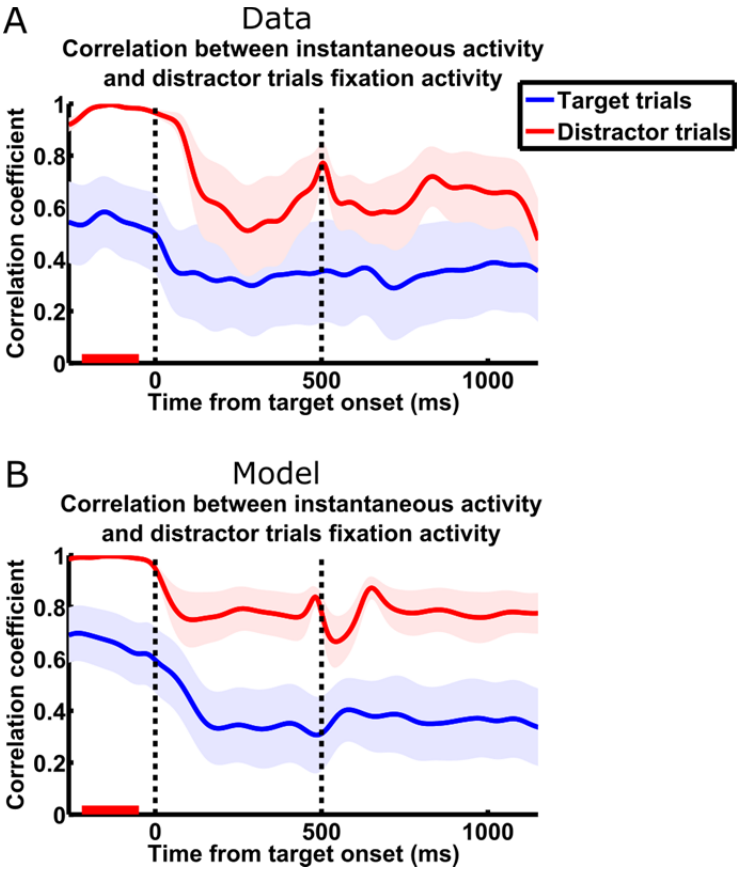
Inventory of Supplemental Items

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”



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.

