *299*, 81–86.

Bisley, J.W., and Goldberg, M.E. (2006). Neural correlates of attention and distractibility in the lateral intraparietal area. J. Neurophysiol. *95*, 1696–1717.

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

Blatt, G.J., Andersen, R.A., and Stoner, G.R. (1990). Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. J. Comp. Neurol. *299*, 421–445.

Carandini, M., and Heeger, D.J. (2011). Normalization as a canonical neural computation. Nat. Rev. Neurosci. *13*, 51–62.

Cavanaugh, J., Joiner, W.M., and Wurtz, R.H. (2012). Suppressive surrounds of receptive fields in monkey frontal eye field. J. Neurosci. *32*, 12284–12293.

Clower, D.M., West, R.A., Lynch, J.C., and Strick, P.L. (2001). The inferior parietal lobule is the target of output from the superior colliculus, hippocampus, and cerebellum. J. Neurosci. *21*, 6283–6291.

Coe, B., Tomihara, K., Matsuzawa, M., and Hikosaka, O. (2002). Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J. Neurosci. *22*, 5081–5090.

Constantinidis, C., and Wang, X.-J. (2004). A neural circuit basis for spatial working memory. Neuroscientist *10*, 553–565.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500–1509.

Dayan, P., and Abbott, L.F. (2005). Theoretical Neuroscience (The MIT Press).

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

Dorris, M.C., and Glimcher, P.W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron *44*, 365–378.

Dorris, M.C., Olivier, E., and Munoz, D.P. (2007). Competitive integration of visual and preparatory signals in the superior colliculus during saccadic programming. J. Neurosci. *27*, 5053–5062.

Elston, G.N., and Rosa, M.G. (1997). The occipitoparietal pathway of the macaque monkey: comparison of pyramidal cell morphology in layer III of functionally related cortical visual areas. Cereb. Cortex *7*, 432–452.

Falkner, A.L., Krishna, B.S., and Goldberg, M.E. (2010). Surround suppression sharpens the priority map in the lateral intraparietal area. J. Neurosci. *30*, 12787–12797.

Fitzgerald, J.K., Freedman, D.J., Fanini, A., Bennur, S., Gold, J.I., and Assad, J.A. (2013). Biased associative representations in parietal cortex. Neuron *77*, 180–191.

Freedman, D.J., and Assad, J.A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. Nat. Neurosci. *14*, 143–146.

Ganguli, S., Bisley, J.W., Roitman, J.D., Shadlen, M.N., Goldberg, M.E., and Miller, K.D. (2008). One-dimensional dynamics of attention and decision making in LIP. Neuron *58*, 15–25.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574.

Goldman, M.S. (2009). Memory without feedback in a neural network. Neuron *61*, 621–634.

Hunter, J.N., and Born, R.T. (2011). Stimulus-dependent modulation of suppressive influences in MT. J. Neurosci. *31*, 678–686.

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. Nat. Rev. Neurosci. *2*, 194–203.

Kable, J.W., and Glimcher, P.W. (2009). The neurobiology of decision: consensus and controversy. Neuron *63*, 733–745.

Katz, L.N., Yates, J.L., Pillow, J.W., and Huk, A.C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature *535*, 285–288.

Kennerley, S.W., and Wallis, J.D. (2009). Reward-dependent modulation of working memory in lateral prefrontal cortex. J. Neurosci. *29*, 3259–3270.

Leon, M.I., and Shadlen, M.N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron *24*, 415–425.

Lewis, J.W., and Van Essen, D.C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. J. Comp. Neurol. *428*, 112–137.

LoFaro, T., Louie, K., Webb, R., and Glimcher, P.W. (2014). The temporal dynamics of cortical normalization models of decision-making. Lett. Biomath. *1*, 209–220.

Louie, K., Grattan, L.E., and Glimcher, P.W. (2011). Reward value-based gain control: divisive normalization in parietal cortex. J. Neurosci. *31*, 10627–10639.

Louie, K., LoFaro, T., Webb, R., and Glimcher, P.W. (2014). Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. J. Neurosci. *34*, 16046–16057.

Mazurek, M.E., Roitman, J.D., Ditterich, J., and Shadlen, M.N. (2003). A role for neural integrators in perceptual decision making. Cereb. Cortex *13*, 1257–1269.

Miller, E.K., and Wilson, M.A. (2008). All my circuits: using multiple electrodes to understand functioning neural networks. Neuron *60*, 483–488.

Murphy, B.K., and Miller, K.D. (2009). Balanced amplification: a new mechanism of selective amplification of neural activity patterns. Neuron *61*, 635–648.

Nurminen, L., and Angelucci, A. (2014). Multiple components of surround modulation in primary visual cortex: multiple neural circuits with multiple functions? Vision Res. *104*, 47–56.

Ozeki, H., Finn, I.M., Schaffer, E.S., Miller, K.D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. Neuron *62*, 578–592.

Patel, G.H., Shulman, G.L., Baker, J.T., Akbudak, E., Snyder, A.Z., Snyder, L.H., and Corbetta, M. (2010). Topographic organization of macaque area LIP. Proc. Natl. Acad. Sci. USA *107*, 4728–4733.

Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.

Roesch, M.R., and Olson, C.R. (2003). Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. J. Neurophysiol. *90*, 1766–1789.

Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. *22*, 9475–9489.

Rubin, D.B., Van Hooser, S.D., and Miller, K.D. (2015). The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. Neuron *85*, 402–417.

Schall, J.D., and Hanes, D.P. (1993). Neural basis of saccade target selection in frontal eye field during visual search. Nature *366*, 467–469.

Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. Annu. Rev. Neurosci. *36*, 337–359.

Stanton, G.B., Bruce, C.J., and Goldberg, M.E. (1995). Topography of projections to posterior cortical areas from the macaque frontal eye fields. J. Comp. Neurol. *353*, 291–305.

Sundberg, K.A., Mitchell, J.F., and Reynolds, J.H. (2009). Spatial attention modulates center-surround interactions in macaque visual area v4. Neuron *61*, 952–963.

Suzuki, M., and Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. Nat. Neurosci. *16*, 98–104.

Tsui, J.M.G., and Pack, C.C. (2011). Contrast sensitivity of MT receptive field centers and surrounds. J. Neurophysiol. *106*, 1888–1900.

Usher, M., and McClelland, J.L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. Psychol. Rev. *108*, 550–592.

Wong, K.-F., and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. *26*, 1314–1328.

Xing, J., and Andersen, R.A. (2000). Memory activity of LIP neurons for sequential eye movements simulated with neural networks. J. Neurophysiol. *84*, 651–665.

Neuron, Volume *93*

Supplemental Information

Coupling between One-Dimensional Networks Reconciles Conflicting Dynamics in LIP and Reveals Its Recurrent Circuitry

Wujie Zhang, Annegret L. Falkner, B. Suresh Krishna, Michael E. Goldberg, and Kenneth D. Miller

Inventory of Supplemental Items

1. Figure S1, related to Figures 1 and 2
2. Figure S2, related to Figures 1 and 2
3. Figure S3, related to Figures 1 and 2
4. Figure S4, related to Figure 3
5. Figure S5, related to Figure 3
6. Figure S6, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"
7. Figure S7, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"
8. Figure S8, related to Figure 4
9. Figure S9, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"
10. Figure S10, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"
11. Section 1: Task details; related to the Results sections "One-dimensional dynamics in LIP" and "Surround suppression and violations of one-dimensional dynamics"
12. Section 2: Modeling and analysis procedures; related to Experimental Procedures
13. Section 3: Implications of different mechanisms of persistent activity for two-dimensional dynamics; related to the Results section "Simple model of coupled local networks reconciles the results"
14. Section 4: Analysis of feedforward connections in the Schur form of the connectivity matrix; related to Figure 3
15. Section 5: The eigenvalues of the sum and difference patterns; related to Figure 3
16. Section 6: Equivalence of complex sum pattern pairs with single real sum patterns; related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"

17. Section 7: The consequences of low-dimensional dynamics for attentional switching; related to the Results section “Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes”
18. Section 8: Unconnected neurons behave like neurons in a single local network; related to the Results section “Detailed analysis: two-dimensional dynamics explain correlation patterns”
19. Section 9: Discrepancies between the magnitudes of activity patterns in Fig. 4C and G and their inputs in Fig. 4D and H; related to Figure 4
20. Section 10: Difference in correlation drop evoked by transient visual stimulation between the Bisley and Goldberg and the Falkner, Krishna et al. datasets; related to Figures 1 and 2
21. Section 11: Network dynamics underlying different levels of surround suppression; related to the Results section “Detailed analysis: two-dimensional dynamics explain correlation patterns”
22. Section 12: Differences in PCA results between the FK data and model; related to Figure 5
23. Section 13: Dynamics and dimensionality of excitatory populations and inhibitory populations; related to the Results section “Direct evidence for two-dimensional dynamics in the Falkner, Krishna et al. dataset”
24. Section 14: Alternative mechanisms for surround suppression and 2D dynamics; related to the Results section “Two-dimensional dynamics suggest a recurrent origin for LIP surround suppression”

Supplemental Figures and Legends

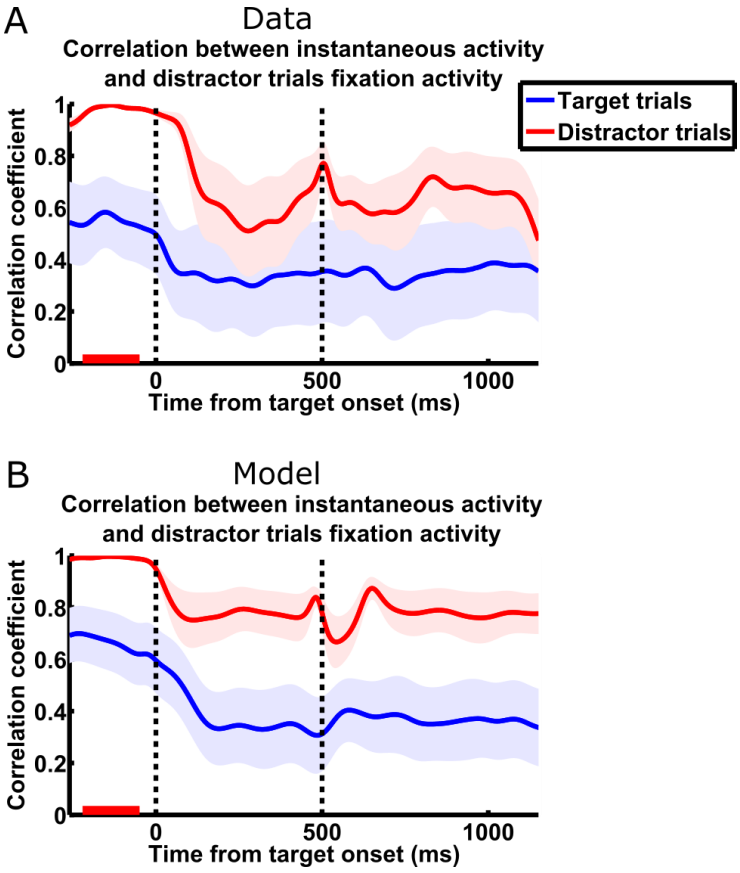

Figure S1, related to Figures 1 and 2

Correlations between instantaneous activity and distractor trial fixation activity of the Falkner, Krishna et al. (FK) data and simulation.

(A) Correlation analysis on the FK dataset, calculated using distractor trial fixation activity. The correlations are calculated similarly to that in Fig. 1F, except that fixation activity is averaged over distractor trials (over the period from 220 ms to 50 ms before target onset, marked by the red bar) instead of target trials. Same conventions as Fig. 1F.

(B) Correlation analysis on the FK simulation results (same simulated dataset as that in Fig. 2D and F), calculated using distractor trial fixation activity. Same conventions as Fig. 1F.