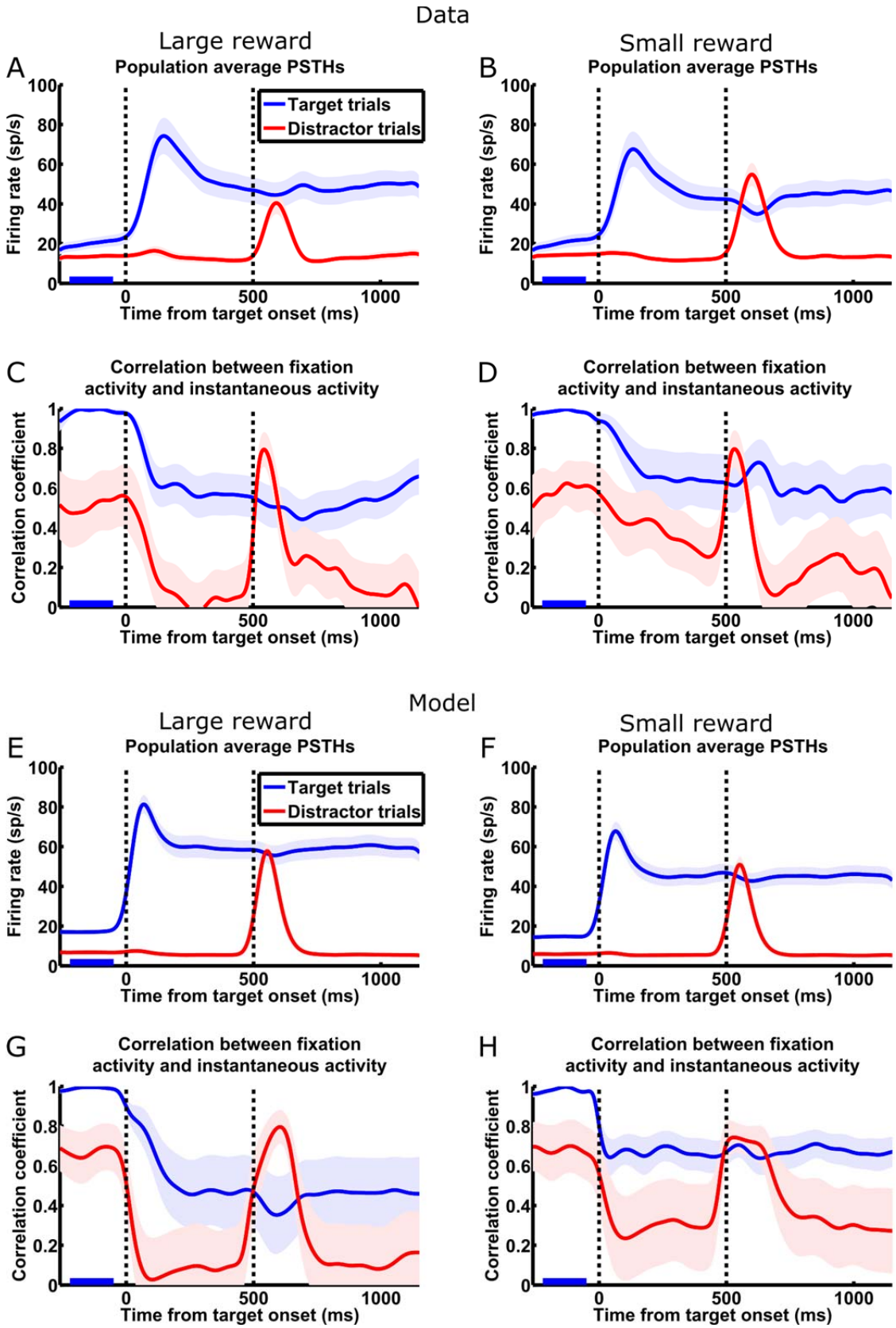

Data
Large reward
Small reward
A
B
Population average PSTHs
Target trials
Distractor trials
Firing rate (sp/s)
Time from target onset (ms)
C
D
Correlation between fixation activity and instantaneous activity
Correlation coefficient
Model
E
F
G
H

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.

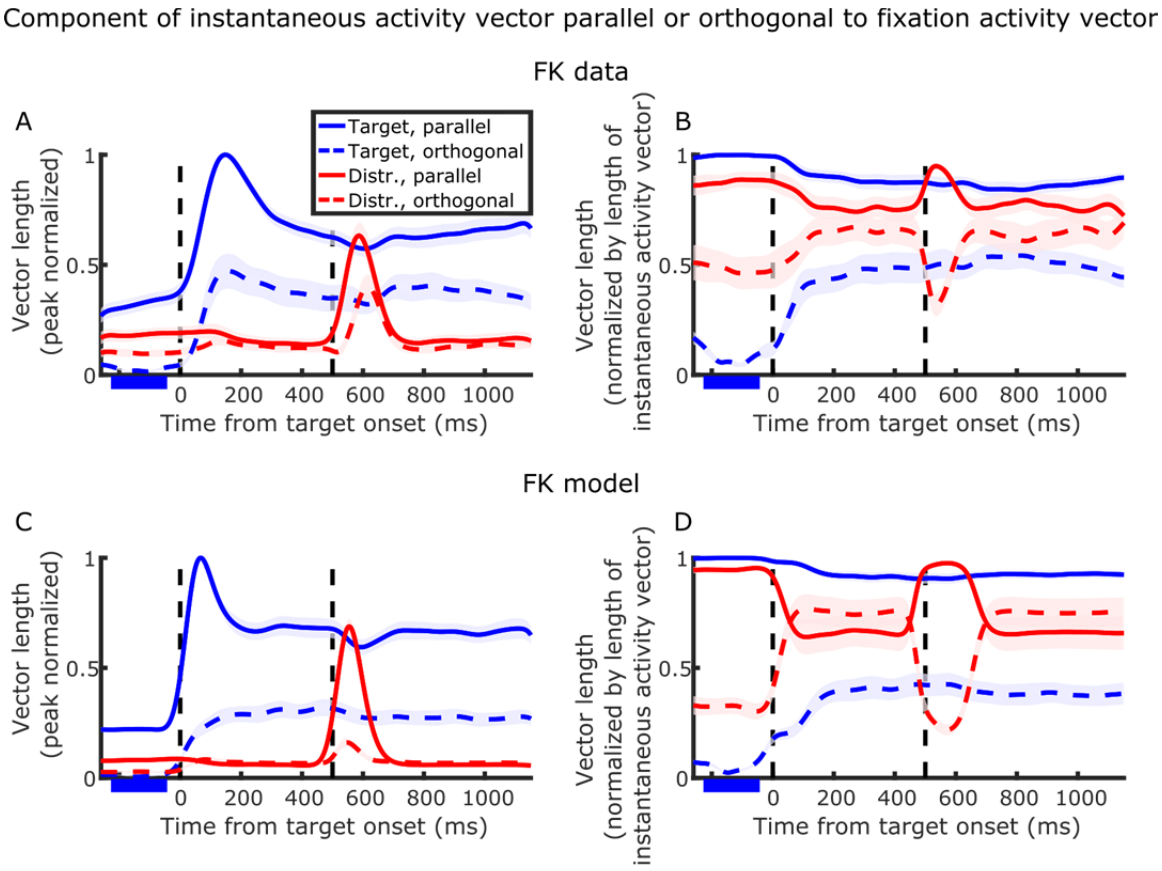


Figure S3, related to Figures 1 and 2

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.