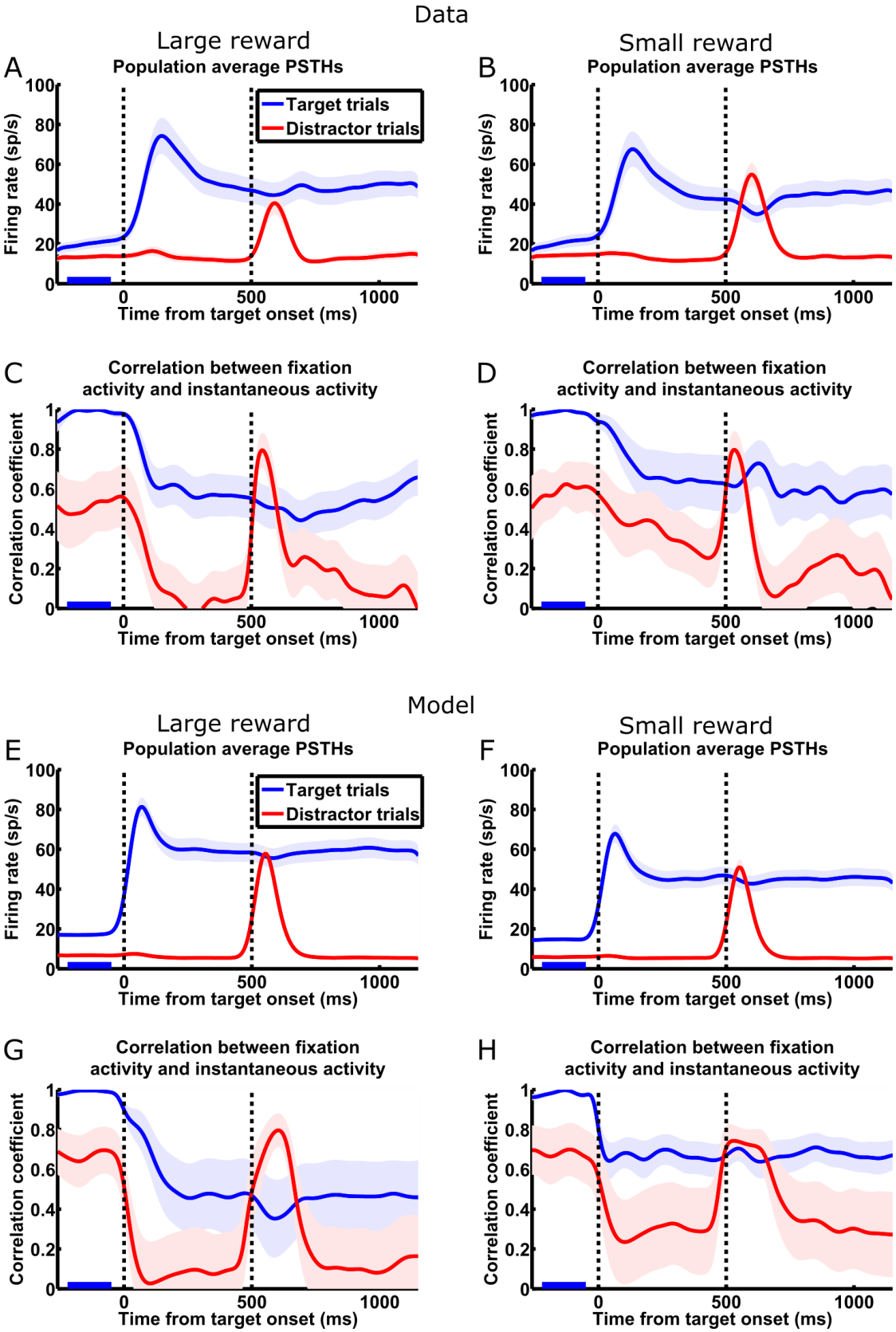
Data
Large reward
Small reward
A
Population average PSTHs
Target trials
Distractor trials
Firing rate (sp/s)
Time from target onset (ms)
B
Population average PSTHs
C
Correlation between fixation activity and instantaneous activity
Correlation coefficient
D
Correlation between fixation activity and instantaneous activity
Model
E
F
G
H

Figure S2, related to Figures 1 and 2

The Falkner, Krishna et al. data and simulation results, plotted separately for different reward conditions.

(A and B) Population average PSTHs on large reward (A) and small reward (B) trials (n = 27 cells) in the FK dataset. Same conventions as Fig. 1D.

(C and D) Correlation analysis on large reward (C) and small reward (D) trials in the FK dataset. Correlations are calculated similarly to that in Fig. 1F, except that fixation activity is averaged over only target trials with large reward (C) or small reward (D). Same conventions as Fig. 1F.

(E and F) Activity in separate simulations of the large reward (E) and small reward (F) conditions of the FK experiment (n = 27 cells). Large and small rewards were modeled by using delay input ranges (parameters I_{D1} and I_{D2}) of 7 – 67 and 2 – 62, respectively. Population average PSTHs with same conventions as Fig. 1D.

(G and H) Correlations from the large reward (G) and small reward (H) simulations shown in E and F. Correlations are calculated similarly to that in Fig. 1F, except that fixation activity is averaged over only target trials with large reward (G) or small reward (H). Same conventions as Fig. 1F.

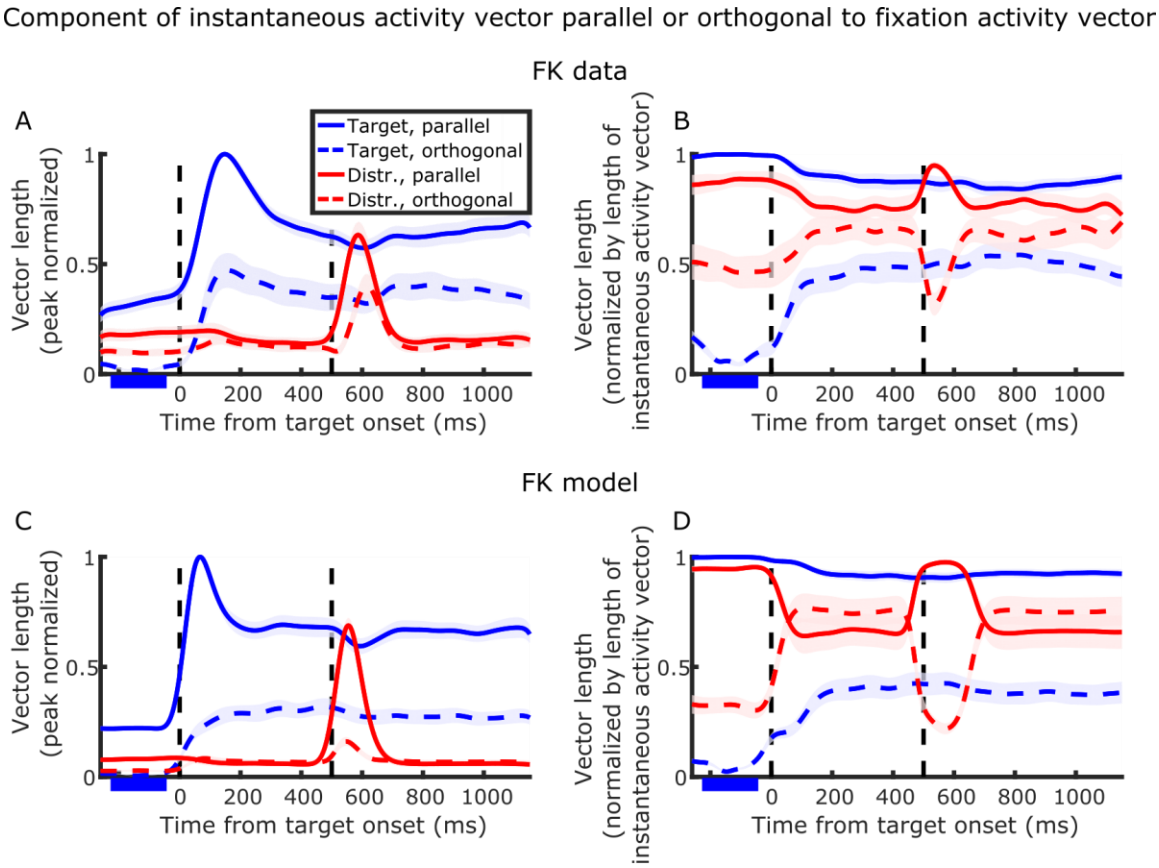

Figure S3, related to Figures 1 and 2

For each instantaneous activity vector, the lengths of its component parallel to $\vec{F}$ (solid traces) and its component orthogonal to $\vec{F}$ (dashed traces), on target trials (blue) and distractor trials (red), for FK data (A-B) and model (C-D). In A and C, the parallel and orthogonal components are both normalized to the peak length of the component parallel to $\vec{F}$ on target trials in the respective panel, so that the units in each panel are constant across time. In B and D, at each time point for a given trial type, the parallel and orthogonal components are normalized by the length of the instantaneous activity vector, so that the sum of squares of the two components always equals 1. In each panel, the first and second vertical dashed lines denote the onset of the target and the distractor, respectively; the period over which activity on target trials is averaged to calculate $\vec{F}$ is marked by a blue bar. Shading around traces indicates standard error estimated from 1000 bootstrap samples.

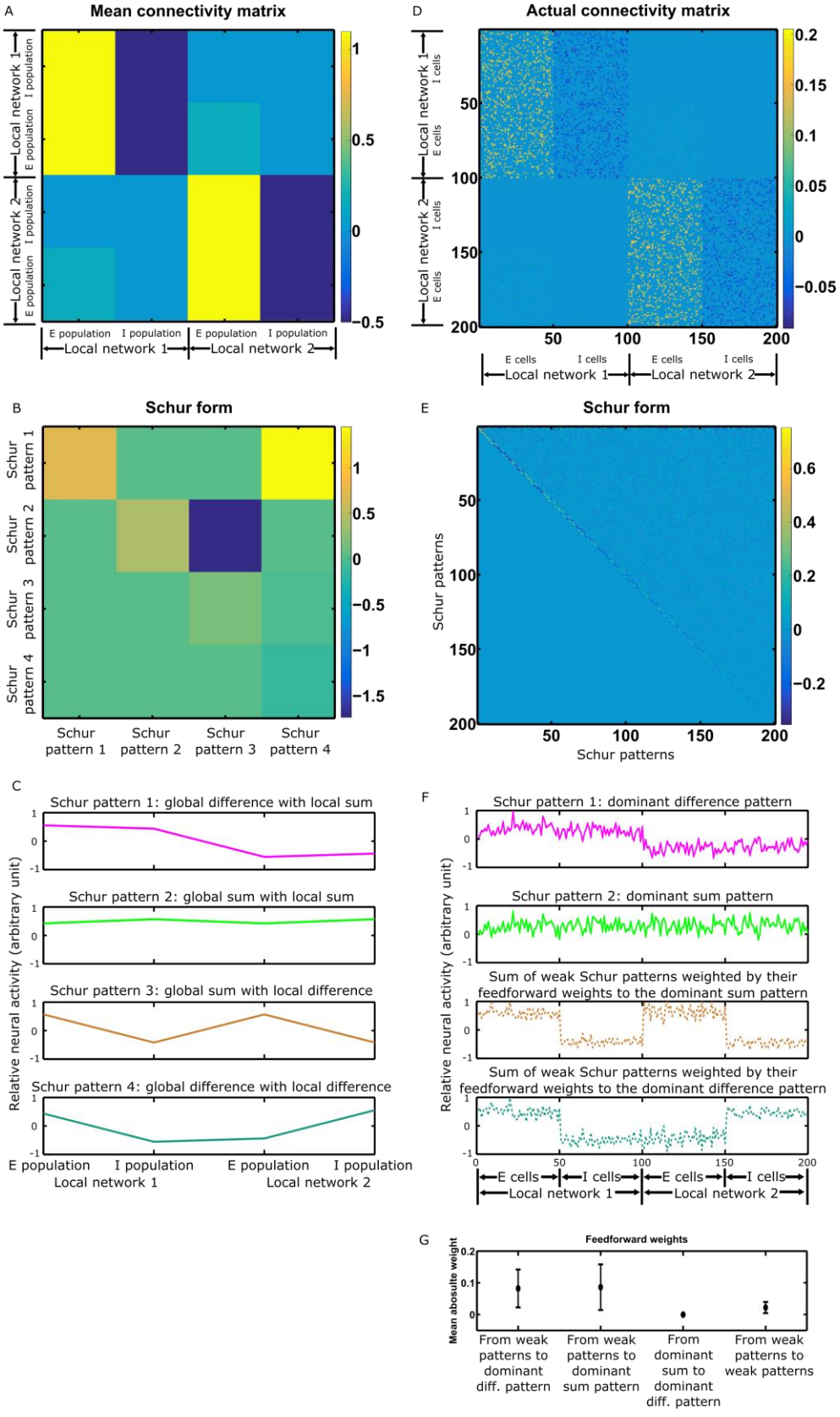
A
Mean connectivity matrix
Local network 1
Local network 2
E population
I population
1
0.5
0
−0.5
D
Actual connectivity matrix
E cells
I cells
0.2
0.15
0.1
0.05
0
−0.05
50
100
150
200
B
Schur form
Schur pattern 1
Schur pattern 2
Schur pattern 3
Schur pattern 4
1
0.5
0
−0.5
−1
−1.5
E
Schur form
Schur patterns
0.6
0.4
0.2
0
−0.2
C
Schur pattern 1: global difference with local sum
Schur pattern 2: global sum with local sum
Schur pattern 3: global sum with local difference
Schur pattern 4: global difference with local difference
Relative neural activity (arbitrary unit)
F
Schur pattern 1: dominant difference pattern
Schur pattern 2: dominant sum pattern
Sum of weak Schur patterns weighted by their feedforward weights to the dominant sum pattern
Sum of weak Schur patterns weighted by their feedforward weights to the dominant difference pattern
G
Feedforward weights
Mean absoulte weight
From weak patterns to dominant diff. pattern
From weak patterns to dominant sum pattern
From dominant sum to dominant diff. pattern
From weak patterns to weak patterns

Figure S4, related to Figure 3

Analysis of the Schur form of the connectivity matrix. See SI section 4 for details.

(A) A mean population connectivity matrix between the E and I populations of two LNs.

(B) The Schur form of the mean population connectivity matrix.

(C) The four Schur patterns of the mean population connectivity matrix.

(D) An actual connectivity matrix. The same one analyzed in Fig. 3.

(E) The Schur form of the actual connectivity matrix.

(F) The two leading Schur patterns (the dominant difference and sum patterns in Fig. 3), and sums of all other Schur patterns weighted by their feedforward weights to each of the two leading Schur patterns, respectively. The leading Schur patterns and the weighted sums correspond to the Schur patterns of the mean population connectivity matrix (C).

(G) Comparison of mean absolute feedforward weights in E, with standard deviations. The strongest feedforward connections are those from the weak patterns to the leading patterns. The feedforward weight from the dominant sum pattern to the dominant difference pattern is small, making these two patterns effectively independent.

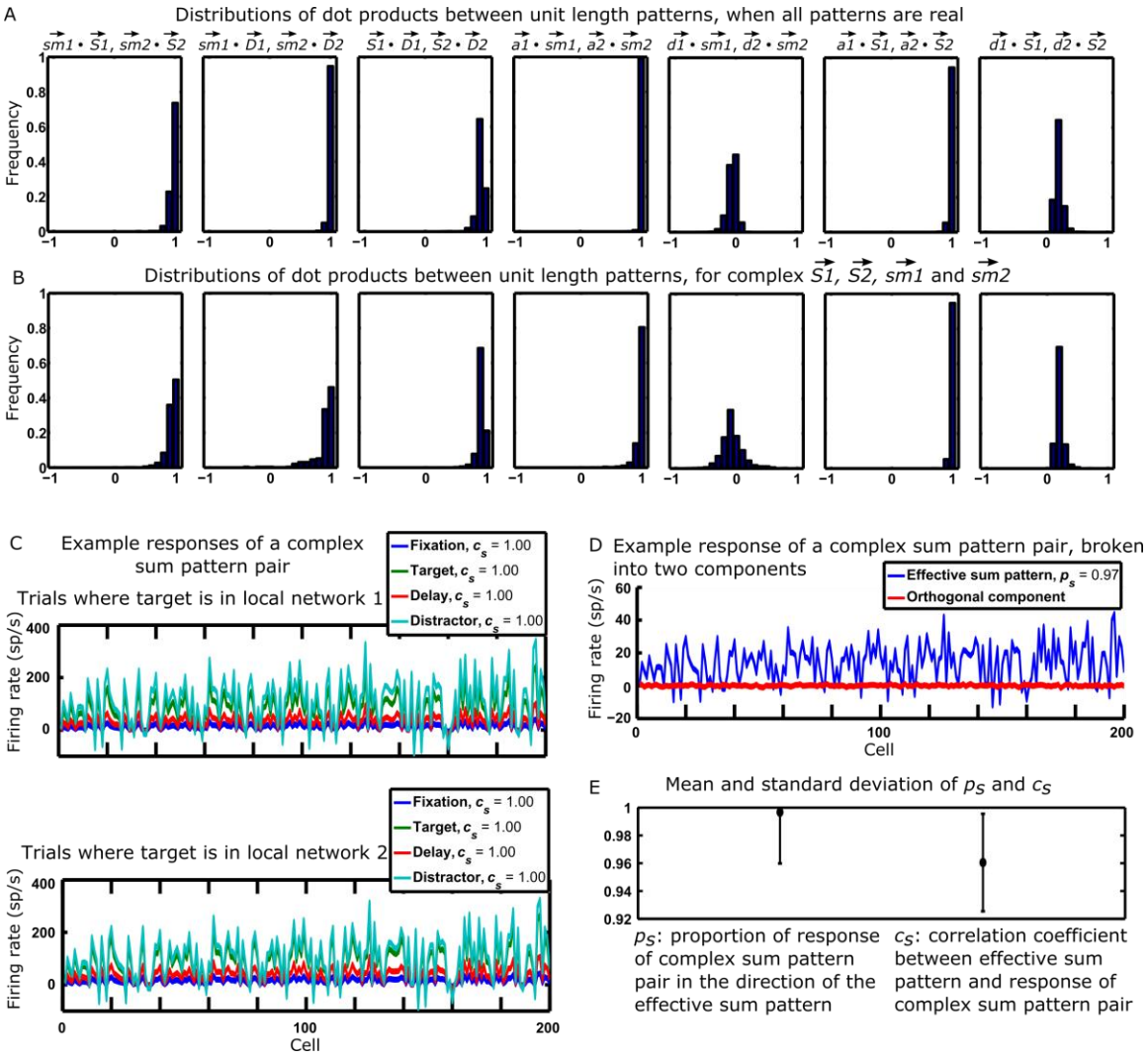

Figure S5, related to Figure 3

Comparisons of the directions of dominant activity patterns, and demonstrations of the equivalence of complex sum pattern pairs with single real sum patterns.

(A) Distributions of dot products over 1000 random instantiations of weight matrices where the vectors in the dot products are real. In the following definitions, *x* can be 1 or 2 to specify LN1 or LN2. $\overrightarrow{smx}$: the slow mode of an LN if it were not connected to the other LN. $\overrightarrow{Sx}$ (or $\overrightarrow{Dx}$): the portion of the global sum (or difference) pattern restricted to cells of a single LN. $\overrightarrow{ax}$ (or $\overrightarrow{dx}$): the average (or difference) of $\overrightarrow{Sx}$ and $\overrightarrow{Dx}$. All patterns are normalized to have unit vector length. The overall sign of each $\overrightarrow{Sx}$, $\overrightarrow{Dx}$, and $\overrightarrow{smx}$ vector is

defined such that the mean of the vector is positive. Note that, for a given LN *x*, $\overrightarrow{Sx}$, $\overrightarrow{Dx}$, and $\overrightarrow{smx}$ are all very similar to one another; $\overrightarrow{ax}$ and $\overrightarrow{smx}$ are virtually identical; and $\overrightarrow{dx}$ is roughly orthogonal to $\overrightarrow{smx}$. Parameters of the weight matrices are given in the SI section 2.2.

(B) Same as A, but over 1000 random instantiations of weight matrices where at least one of $\overrightarrow{S1}$, $\overrightarrow{S2}$, $\overrightarrow{sm1}$, and $\overrightarrow{sm2}$ is a pair of complex patterns. Only dot products involving at least one of these complex patterns went into the distributions here. For each complex pattern pair, we calculate the effective real pattern as the steady-state response of the complex pair to uniform input across cells (i.e., a vector of all ones, for reasons described in SI section 6), normalized to unit length. The effective real patterns are then used to calculate the dot products.

(C) Example responses of a complex sum pattern pair to the eight different inputs in the task. c_s is calculated from each response as the correlation coefficient between it and the effective sum pattern. High firing rates in the target and distractor responses result from hypothetically sustaining the strong visual input to let the responses reach steady state.

(D) Example response of a complex sum pattern pair to fixation input, broken into response in the effective sum pattern and response in the orthogonal direction. p_s is calculated as the proportion of the total response in the direction of the effective sum pattern.

(E) Means and standard deviations of c_s and p_s, calculated from 8000 responses (8 responses for each of 1000 weight matrices) of complex sum pattern pairs.

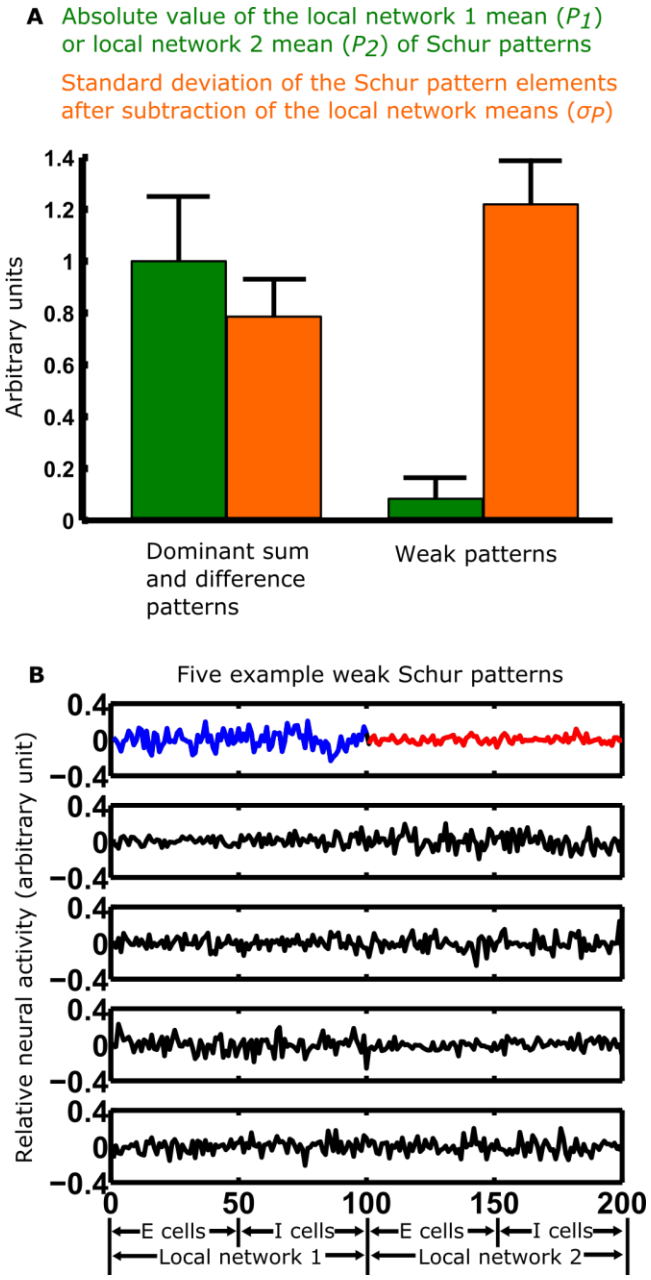

Figure S6, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"

Weak patterns are not driven strongly by the mean input to an LN.

(A) 100 global networks were generated. For each Schur pattern of each global network, we examined its two LN portions: the elements corresponding to LN1 and the elements corresponding to LN2. For each

portion we calculated its mean (P_1 or P_2). The green bars plot the mean and standard deviation of the absolute values of all P_1 and P_2, for all the dominant patterns, and for all the weak patterns. For example, the absolute values of the mean over the blue portion and the mean over the red portion of the first example Schur pattern in B are two numbers that went into the green bar for weak patterns plotted here. For each Schur pattern we also calculated σ_P, the standard deviation over the elements after P_1 and P_2 are subtracted from the respective LN portions. The orange bars plot the mean and standard deviation of all such σ_P for all the dominant patterns, and for all the weak patterns. The small P_1 and P_2 relative to σ_P for the weak patterns compared to the dominant patterns mean that the weak patterns are not strongly driven by the mean input to an LN, but are instead driven by the fluctuations across cells around the mean input.

(B) Five example weak patterns. Note that they represent "random" activation of the neurons (i.e. some neurons increase firing and others decrease firing), unlike the sum and difference patterns (Fig. 3B) which represent concerted activation of most neurons of the same LN (i.e. either all increases firing or all decreases firing).

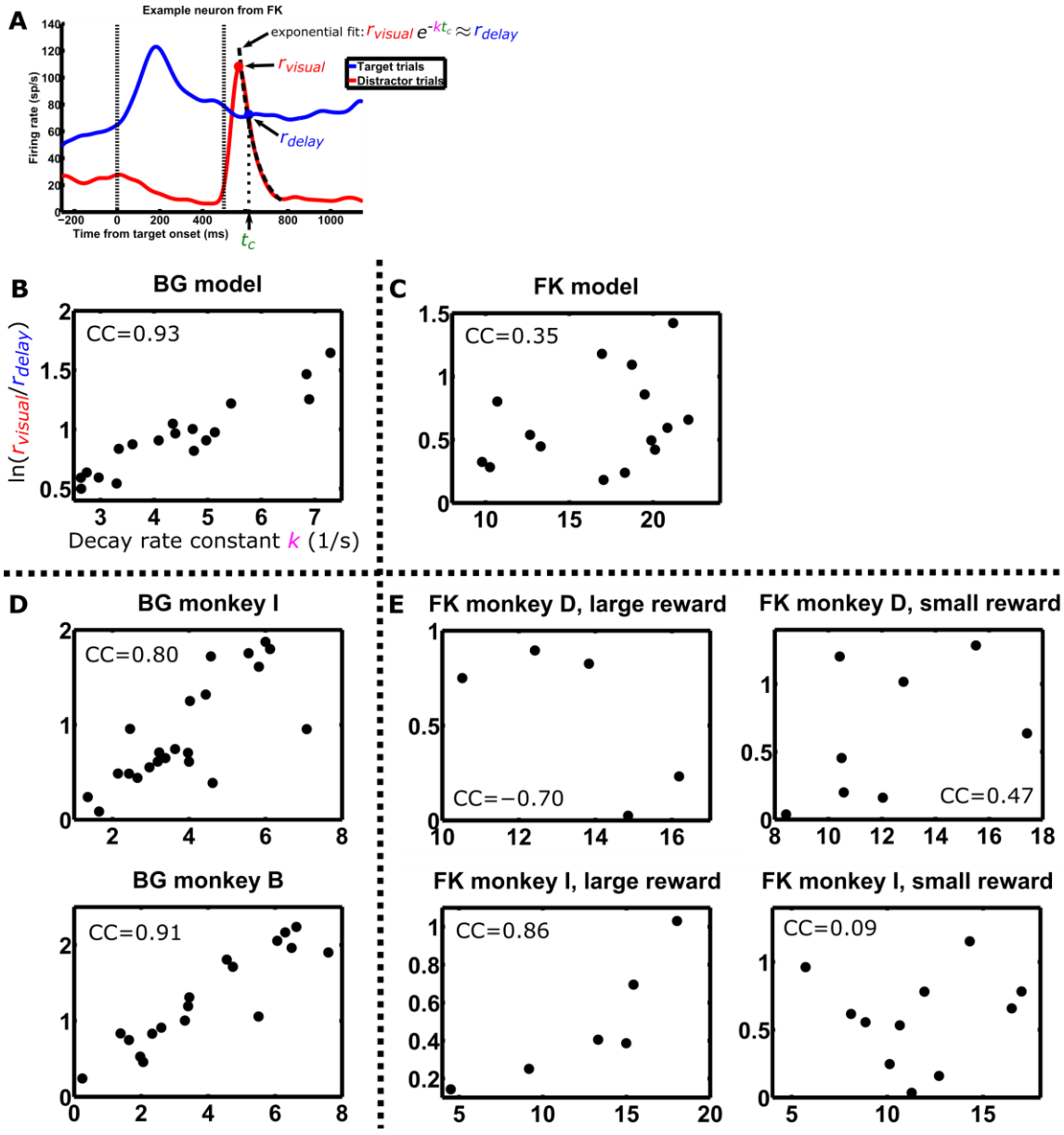

Figure S7, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"

The crossing dynamics of single neurons. Here 'crossing' refers to a neuron's visual response to a distractor, after rising to a peak above the level of its own delay activity, crossing that delay activity level as it decreases, as illustrated in (A) and described in the Results and SI section 7. This analysis follows

Bisley and Goldberg (2006) and Ganguli et al. (2008).