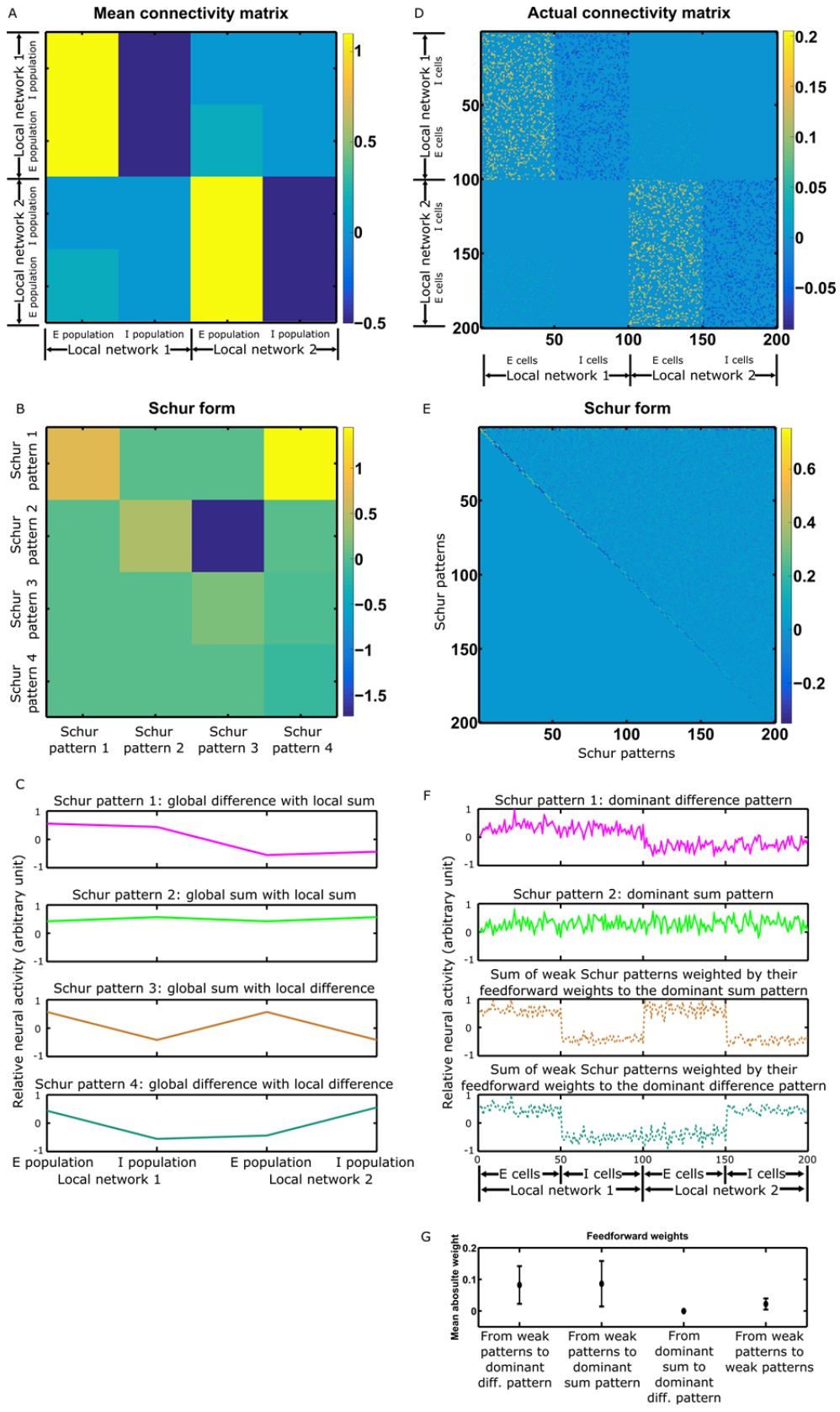

A
Mean connectivity matrix
Local network 1
Local network 2
E population
I population
D
Actual connectivity matrix
E cells
I cells
B
Schur form
Schur pattern 1
Schur pattern 2
Schur pattern 3
Schur pattern 4
E
Schur form
Schur patterns
C
Schur pattern 1: global difference with local sum
Schur pattern 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 absolute 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.