(A) The quantities relevant to the crossing dynamics, illustrated for one example neuron. The decay of the distractor visual response is fit with an exponential function: the peak visual response, r_{visual}, decays exponentially with time constant k, and crosses the delay activity, r_{delay}, at the crossing time, t_c. Single neuron PSTHs plotted with the same conventions as Fig. 1D.

(B-E) $ln\ (\ r_{visual} / r_{delay}\ )$ is plotted against k for BG and FK model and data. Each dot is a single neuron, where the plotted quantities are measured as illustrated in A. Only cells that had a crossing, meaning $r_{visual} > r_{delay}$, are included. Rearranging the equation in A gives $ln\ (\ r_{visual} / r_{delay}\ ) \approx t_c\ k$; thus, the slope of the line connecting each dot to the origin is t_c, the crossing point of that neuron. When $ln\ (\ r_{visual} / r_{delay}\ )$ and k are highly correlated as in the BG model and data (B and D), the slopes are similar, meaning that single neurons have similar crossing times. $ln\ (\ r_{visual} / r_{delay}\ )$ and k are less correlated in the FK model and data (C and E), indicating that single neuron crossing times are more variable. D is replotted from Fig. 1E-F of Ganguli et al. (2008). One of the FK monkeys has too few cells (1 cell for large reward, 3 cells for small reward) and is not included in E. FK had large and small reward conditions, which have different levels of both visual and delay activity and so different crossing times (Fig. S2A-B), therefore we have plotted the conditions separately. In fact, from Fig. S2A, for large reward the average distractor activity never reaches the level of the average delay activity, meaning that there is no population crossing time in this case. For completeness we nonetheless show those cells that showed a crossing in their individual activities for the large reward case.

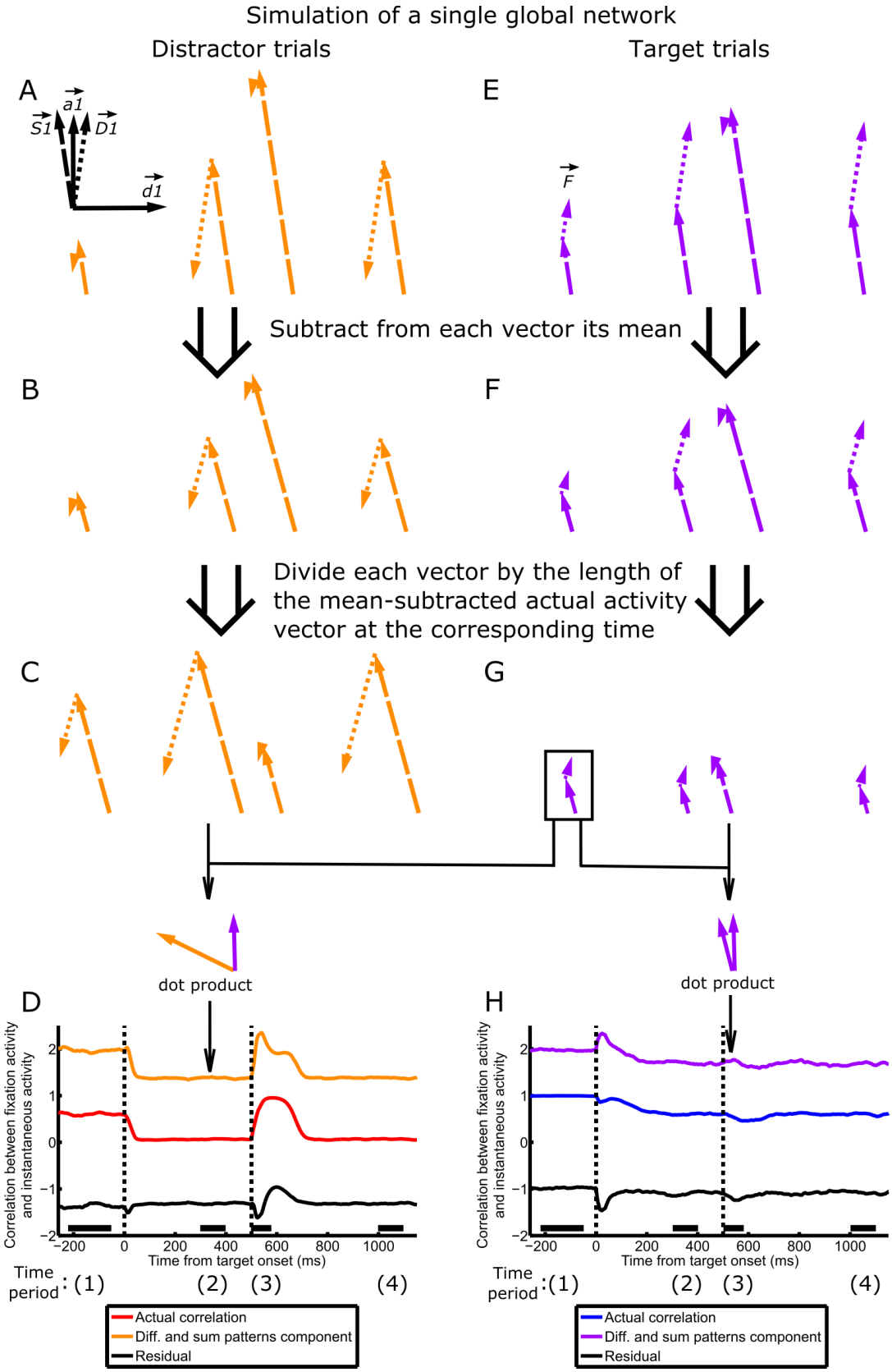
Simulation of a single global network
Distractor trials
Target trials
A
a1
S1
D1
d1
E
F
Subtract from each vector its mean
B
F
Divide each vector by the length of
the mean-subtracted actual activity
vector at the corresponding time
C
G
dot product
dot product
D
H
Correlation between fixation activity and instantaneous activity
Time from target onset (ms)
Time period: (1) (2) (3) (4)
Actual correlation
Diff. and sum patterns component
Residual

Figure S8, related to Figure 4

Details of the relationship between $\overrightarrow{S1}$ and $\overrightarrow{D1}$ and the correlation between fixation and instantaneous activity during a given time period. A-D are distractor trials; E-H are target trials.

(A inset) The two-dimensional space spanned by the two dominant activity patterns of one LN, $\overrightarrow{S1}$ (dashed vector) and $\overrightarrow{D1}$ (dotted vector). Replotted from Fig. 4C inset.

(A and E) The evolution of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities. For each trial type, $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities are each averaged over each of four time periods (spanned by black bars in D and H), and are illustrated in their two-dimensional subspace, where the relative lengths of and the angle between $\overrightarrow{S1}$ and $\overrightarrow{D1}$ activities are preserved and accurately rendered. The $\overrightarrow{S1}$ and $\overrightarrow{D1}$ components of $\vec{F}$, the vector of target trial fixation activities, are labeled in E. Replotted from Fig. 4C and G.

(B and F) For each vector in A and E, its mean was subtracted. The resulting mean-subtracted vectors are illustrated in their two-dimensional subspace. Note that the scales of A and E and of B and F are different.

(C and G) Each vector in B and F is normalized by the length of the mean-subtracted actual activity vector at its respective time. Note that B, F, C, and G share the same space and scale. To calculate $Corr_{sum,diff}$ (the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ component of the correlation coefficient between instantaneous and fixation activity) at a given time period and for a given trial type, first add the two vectors derived from $\overrightarrow{S1}$ and $\overrightarrow{D1}$ for that time and trial type, and likewise add the two vectors for the fixation period on target trials (boxed in G). Then, $Corr_{sum,diff}$ at that time and on that trial type is the dot product between the two resultant vectors (illustrated for the second time period on distractor trials and the third time period on target trials).

(D and H) Actual correlation (red/blue), $Corr_{sum,diff}$ (orange/purple, the component of correlation due to the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ patterns alone), and $Corr_{residual}$ (black, the residual component) on distractor/target (D/H) trials. The orange and black traces add up to the red trace, and the purple and black traces add up to the blue trace. See Results for how the correlation was broken down into the two components. Replotted from Fig. 4B and F. Note that the actual correlation, but not $Corr_{sum,diff}$ or $Corr_{residual}$, is restricted to lie within -1 and 1.

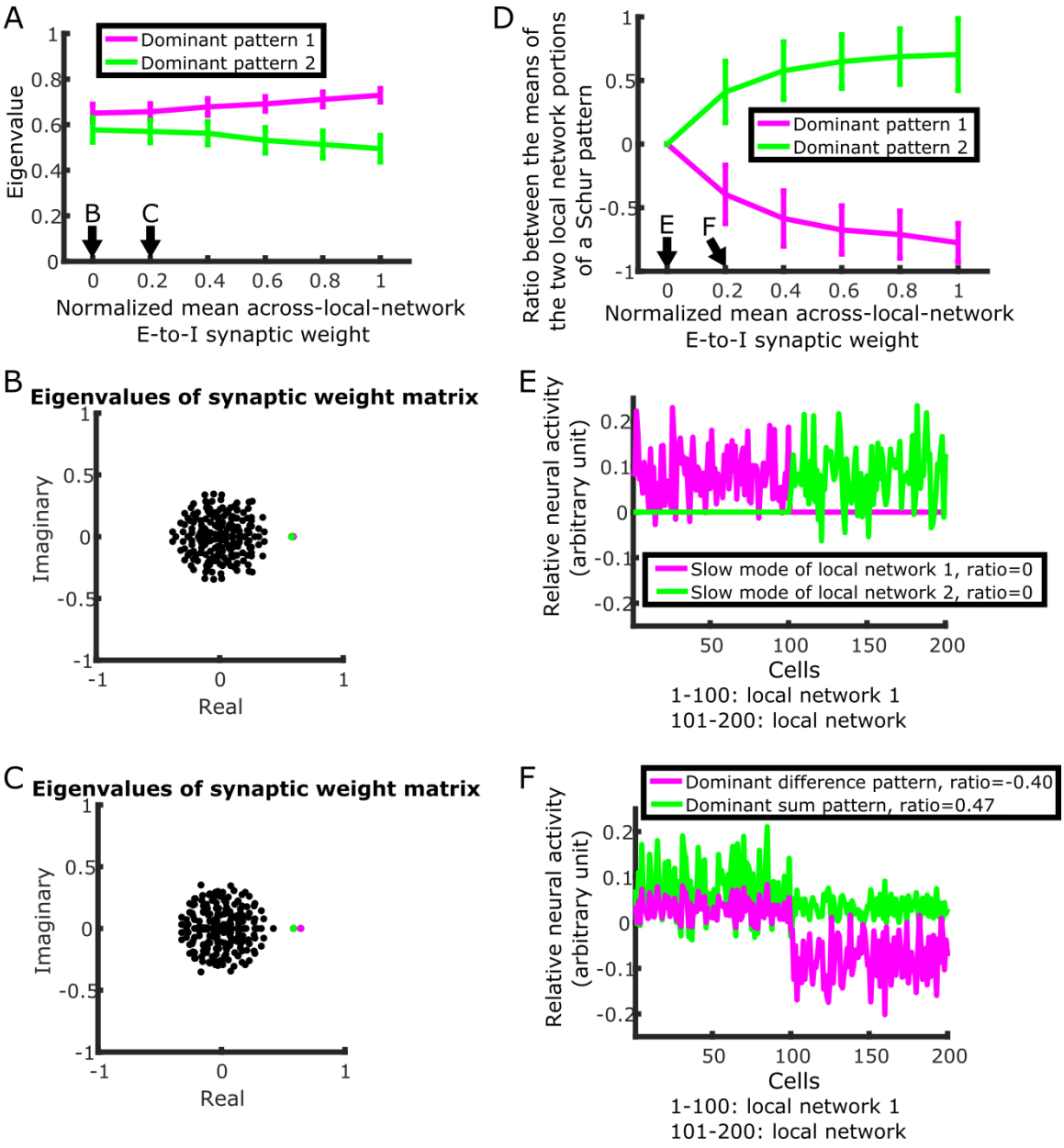

Figure S9, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"

Independent slow modes gradually morph into sum and difference patterns as coupling between LNs strengthens.

(A) The two leading eigenvalues of the global network as functions of the across-local-network E-to-I synaptic weights. As coupling strengthens, one eigenvalue (that of the difference pattern) increases while

the other eigenvalue (that of the sum pattern) decreases. Error bars are standard deviations across simulations (n = 100 global networks for each value of mean synaptic weight). The normalized mean weights of 0 and 1 are used in our BG and FK models, respectively (e.g. Fig. 2). Weights in-between produce intermediate levels of surround suppression (data not shown). Equations (1) and (2) in SI section 5 show the dependence of the two eigenvalues on the across-local-network weight for the mean population connectivity matrix, which agrees with the eigenvalues of actual connectivity matrices plotted here. Note that with a mean weight of zero, the difference between the two eigenvalues reflect stochastic differences between the connectivity of the LNs, instead of deterministic differences between sum and difference patterns, as is the case with nonzero mean weights.

(B-C) Representative eigenvalue spectra of networks with the different levels of coupling indicated by arrows in A.

(D) For each dominant pattern (which has 200 elements), we first calculated P_1, the mean over its LN1 portion (elements 1-100), and P_2, the mean over its LN2 portion (elements 101-200). Then we calculated the ratio between the P_1 and P_2, with the one that has the larger absolute value in the denominator, so that the ratio ranges between -1 and 1. The ratio for each of the two dominant patterns are plotted as a function of the across-local-network E-to-I synaptic weights. A ratio of 0 indicates that the pattern represents activation of one LN independent of the other LN, i.e. the slow mode in BG. A positive/negative ratio indicates common/differential activation of the two LNs, i.e. the sum/difference pattern. As coupling between the LNs strengthens, the two slow modes morph into the sum and difference patterns. Error bars are standard deviations across simulations (n = 100 global networks for each value of mean synaptic weight).

(E-F) Representative dominant patterns of networks for the different levels of coupling indicated by arrows in D.

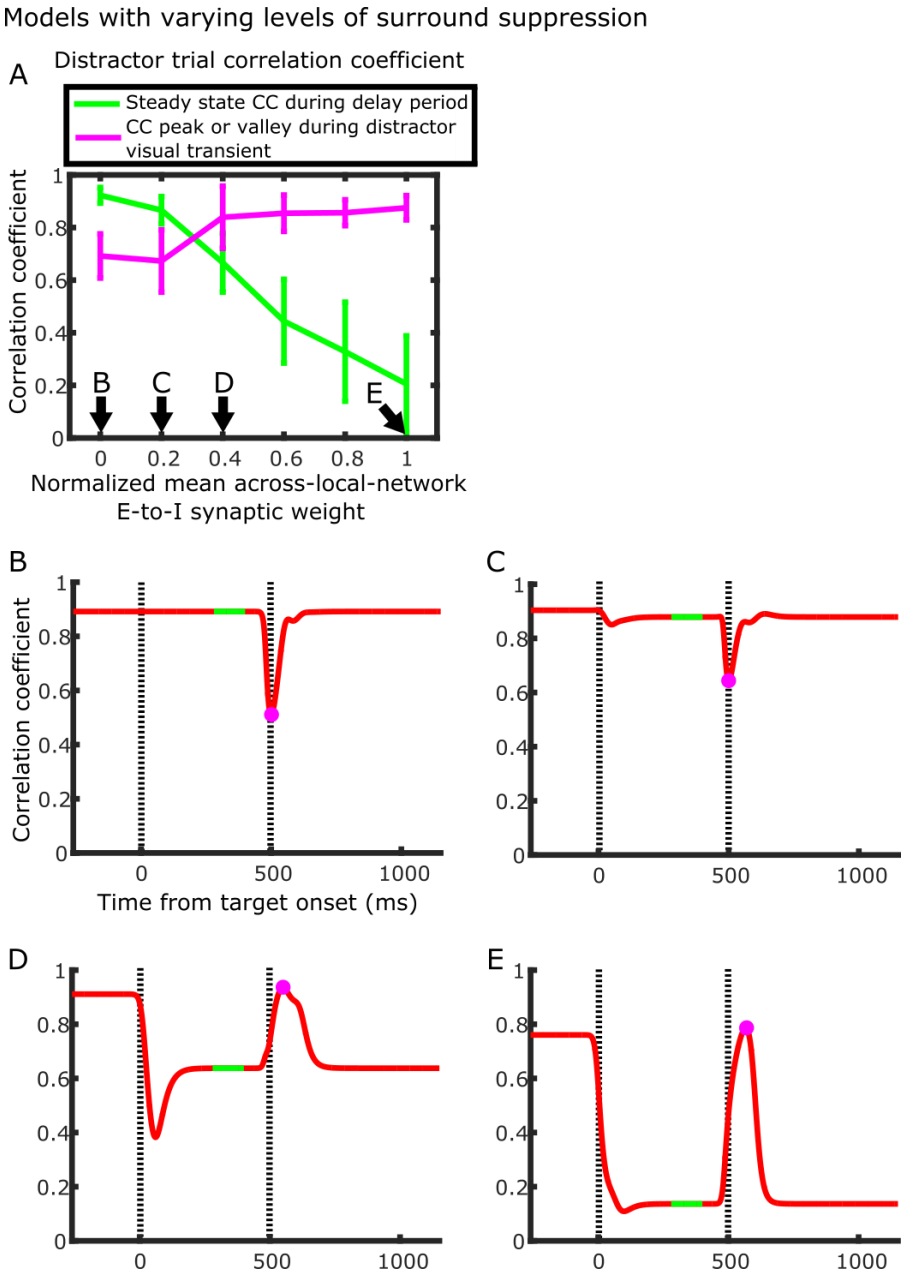

Figure S10, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"

Model predictions for the network dynamics underlying different levels of surround suppression.

(A) Two salient features (illustrated in B-E) of distractor trials correlation as functions of the mean across-local-network E-to-I synaptic weight. Normalized mean weights of 0 and 1 are the values used in our BG

and FK models, respectively (e.g., Fig. 2); intermediate weight values produce intermediate levels of surround suppression (data not shown). The delay period steady-state correlation coefficient is defined as the average correlation from 280 to 400 ms after target onset. The correlation coefficient peak/valley is defined as the maximum correlation from 500 to 600 ms when the correlation is transiently rising, or the minimum correlation from 500 to 600 ms when the correlation is transiently dropping. Error bars are standard deviations across simulations (n = 100 simulations for each value of mean synaptic weight; the parameters of each simulation are independently and randomly drawn). Note that for the mean weight of 0, the correlation coefficient valley plotted here is less deep than that in our BG model (Fig. 2E), because all simulations in this figure use the FK visual input parameters (see SI section 10 for the effects of visual input on correlations).

(B-E) Distractor trials correlations from representative simulations of networks with the different levels of coupling strength indicated by arrows in A. Green traces denote the interval over which the steady-state correlation coefficient in A is calculated, and the magenta dots denote the correlation coefficient peaks/valleys in A. Plotted with same conventions as Fig. 1F.

Supplemental Information

Section 1: Task details

At the beginning of each recording session, before task performance, both Bisley and Goldberg (BG) and Falkner, Krishna et al. (FK) isolate an LIP neuron and map out its receptive field (RF). In addition, FK map out a location in the visual field where a stimulus evokes maximum suppression.

In both studies, a monkey initiates a trial by fixating a central spot. After some time (BG: variable between 1 s and 2 s; FK: 500 ms) the saccade target appears. The target disappears after 100 ms in the BG task version, and it stays on in the FK version. After a delay (BG: 600 ms; FK: 500 ms), a task-irrelevant distractor stimulus is flashed (duration: BG, 100 ms; FK, <50 ms). After another delay (BG: variable between 700 ms and 1700 ms; FK: 550 ms), the fixation point disappears, and the monkey saccades to the target location for a reward.

In the BG version of the task, the target and the distractor, one of which is in the RF of the neuron being recorded, are in opposite visual quadrants and equidistant from the fixation point (i.e. they are at equal radii from the fixation point, and one is at a location rotated 180 degrees from the other's location). In the FK version of the task, either the target or the distractor is in the RF, and the other stimulus is at the location previously determined to elicit maximum surround suppression. On a given trial, either the target or the distractor is in the RF of the neuron being recorded. In the BG task, the two different trial types are randomly interleaved; in the FK task, the two types of trials were run in blocks.

In the BG version of the task, during the delay period between distractor presentation and fixation point disappearance, a Landolt ring (a ring with a small segment missing) and three complete rings are flashed simultaneously for 17 ms. These four stimuli are at the target and distractor locations and at the locations rotated 90 degrees about the fixation point from those two locations, so that one is in each of the visual quadrants and all are equidistant from the fixation point. The Landolt ring appeared at either the target or the distractor location. The monkey is required to detect the orientation of the Landolt ring: if the gap is on the right, the monkey needs to cancel the planned saccade and maintain fixation after the fixation point disappears; if the gap is on the left, the monkey can proceed with the planned saccade after the fixation point disappears. The rings were shown at high contrast during neural recordings; they were shown

at varying contrasts in separate psychophysical experiments to map contrast thresholds and thus the allocation of attention. In this paper, we only analyzed trials in which the rings appeared more than 700 ms after distractor onset.

In the FK version of the task, in each trial the target is one of two colors, indicating that the reward amount for that trial would be large or small.

Section 2: Modeling and analysis procedures

Section 2.1: Data analysis

To estimate the standard error of correlations between instantaneous and fixation activities from an actual population or a simulated population, we formed 1000 bootstrap sample populations by sampling cells with replacement from the given population. Then the standard error is the standard deviation of the correlations calculated from different bootstrap sample populations. The standard error of the vector lengths in Fig. S3 was computed the same way.

To perform principal component analysis, we form an N x T matrix, each row of which is the trial-averaged firing rates of one cell at T time points. The times include all millisecond time points during distractor trials except the times of the visual response (details in Fig. 5A-B legend). We then subtracted from each row its row mean (i.e. the mean rate of the cell across the T time points), obtaining a matrix $\boldsymbol{R}$. The PCs are the eigenvectors of the N x N matrix $\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}}$, and the proportion of variance explained by a PC is its corresponding eigenvalue divided by the sum of all eigenvalues.

Section 2.2: Modeling details

The model network consists of two LNs of N neurons each (N/2 E cells and N/2 I cells). We included I cells, unlike the E-cells-only model of Ganguli et al. (2008), because we aimed to model surround suppression. We chose to model equal numbers of E and I cells for simplicity, but modeling more realistic ratios of the number of E and I cells does not change our results (data not shown). Within an LN, the mean excitatory and inhibitory synaptic weights, onto both E and I cells, are $\frac{a}{N/2}$ and $-\frac{b}{N/2}$, respectively. We choose $a > 1$ and $a - b < 1$, such that each LN operates as an inhibition-stabilized

network, a network regime underlying surround suppression in V1 (Ozeki et al., 2009; Rubin et al., 2015). Furthermore, $a > b$, so that each LN strongly amplifies a pattern of increased activity across neurons. The mean synaptic weight of excitatory projections from the E cells of each LN to the I cells of the other LN is $\frac{c}{N/2}$: $c = 0$ for the BG model network, and $c > 0$ for the FK model network. We model sparse and random connectivity: a small fraction p of the weights are non-zero, and each non-zero weight is independently drawn from a normal distribution with mean $\frac{x}{pN/2}$ and standard deviation $\frac{x}{2pN}$, where $x = a$ for local excitatory synapses, $x = -b$ for local inhibitory synapses, and $x = c$ for across-network excitatory synapses. We have chosen the standard deviations of the weight distributions to be small enough that we have not observed weights that violate Dale's Law; if observed, such weights would be set to zero.

We model the dynamics of the neurons with the following linear differential equation:

$$\boldsymbol{T}\frac{d\vec{r}}{dt} = -\vec{r}(t) + \boldsymbol{W}\vec{r}(t) + \vec{I}(t)$$

where $\boldsymbol{T}$ is a diagonal matrix of the time constants of the neurons (normally distributed with mean τ and standard deviation τ/k; again, negative time constants were not observed, but would be set to 1 ms if observed), $\vec{r}$ is a vector of the activity of the neurons, $\boldsymbol{W}$ is the synaptic weight matrix, and $\vec{I}$ is a vector of the input to the neurons from areas outside LIP. For each trial type, the initial condition is the steady state response to the deterministic part of the input during the fixation period on the respective trial type (i.e. $\vec{I}_{determ.}$; see below). Negative firing rates are not allowed and are rectified to zero (in our simulations, firing rates generally stay positive and do not reach zero). This is a standard phenomenological firing rate model that can be derived as an approximation to biophysically realistic spiking models (Dayan and Abbott, 2005). These dynamics are taken to be modeling trial-averaged firing rates, as we have no knowledge of the single-trial population dynamics during our tasks.

The input at any time t has two components:

$$\vec{I}(t) = \vec{I}_{determ.}(t) + \vec{I}_{noise}(t)$$

where $\vec{I}_{determ.}(t)$ is the deterministic input, and $\vec{I}_{noise}(t)$ is the noise.

At a given time t, each element of $\vec{I}_{determ.}(t)$ is the sum of one or more of the four types of input described in the Results. For each of the four input types, the input to each cell is independently drawn

from a uniform distribution, with range of the distribution picked to qualitatively fit the experimentally observed firing rates. The range parameters of the uniform distribution for fixation input are: (I_{F1}, I_{F2}); transient visual input: (I_{V1}, I_{V2}); sustained visual input: (I_{V1}', I_{V2}'); delay input: (I_{D1}, I_{D2}); expectation input: (I_{E1}, I_{E2}). For a given cell, its fixation inputs on the two trial types are the same, and so are its transient visual inputs. The transient visual input lasts for 100 ms for the BG model and 40 ms for the FK model. The onset of delay input, as well as the sustained visual input in the FK model, is at the offset of the transient visual input evoked by a target. For simplicity, we model inputs with instantaneous onset, e.g., visual input is turned on to full strength at the onset of a visual stimulus. The instantaneous onset of visual input results in the more rapid drop in correlation following target and distractor onset in the BG model compared to the BG data (Fig. 2E and Fig. 1E). If we let inputs increase gradually to their full strength, our BG model can reproduce the slower rate of correlation drop (data not shown).

$\vec{I}_{noise}(t)$ is calculated as follows:

$$\vec{I}_{noise}(t) = v\vec{I}_{noise}(t - \Delta t) + \vec{I}_{random}(t)$$

v is a parameter between 0 and 1, which determines how much the noise is temporally correlated; Δt = 1 ms is the discrete time step used in our numerical simulations; $\vec{I}_{random}(t)$ is the new noise at time *t*, each element of which is independently drawn at each time step from a normal distribution with zero mean and standard deviation equal to a fraction *z* times the corresponding element in $\vec{I}_{determ.}(t)$.

The inherited surround suppression model is identical to the FK model except in two ways. First, the two LNs are unconnected. Second, whenever one LN receives visual or delay external input, the mean external input to the other LN is reduced by an amount proportional to the mean visual or delay input: the decrease in input to each cell at time *t* is independently picked from a uniform distribution, whose mean is a fraction *u* of the mean visual and/or delay input at time *t* to the activated LN, and whose range is from 0 to twice its mean.

To simulate the experiments, the simulation was run multiple times (41 times for the BG simulation and 27 times for the FK simulation), each time with random instantiations of connectivity matrices, neuronal time constants, and inputs. One cell is randomly picked from each simulation to form populations the same sizes as the experimental populations.

The model parameters are: $N = 100$, $a = 1.1$, $b = 0.5$, $c = 0.15$, $p = 0.2$, $\tau = 10$, $k = 10/3$, $I_{F1} = 4$, $I_{F2} = 6$, $I_{V1} = 30$ (BG) or 60 (FK), $I_{V2} = 160$ (BG) or 130 (FK), $I_{V1}' = 2$, $I_{V2}' = 4$, $I_{D1} = 5$, $I_{D2} = 65$, $I_{E1} = 2$, $I_{E2} = 10$, $v = 0.97$, $z = 1/30$, $u = 1/30$. In the model, firing rates are in units of sp/s and time in units of ms. The ranges of external inputs were chosen to be consistent with firing rates in the respective top-down and bottom-up areas and to roughly match the simulated LIP firing rates to the data. Note that these parameters were not fine-tuned to quantitatively reproduce the data; our model is robust and can qualitatively reproduce the data with a range of parameters.

Section 3: Implications of different mechanisms of persistent activity for two-dimensional dynamics

In both our model and that of Ganguli et al. (2008), LIP persistent activity during the delay period results from sustained top-down input from prefrontal cortex. This is a simplifying assumption, made because the focus of both studies was on the recurrent interactions within LIP. Possibilities for the actual mechanisms behind LIP persistent activity were discussed by Ganguli et al. (2008) in their Discussion. As they discussed in more detail, LIP is not likely to have attractor dynamics and sustain persistent activity by itself, since in the BG task, the strong visual response to the distractor is not able to trigger persistent activity. Therefore, they suggested ways whereby different oculomotor areas (LIP, FEF, dlPFC, SC, etc.) can recurrently interact with each other to generate distractor-resistant persistent activity in each area. One possibility is that each area acts as a "leaky attractor," but the areas recurrently excite each other to balance out the leak so that each area has persistent activity. Or, one area might be able to produce persistent activity by itself, but needs transient "gating" signals from other areas to be able to ignore distractors.

Because we do not have detailed knowledge of the connectivity between LIP and PFC, nor knowledge of the activity patterns across PFC neurons on the tasks we studied, attempts to include the recurrent interactions between LIP and PFC in our BG or FK models would be very under-constrained. However, we note that if persistent delay activity in LIP is generated through recurrent interaction with PFC, the conclusions of our study do not change. The dynamics of an LIP LN is still dominated by a small number of dominant patterns, but interaction with PFC effectively modulates the strength of self-excitation of the LIP dominant patterns, possibly allowing them to be persistently active or decay based on the

requirements of the task.

Section 4: Analysis of feedforward connections in the Schur form of the connectivity matrix

One common way to examine the influence of the connectivity of a network on its dynamics is through determining the eigenvectors and eigenvalues of the connectivity matrix. The eigenvectors are a set of activity patterns that each excite or inhibit itself but not any of the other patterns. Thus, in a linear model these patterns evolve independently: each evolves according to its own self-connection, independent of the other patterns. The strength of self-connection of each eigenvector is given by the real part of its corresponding eigenvalue, and so one may expect the eigenvectors whose eigenvalues have the largest real part to dominate the activity of the network.

However, for biological connection matrices composed of separate excitatory and inhibitory neurons, the eigenvectors are not orthogonal (Murphy and Miller, 2009), meaning for example that two eigenvectors with large amplitude can cancel, resulting in small overall activity. These cancellations and related effects can make it difficult to understand neural activities from the independent dynamics of the eigenvectors. Instead, it can be more illuminating to analyze the Schur patterns: an ordered set of *orthogonal* activity patterns derived by orthogonalizing the eigenvectors (Murphy and Miller, 2009; Goldman, 2009). For a given connectivity matrix, there are different sets of Schur patterns, obtained by orthogonalizing the eigenvectors in different orders. For our purpose of finding the dominant activity patterns, we choose the set of Schur patterns that are ordered by their strength of self-connections, from the most self-excitatory to the most self-inhibitory. The self-connections are examined in the main text and in SI section 5; here we examine the rest of the connections between the Schur patterns, a set of purely feedforward connections.

To understand the structure of feedforward connections in our connectivity matrices, we first examine the mean population connectivity matrix. This is a 4-by-4 matrix, whose rows and columns denote the excitatory (E) and inhibitory (I) populations of the two LNs, and whose elements are the mean connection strengths between them multiplied by $N/2$ (the number of E or I neurons in each LN). Fig. S4A plots an example mean population connectivity matrix. The four rows/columns denote: the E population of

LN1, the I population of LN1, the E population of LN2, and the I population of LN2. Each row shows the input weight to the given population from each of the four populations, while each column shows the projection weight from the given population to each of the four populations. Fig. S4B plots the Schur form of this matrix, which shows the connections between the Schur activity patterns or basis vectors (each representing a pattern of activity across the four populations). It shows that in addition to self-connections (non-zero entries on the diagonal, which are the eigenvalues associated with the patterns), there are feedforward connections from activity pattern 3 to pattern 2, and from pattern 4 to pattern 1 (non-zero entries on the upper triangle). What are these activity patterns? Fig. S4C plots the Schur basis vectors. To describe these we will introduce the following terminology. By global sum or difference we mean that the activity patterns of the two LNs are the same or opposite, respectively. By local sum or difference we mean that the activities of the E and I populations within an LN are the same or opposite, respectively. We can see that patterns 1 to 4 represent: global difference with local sum, global sum with local sum, global sum with local difference, and global difference with local difference. The connections from pattern 3 to pattern 2 and from pattern 4 to pattern 1 thus represent local difference patterns feeding into local sum patterns, a manifestation of balanced amplification, which we investigated in Murphy and Miller (2009).

The dominant activity patterns of the mean population connectivity matrix are patterns 1 and 2 (corresponding to the sum and difference patterns discussed in the main text), because they are amplified both by strong self-excitation and by receipt of feedforward excitation. Does this structure also hold for the actual connectivity matrix, in which each population consists of many neurons, with weights between neurons chosen stochastically? We analyze one actual connectivity matrix, the one examined in Fig. 3 of the main text. In Fig. S4D-E, we plot the actual connectivity matrix and its real-valued Schur form. As we have seen in the main paper, the two most strongly self-excitatory patterns of the actual connectivity matrix (plotted in Fig. 3B and again in Fig. S4F) are still the patterns of global difference with local sum and global sum with local sum, as predicted by the mean population connectivity matrix. We will refer to them here as the dominant difference and dominant sum patterns. The two weaker patterns of the mean population connectivity matrix—the patterns of global sum with local difference and global difference with local difference—are dispersed in the many weakly self-excitatory patterns that are a manifestation of the

sparse and random connectivity of the actual connectivity matrix; the feedforward structure of these patterns to the two dominant patterns are hidden, but unchanged. We can reveal the feedforward structure to the dominant difference or sum pattern by summing the less self-excitatory Schur basis vectors (i.e., all of the patterns except the dominant difference and sum patterns), each weighted by its feedforward weight to the dominant difference or sum pattern, respectively. The resulting weighted sums are a pattern of global difference with local difference, which feeds into the dominant difference pattern, and a pattern of global sum with local difference, which feeds into the dominant sum pattern, just as predicted by the mean population connectivity matrix (Fig. S4F). Furthermore, a comparison of the magnitudes of feedforward weights show that the only strong feedforward connections are those from the less self-excitatory patterns to the two dominant patterns; in particular, the feedforward connections from the dominant sum pattern to the dominant difference pattern is very weak, making these two dominant patterns essentially independent (Fig. S4G). Thus, based on the structure of the weight matrix, we can see that the difference and sum patterns would dominate the dynamics of the network.

Section 5: The eigenvalues of the sum and difference patterns

Here we calculate the eigenvalues of the mean population connectivity matrix examined in the last section (e.g. Fig. S4A). In this matrix,

$$\begin{pmatrix} a & -b & 0 & 0 \\ a & -b & c & 0 \\ 0 & 0 & a & -b \\ c & 0 & a & -b \end{pmatrix}$$

a is the strength of the E weights and *-b* the I weights within an LN, and *c* is the weight of the between-network E-to-I connections that mediate surround suppression; *a*, *b*, and *c* are all positive. The eigenvalues of this matrix are, from the most positive to the most negative,

$$\lambda_D = \frac{1}{2}\left(a - b + \sqrt{(a-b)^2 + 4bc}\right) \quad (1)$$

$$\lambda_S = \frac{1}{2}\left(a - b + \sqrt{(a-b)^2 - 4bc}\right) \quad (2)$$

$$\lambda_3 = \frac{1}{2}\left(a - b - \sqrt{(a-b)^2 - 4bc}\right)$$

$$\lambda_4 = \frac{1}{2}\left(a - b - \sqrt{(a-b)^2 + 4bc}\right)$$

Each LN by itself has a slow mode when its recurrent excitation dominates recurrent inhibition (i.e. $a > b$). When the two LNs are uncoupled (i.e. $c = 0$, the BG case), λ_D and λ_S are equal and are the slow mode eigenvalues of the independent LNs, while λ_3 and λ_4 are zero. The weak suppressive coupling between the two LNs in the FK case (small, positive c) perturbs these eigenvalues. λ_D and λ_S remain large and positive, and become the eigenvalues of the difference and sum patterns, respectively, while λ_3 and λ_4 remain close to zero, lying to either side of zero and separated by the same distance that separates λ_D from λ_S. Because the LNs mutually suppress each other, $\lambda_D > \lambda_S$, i.e. the difference pattern is more strongly amplified than the sum pattern by the connectivity. If we then expand the matrix to $N/2$ excitatory and $N/2$ inhibitory neurons in each LN (so the matrix is $2N$ x $2N$), with weights uniformly $a/(N/2)$, $b/(N/2)$, $c/(N/2)$, and 0 in the blocks that replace each a, b, c, and 0 respectively of the 4 x 4 matrix, then the matrix has four Schur vectors similar to those of Fig. S4C (elements over each set of $N/2$ E or I neurons within an LN are uniform), and with the four eigenvalues as given above; and $2N$-4 Schur vectors that have eigenvalue 0, which are orthogonal to the first four and so sum to zero over each set of $N/2$ E or I neurons within each LN. When this matrix is then replaced with the actual connectivity matrix, which has sparse random connectivity within each nonzero $N/2$ x $N/2$ block with the same mean connection strength (e.g. Fig. S4D), the two dominant eigenvalues remain close to λ_D and λ_S respectively, with Schur vectors having mean values over each set of $N/2$ E or I neurons within an LN similar to those of the previous sum and difference Schur vectors; while the two weaker patterns associated with near-zero eigenvalues λ_3 and λ_4 and the remaining patterns with zero eigenvalues are dispersed among the many weak patterns of the actual connectivity matrix (Fig. S4F) A similar process was described in more detail for all-excitatory connectivity of a single LN in the Supplemental Materials of Ganguli et al. (2008).

If we model the mean population connectivity matrix with more parameters (e.g., separate weight parameters for the E-to-E, E-to-I, I-to-E, and I-to-I connections within an LN, and additional across-local-network E-to-E connections), our formulas for the eigenvalues would become much more complex, but the simple intuition presented above do not change. With parameters of within-local-network weights that result in the isolated LN having a slow mode, the global network would have two and only two dominant

patterns. The addition of weak across-local-network mean weights, which are consistent with the weak suppression observed by FK and with the fact that cortical connection density decreases with distance (Markov et al., 2011), as well as the transition from uniform connectivity to sparse random connectivity, act as relatively small perturbations, and the sum and difference patterns remain the only two dominant patterns.

Section 6: Equivalence of complex sum pattern pairs with single real sum patterns

With the connectivity parameters in the main text, in a small proportion of random instantiations of connectivity matrices, two complex patterns (which are complex conjugates in the eigenvector basis) take the place of the single real global sum pattern. When recurrent excitation is sufficiently stronger than inhibition, all random instantiations of connectivity matrices have real sum patterns, and when excitation is weaker (while still being stronger than inhibition, ensuring the existence of slow modes), complex sum pattern pairs are more frequent. Similarly, the slow mode of an isolated LN can also be a complex pattern pair.

A complex conjugate pair introduces two slowly-decaying patterns of neural activation in place of the single pattern corresponding to a real sum pattern or a real slow mode. However, our analysis remains unchanged, because activation of a complex conjugate pair in response to our various input patterns is very largely confined to a single dimension, which we call the effective sum pattern or the effective slow mode. We define the effective sum pattern for a complex sum pattern pair (or effective slow mode for a complex slow mode pair) to be the steady-state response of the complex pattern pair to a uniform input across cells of the network (i.e., a vector of $2N$ ones for the sum pattern pair or N ones for the slow mode pair), normalized to a vector length of one. The near complete overlap of dot product distributions calculated with real patterns and effective patterns (Fig. S5A and B) shows that the effective patterns would behave the same way as the real patterns analyzed in the main text.

In response to inputs used to simulate the experiments, the response of complex sum pattern pairs or complex slow mode pairs corresponds almost perfectly to their effective patterns. To illustrate this, we simulated 8000 such responses for complex sum patterns (for each of 1000 weight matrices, 8 responses

were calculated, see Fig. S5C-D; responses for complex slow mode pairs were entirely similar) and used two metrics to quantify their resemblance to effective sum patterns. For each response, we calculate c_s, the correlation coefficient between the effective sum pattern and the response of the complex sum pair, and p_s, the proportion of the total response of the complex sum pattern pair in the direction of the effective sum pattern (equal to the dot product of the response of the complex sum pair with the effective sum pattern, each normalized to unit vector length). Fig. S5E shows that c_s and p_s are indeed very high, demonstrating the equivalence of complex sum pattern pairs with single real sum patterns. Simulations of networks with complex pattern pairs show similar firing rates and correlation patterns as Fig. 2 (data not shown), further confirming the equivalence.

Section 7: The consequences of low-dimensional dynamics for attentional switching

As described in the Results section "One-dimensional dynamics in LIP," BG found that a monkey's attention switched from the target location to the distractor location upon presentation of the distractor, and then switched back to the target location at an attentional switching time that coincided with the time at which the LIP population mean response to the distractor crossed below that to the target. Furthermore the crossing times of single neurons coincided with this population crossing time. LIP single neurons having a common crossing time depends on one-dimensional dynamics: the slowly-decaying population visual response to the distractor and the population delay activity are both in the same dimension (Ganguli et al., 2008).

In this section, we first examine the factors that determine the crossing time of the decaying distractor visual response and delay activity in FK, then examine the crossing of single neurons in both model and data.

The common crossing time of single neurons in BG can be explained by the one-dimensionality of LIP local dynamics around the time of the crossing (Ganguli et al., 2008). In state space, the multi-neuronal delay activity is a point on the one-dimensional line which is the direction of the slow mode, and the multi-neuronal visual response moves on this line towards the delay activity point as it decays. At the time that the multi-neuronal visual response meets the delay activity, the visual response of each neuron is equal to

its delay activity, and thus this is the common crossing time.

In FK, the dynamics of an LIP LN are dominated by two activity patterns, the sum and difference patterns. If the inputs to the two LNs were exactly interchanged between trials that were target trials or distractor trials for LN1, and the two LNs had identical connectivity (i.e., if the two LNs and their inputs were perfectly symmetric), then the activation of the sum pattern as a function of time would be exactly the same on target and distractor trials of LN1, while that of the difference pattern would be exactly opposite. In this ideal case, after distractor offset, the decaying visual response on distractor trials and the delay activity on target trials only differ in their difference pattern activity, and so the crossing time is the time when the difference pattern activity is zero.

We can write the dynamics of the difference pattern activity as:

$$\tau \frac{d}{dt} r_{diff} = -r_{diff} + \lambda_{diff} r_{diff} + I_{diff} \tag{3}$$

where τ is the neuronal time constant, r_{diff} is the activation of the difference pattern, λ_{diff} is the eigenvalue of the difference pattern, and I_{diff} is the external input to the difference pattern. For a given random instantiation of a global network, λ_{diff} is close to λ_D, the difference pattern eigenvalue of the mean population connectivity matrix, calculated in equation (1) in section 5 above. The difference pattern activity on distractor trials during the decay of the visual response is given by a solution to equation (3):

$$r_{diff}(t) = r_{diff}(0)\, e^{-\frac{t}{\tau_{diff}}} + \left(1 - e^{-\frac{t}{\tau_{diff}}}\right) r_{diff}^{delay} \tag{4}$$

Here $r_{diff}(t)$ is the difference pattern activity as a function of time since the peak of the visual response (i.e. the offset of the visual stimulus occurs at $t = 0$), and $\tau_{diff} = \frac{\tau}{1-\lambda_{diff}}$ is the time constant of the difference pattern. The steady-state difference activation in the delay period is $r_{diff}^{delay} = \frac{-I_{diff}^{delay}}{1-\lambda_{diff}}$, where $-I_{diff}^{delay}$ is the input to the difference pattern during the delay period before and after visual stimulation by the distractor. For clarity we define I_{diff}^{delay} to be positive, and thus the negative sign before I_{diff}^{delay} signifies that it drives the difference pattern negatively during distractor trials. The difference pattern activation at the offset of the visual stimulus (the peak activation by the visual stimulus) is

$$r_{diff}(0) = \left[\left(1 - e^{-\frac{t_0}{\tau_{diff}}}\right)\frac{I_{diff}^{visual}}{1-\lambda_{diff}} + e^{-\frac{t_0}{\tau_{diff}}} r_{diff}^{delay}\right]$$

t_0 is the amount of time that visual stimulation was on and I_{diff}^{visual} is the input to the difference pattern during visual stimulation. Here we have assumed that the difference pattern was at its steady-state activation for delay period input, r_{diff}^{delay}, at the onset of the visual stimulus. In the case that the two LNs are perfectly symmetric, the difference pattern component of delay activity on target trials after distractor offset is simply the negative of equation (4).

In this symmetric case, the crossing time T_c, the time when the decaying distractor trials visual response and the target trials delay activity are the same, is the time when they both have zero difference pattern activity. This time is found by setting equation (4) to zero and solving for *t*:

$$T_c = \tau_{diff} \ln\left[1 + \frac{r_{diff}(0)}{-r_{diff}^{delay}}\right] = \tau_{diff} \ln\left[\left(1 + \frac{I_{diff}^{visual}}{I_{diff}^{delay}}\right)\left(1 - e^{-\frac{t_0}{\tau_{diff}}}\right)\right]$$

This shows that first, the crossing time is simply proportional to the time constant of the difference pattern. Second, this time constant is multiplied by a term that weakly (logarithmically) increases with the ratio of the peak visual activation of the difference pattern on distractor trials, $r_{diff}(0)$, to its delay period activation on target trials, $-r_{diff}^{delay}$.

The above expression for T_c depends crucially on the two LNs being symmetric, restricting two-dimensional dynamics to the single dimension of the difference pattern. However, as discussed in the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns," the stochastic components of connectivity and inputs means that the two LNs and the two trial types are not symmetric. This in turn means that sum pattern activities are not exactly the same on the two trial types, and difference pattern activities are not exactly opposite. Thus the decaying visual response and delay activity evolve in a two-dimensional space instead of one dimension, and they do not meet in general. The crossing of their mean population activities is not the crossing of their multi-neuronal activity patterns and not the common crossing of single neurons. As a result, single neuron crossing times should be considerably more variable in FK than in BG. The above expression for T_c , using the mean difference eigenvalue and inputs across network instantiations, should reasonably approximate the mean population crossing time. However, there

will be variability across network instantiations not only due to variability in the eigenvalue and inputs but also due to variability in the amplitude of the difference pattern at the time the mean population activities cross.

Now we proceed to analyze the crossing dynamics of single neurons. We follow Bisley and Goldberg (2006) and Ganguli et al. (2008) and fit (by minimizing squared error) a single neuron's decaying distractor visual response $r(t)$ from the time of the peak response, taken as $t = 0$ with peak response $r(0) = r_{visual}$, to the time the response decayed to baseline (identified as the time of minimum response in the 200 ms window after the peak response), with an exponential decay function, $r(t) = r_{visual}e^{-kt}$ (Fig. S7A). The neuron's inverse decay time constant, k, is the fit parameter. The FK responses are well-fit by exponentials (average R^2 across neurons and reward conditions: 0.97) as in the BG data (Bisley and Goldberg, 2006), and the model data was also well fit (average R^2: 0.97). Then the crossing time t_c for the neuron is defined as the time at which this exponentially decaying activity equals r_{delay}, the neuron's average delay activity: $r_{visual}e^{-kt_c} = r_{delay}$ or $ln\ (r_{visual} / r_{delay}) / k = t_c$. Thus, if we show each cell as a point in a plot of $ln\ (r_{visual} / r_{delay})$ vs. k, a given cell's crossing time can be read off as the slope of the line from the origin through the cell's point. Bisley and Goldberg (2006) and Ganguli et al. (2008) found that, although $ln\ (r_{visual} / r_{delay})$ and k each varies widely across neurons, these two quantities are highly correlated across neurons with a roughly common slope through the origin (Fig. S7D). That is, t_c is approximately the same across neurons in the BG data —there is a common crossing time. Our model of the BG data replicates this behavior (Fig. S7B), but, in accord with the above discussion, our FK model shows much weaker correlation (Fig. S7C), i.e. a lack of a common single-neuron crossing time. In accord with this model prediction, the FK data also generally shows little correlation (Fig. S7E).

In FK, the population crossing time depends on the dynamics in two activity dimensions, as opposed to one. Thus, on the hypothesis that the attentional switching time corresponds to the population crossing time, we would expect greater variability in the attentional switching time in the FK case than in the BG case, both across trials and across spatial locations, because each activity dimension will have some independent sources of variance in its structure across space and its activations from trial to trial. Furthermore, in circumstances of strong surround suppression, such as the large reward condition in FK,

the mean population response to the distractor may never be larger than the delay activity (Fig. S2A). In these circumstances, we would predict that attention on many trials may stay fixed at the target location, never switching to the distractor location (on some trials, fluctuations might lead peak distractor activity to exceed target delay activity and thus lead to attentional shifts). These predictions could be tested using the psychophysical methods of BG.

The original salience map hypothesis of Bisley and Goldberg (2003) is that, at any given time, the locus of attention is the RF of the LN with the highest average activity in LIP. We note that this remains valid regardless of whether single neurons have a common crossing time and whether population crossing times are variable across trials.

Section 8: Unconnected neurons behave like neurons in a single local network

In Fig. 4, we modeled the results of recording from a neuronal population belonging to a single LN. In our main simulations (Fig. 2), we instead reproduce the experimental procedure, by modeling cells recorded during different experimental sessions as coming from independent sub-networks of LIP, i.e., from independent random instantiations of the global network and its inputs. However, the analysis of a single global network in the "Detailed Analysis" sections in the main text still applies.

A neuron tends to have similar activation in its network's sum and difference patterns; this activation is determined by the particular instantiation of the probabilistic connectivity. Now consider the "virtual" sum or difference pattern of a population of neurons from different networks, determined by setting each neuron's activity to its activity in its own network's sum or difference pattern, respectively. Note that these virtual patterns are not actual Schur patterns of any network (the cells in the virtual patterns are not connected to each other, the virtual patterns are not orthogonal, etc.). Although external inputs to individual cells are variable and noisy across networks and sessions, the sum or difference patterns of each network, and thus the virtual patterns, are primarily driven by the mean inputs across LNs, which are consistent across networks and sessions. Therefore the virtual dominant patterns are activated in roughly the same manner during a trial as the dominant patterns of a single network.

Then, during steady-state activity (i.e., fixation activity and delay activity) the correlation pattern

of the population drawn from different networks behaves in the same way as a population from a single network. Outside of the steady states (i.e., transient visual activity), the activations of the virtual dominant patterns are consistent with activations of the actual dominant patterns, as long as the actual dominant patterns of different networks have similar time constants. These time constants are determined by the neuronal time constants as well as the eigenvalues and other properties of the connectivity within a given network, with the dominant eigenvalues largely determined by the mean connection strengths within and between E and I populations. Because we model different LNs as having the same statistics of neuronal time constants and connectivity parameters, we expect the time constants of the dominant patterns to be reasonably similar across LNs (see the Supplemental Data of Ganguli et al., 2008, which shows the invariance across LNs of the local slow mode decay time). We found that in our model, within a robust range of the variability of these parameters, correlation during transient states as well as steady states is indeed similar between a population of neurons drawn from different networks and a population drawn from a single network.

We have shown that a population of neurons with different RFs recorded at different times show low-dimensional dynamics (1D in BG and 2D in FK). However, we note that if these same neurons are recorded simultaneously, they would *not* show low-dimensional dynamics—they would not be activated together, simply because they have different RFs. A population of simultaneously recorded single neurons would only show the low-dimensional dynamics we observed if they share the same RF.

In this section we provided an explanation for why, in our model, neurons with different RFs show the same low-dimensional dynamics as neurons with the same RF. If our model is correct, then by the reasoning above, the low-dimensional dynamics in the BG and FK data suggest that the connectivity within and between different LIP LNs have similar statistics. This would allow LIP to process different parts of visual space in the same way.

Section 9: Discrepancies between the magnitudes of activity patterns in Fig. 4C and G and their inputs in Fig. 4D and H

In Fig. 4C and G we plotted the activation of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ during different time periods, and in Fig.

4D and H we plotted the inputs underlying those activations. Here we explain the discrepancies between the activations and the inputs plotted.

First, the inputs illustrated in Fig. 4D and H predict the steady state activation of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ *if* the inputs are sustained, which is the case for time periods (1), (2), and (4); however, over time period (3), the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ plotted in Fig. 4C and G are not at the steady state predicted by their input, because the input is transient.

Second, for the same time period, I_1 and I_2 are simply exchanged in distractor trials compared to target trials (Fig. 4D, H), predicting that the two trial types would have the same magnitude and sign of $\overrightarrow{S1}$ activity, and the same magnitude and opposite signs of $\overrightarrow{D1}$ activity. However, the residual, stochastic part of the inputs to the two LNs are not simply exchanged on the two trial types, and their stochastic activations of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ result in different vector lengths for the same time period in Fig. 4C compared to Fig. 4G. For the same reason, during the transient response at time (3) in Fig. 4C and G, the particular random instantiation of stochastic inputs in that simulation happens to make the small $\overrightarrow{D1}$ activity point in the same direction for both trial types (in other instantiations, it might point in either direction for either trial type).

Section 10: Difference in correlation drop evoked by transient visual stimulation between the Bisley and Goldberg and the Falkner, Krishna et al. datasets

During the transient target visual response on target trials, there is a larger drop in correlation in BG than in FK, in both data (Fig. 1E-F) and model (Fig. 2E-F). During the transient distractor visual response on distractor trials, the correlation in the FK model rises to a higher level than the level to which the correlation drops in the corresponding period in the BG model (compare Fig. 2F to Fig. 2E), as is also seen in the data (compare Fig. 1F to Fig. 1E). In the model, these differences do not depend on whether the two LNs are coupled, but rather occur because the variation between the visual inputs to different neurons is smaller in the FK model than in the BG model, which was meant to roughly match the model firing rate variations to those observed in the data. BG had more visual response variations across cells than FK:

distractor visual response standard deviations are 44 and 73 Hz for the two BG monkeys, and 29, 19, and 34 Hz for the three FK monkeys; target visual response standard deviations are 43 and 68 Hz for the BG monkeys, and 46, 26, and 48 Hz for the FK monkeys. The smaller visual input variation in the FK model compared to the BG model means that the weak Schur patterns are less activated relative to the dominant patterns, since the weak patterns are driven by variations in input across neurons while the dominant patterns are driven by mean inputs (see Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes" and Fig. S6). Thus the dominant activity patterns are a larger component of the visual responses in FK, yielding the higher correlations. This finding suggests a prediction: in tasks or monkeys with smaller variations in visual response, this is due to smaller variations in visual input, which will manifest as higher correlations between target fixation activity and visual responses.

Section 11: Network dynamics underlying different levels of surround suppression

The data shown in Fig. 1D and F were collected after the visual location of maximum surround suppression had been identified for the neuron being recorded, so that on each trial one stimulus is always presented in the maximum suppression location. The location of maximum surround suppression was mapped out using a similar task to the one depicted in Fig. 1B, with one stimulus (target or distractor) at the RF, and the other stimulus at a variety of locations in the surround that elicited varying levels of suppression. The correlation patterns for each location would reveal network dynamics underlying different levels of surround suppression. However, because of the small number of trials at each location, we could not reliably calculate correlations. Therefore, we examined this using our model, by modeling pairs of LNs with different across-local-network E-to-I synaptic weights. First, we see that as these weights increase, from the BG case of no connection to the case of maximum suppression in FK, the two independent slow modes of the two LNs gradually morph into the sum and difference patterns coupling the two LNs (Fig. S9). As the dominant activity patterns of the network gradually change, we expect them to lead to gradual changes in the correlation patterns. Fig. S10 shows our model predictions for correlation patterns at intermediate levels of suppression, where we've focused on the correlations on distractor trials because they

show the most salient changes from the BG to the FK case. In particular, as coupling between the LNs increases, the steady-state correlation during the delay period decreases, and the drop in correlation upon distractor onset becomes smaller and eventually turns into a rise in correlation. These effects are due to the gradual emergence of the dominant difference pattern. As the number of neurons that can be simultaneously recorded from LIP increases in the future, these predictions will become easier to test, since each visual location would elicit different levels of surround suppression for different neurons.

Section 12: Differences in PCA results between the FK data and model

We note two differences between the model and the data. First, the second PC in the model, while always well separated from PCs with lesser variance, sometimes has considerably less variance than the first PC; in Fig. 5D-G we chose for illustration a random instantiation in which the first two PCs had similar variance. Second, the variance accounted for by the PCs orthogonal to the top two PCs is considerably greater in the data than in the model. Quantitative adjustments in the model could lead to better match to the data in these respects. Stronger coupling between the two LNs (increasing the strength of across-network E-to-I connections while also adding across-network E-to-E connections to preserve the strength of surround suppression) should increase the difference between $\overrightarrow{S1}$ and $\overrightarrow{D1}$, resulting in larger variance in the $\overrightarrow{d1}$ direction and thus in the second PC. Reducing the net excitation in the network connectivity would reduce the size of the gap between the variance of the top two PCs and that of the other PCs. Because our main point is to qualitatively explain the data, we did not pursue such quantitative model adjustments.

Section 13: Dynamics and dimensionality of excitatory populations and inhibitory populations

$\overrightarrow{S1}$ and $\overrightarrow{D1}$, the two directions that define the strongly amplified 2D space for the slow dynamics of an LIP LN, primarily differ by their relative activation of E and I cells (Fig. 3C). This suggests that among populations of only E cells or only I cells, slow dynamics would be less prominently 2D. When we simulate our FK model, but picking only E cells or only I cells to form the recorded population, the correlation patterns are qualitatively similar to the correlations with both cell types in Fig. 2F (data not

shown). This occurs for two reasons. First, in the Results section "Conceptual picture: coupling of local slow modes explain LIP dynamics", we presented a simplified explanation for the correlation patterns: even if slow dynamics in each local network is one-dimensional, surround activation, whose mean suppresses that 1D slow mode and whose random fluctuations may activate fast modes, can still result in lowering of correlations. Second, when restricted to only the E cells or I cells of a single local network, $\overrightarrow{S1}$ and $\overrightarrow{D1}$ are not exactly the same. $\overrightarrow{S1}$ and $\overrightarrow{D1}$ are two perturbations of the local network slow mode. Although the main difference between them is their relative activations of E vs. I cells, the precise activation patterns of E cells and of I cells also differ between $\overrightarrow{S1}$ and $\overrightarrow{D1}$. For example, for the $\overrightarrow{S1}$ and $\overrightarrow{D1}$ plotted in Fig. 3C, the correlation between their E portions is 0.94, while that between their I portions is 0.88. Thus, restricted to E or I cells alone, there is still weakly 2D dynamics. The I cells alone are more strongly 2D than the E cells—the correlation between the E portions of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ tends to be higher than that between the I portions. We speculate this might be related to the fact that the I portion is perturbed from the local slow mode by inputs from the other local network, which is unrelated to the local slow mode, while the E portion is perturbed from the slow mode via the local connections that shaped the slow mode. PCA on I cells picked from FK simulations shows weakly two-dimensional dynamics (variance of the second PC clearly separated from that of the following PCs but considerably smaller than that of the first PC); PCA on E cells picked from FK simulations shows one-dimensional dynamics (second PC not separated from remaining PCs; data not shown). Therefore, our finding of 2D dynamics in the FK data by PCA suggests that there were at least a few I cells in the recorded population.

For model E cells alone, the dynamics do not appear one-dimensional according to the pattern of correlations across time, which matches the correlations in the FK data. However, their dynamics do appear one-dimensional according to PCA. The reason for this is as follows. For a dataset of firing rates across time for *N* neurons, PCA finds a set of *N* orthogonal *N*-dimensional firing rate patterns, the first of which carries the most variance across time, the next carrying the most variance in the subspace orthogonal to the first, the 3rd carrying the most variance in the subspace orthogonal to the first two, and so on. All *N* dimensions carry a baseline amount of variance, because of such factors as random variations in mean inputs and stochasticity in the dynamics. The PCA indicates that, for E cells alone, only one dimension

carries significantly more variance than this baseline amount, and the other *N*-1 dimensions do not. In essence, the high correlation coefficient of the E portions of $\overrightarrow{S1}$ and $\overrightarrow{D1}$ means that their average (the E component of $\overrightarrow{a1}$) carries most of their variance, and too little of their variance is carried in the orthogonal direction (the E component of $\overrightarrow{d1}$) for any of the *N*-1 dimensions to stand out as carrying significantly more variance than baseline.

Section 14: Alternative mechanisms for surround suppression and 2D dynamics

In the main text, we have shown that simple suppression of external inputs to both E and I cells of an isolated LN cannot account for the FK network dynamics (Fig. 6). This suggests that the suppression arises from direct suppressive interactions between LIP LNs. Note that such interactions could be mediated by projections to and from other areas, as has been argued for surround suppression in the "far surround" in V1 (Angelucci and Bressloff, 2006); the main point is that it should be a coupling by which activity in one LN directly suppresses activity in the other LN.

Two alternative scenarios seem possible. One is that the interacting LNs of our FK model are actually in another area that we will call area Y; each LN of area Y projects to a corresponding LN in LIP in a manner such that the LIP LN inherits not only the mean firing rate over time in area Y, but also multi-neuronal activity patterns and therefore correlation patterns. This scenario is not impossible, but we consider it unlikely for three reasons. (1) Multi-neuronal firing patterns would be inherited if there is little convergence in the projections from area Y to individual LIP neurons (e.g., one-to-one connectivity). However, there is likely considerable convergence in intracortical projections and in projections from subcortical areas to cortex, and thus the input an individual LIP neuron receives from a group of area Y neurons would reflect their average activity, regardless of the activity patterns across them. Highly variable weights at the area Y-to-LIP synapses could allow LIP to inherit area Y correlations to a certain extent, but in our simulations this is insufficient to reproduce the correlations observed by FK (data not shown). (2) The major areas projecting to LIP each show different response properties from LIP, suggesting that LIP activity patterns could not be simply inherited from them. Neurons in sensory areas such as MT and V4 fire weakly to small, stable visual stimuli such as the target present in the delay period of the FK task—they

cannot account for reliable surround suppression in LIP by sustained saccade plans during the delay. The projections from SC to LIP originate mainly from the superficial "visual" layers of SC, which doesn't exhibit delay activity (Clower et al., 2001). Using a saccade task similar to the BG and FK tasks, Suzuki and Gottlieb (2013) found that surround suppression in prefrontal cortex is much stronger than in LIP and exhibits qualitatively different properties. (3) The existence of LNs having one-dimensional dynamics, a prerequisite for our FK model, is well supported in LIP, but not in any other area. In conclusion, for the above reasons and for parsimony, we consider this scenario unlikely.

A second alternative scenario is that a surround might induce suppression by inducing external input from another area that differentially drives the E vs. I cells of an isolated LN: withdrawal of input to E cells; addition of input to I cells; or a combination of the two. In simulations of these scenarios (not shown), we found that $\overrightarrow{d1}$ activity would be driven and the FK correlation patterns could be reproduced. This is because the $\overrightarrow{d1}$ direction, unlike the $\overrightarrow{a1}$ direction, has roughly opposite means for E vs. I cells (Fig. 3C), so changing the mean input balance to E vs. I cells, relative to whatever balance existed in the external inputs prior to suppression, changes the balance of $\overrightarrow{a1}$ vs. $\overrightarrow{d1}$ activation and thus lowers correlation with the pre-suppression activity. One possible source of external input with a different E/I balance from the pre-suppression inputs might be feedback inputs from higher areas: in V1, feedback connections target E relative to I more strongly than feedforward projections (Liu et al., 2013; Yang et al., 2013), though this is not the direction of difference expected for a suppressive input. Despite arguments that V1 "far surround" suppression is mediated by projections to and from higher areas (Angelucci and Bressloff, 2006), feedback inputs contribute only modestly to surround suppression in monkey V1: letting R_{max} and R_{sur} be the response to the optimal and largest stimulus size respectively, cooling V2 and V3 causes a median decrease in surround suppression index ($1-R_{sur}/R_{max}$) of only 0.065 (compare to mean control index of about 0.9 for large stimuli) (Nassi et al., 2013). Furthermore, there is much direct evidence of V1 surround suppression that is directly mediated within V1: the strong, orientation-tuned component of V1 surround suppression is not inherited from feedforward inputs (reviewed in Ozeki et al., 2009), and V1 visual responses are strongly suppressed by activation of a neighboring region of V1 (Sato et al., 2014). Thus in V1, external inputs appear to play a small role compared to internal circuitry in mediating surround suppression, and our

results suggest that this may be a pattern conserved across cortical areas. Furthermore, we have no evidence of the pattern of inputs needed to produce the FK network dynamics in any external input sources to LIP; on the other hand, this evidence is self-contained within the FK dataset—the inputs needed to produce 2D network dynamics on distractor trials are simply the outputs of the same network on target trials, and vice versa. Thus, we conclude that the most likely and parsimonious interpretation is that surround suppression in LIP arises, at least in part, from its internal circuitry.

References

Angelucci, A., and Bressloff, P.C. (2006). Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. Prog. Brain Res. *154*, 93–120.

Liu, Y.J., Ehrengruber, M.U., Negwer, M., Shao, H.J., Cetin, A.H., and Lyon, D.C. (2013). Tracing inputs to inhibitory or excitatory neurons of mouse and cat visual cortex with a targeted rabies virus. Curr. Biol. *23*, 1746-1755.

Markov, N.T., Misery, P., Falchier, A., Lamy, C., Vezoli, J., Quilodran, R., Gariel, M.A., and Giroud, P. (2011). Weight Consistency Specifies Regularities of Macaque Cortical Networks. Cereb. Cortex *21*, 1254–1272.

Nassi, J.J., Lomber, S.G., and Born, R.T. (2013). Corticocortical feedback contributes to surround suppression in V1 of the alert primate. J. Neurosci. *33*, 8504–8517.

Sato, T.K., Häusser, M., and Carandini, M. (2014). Distal connectivity causes summation and division across mouse visual cortex. Nat. Neurosci. *17*, 30–32.

Yang, W., Carrasquillo, Y., Hooks, B.M., Nerbonne, J.M., and Burkhalter, A. (2013). Distinct balance of excitation and inhibition in an interareal feedforward and feedback circuit of mouse visual cortex. J. Neurosci. *33*, 17373-17384.