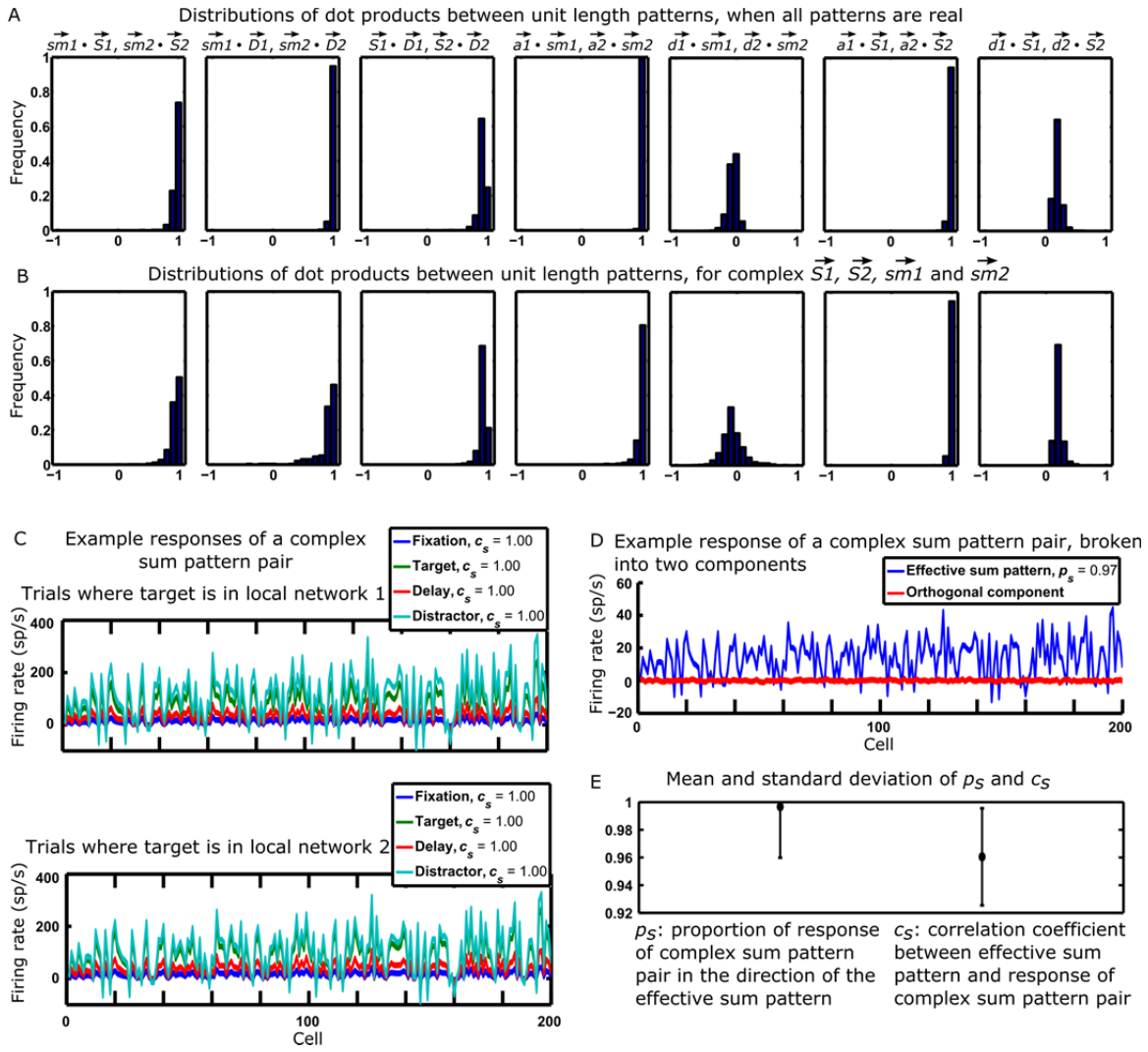


Figure S5, related to Figure 3

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.

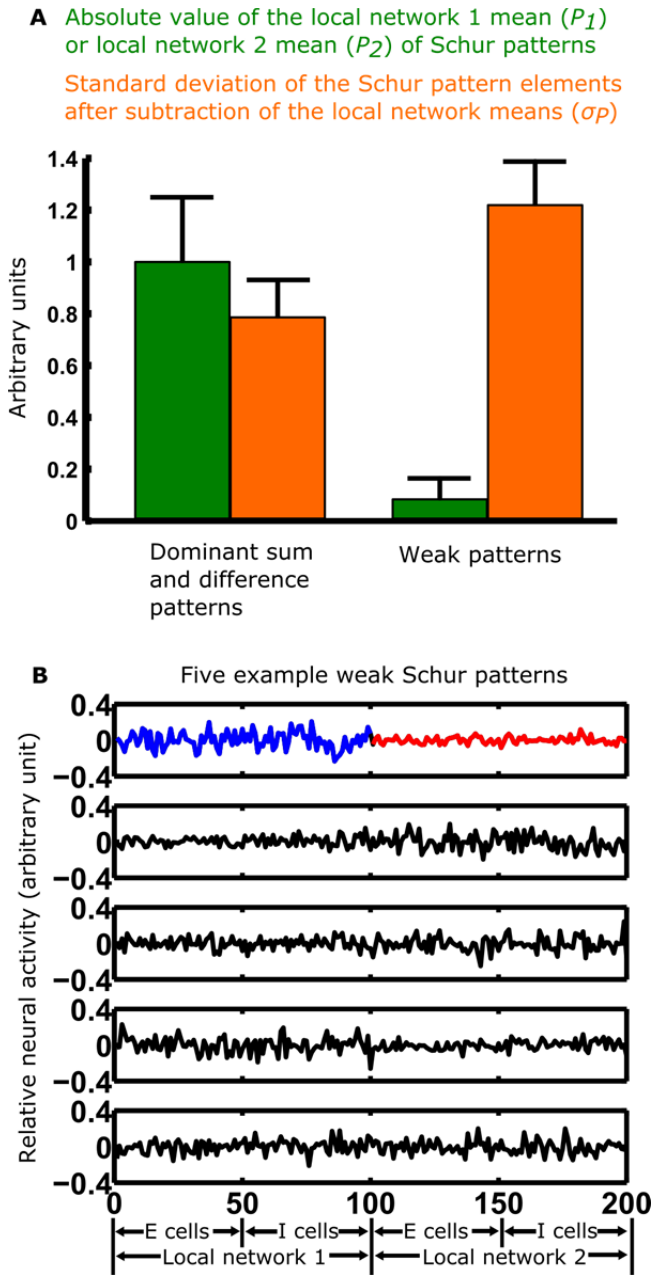


Figure S6, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"

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).

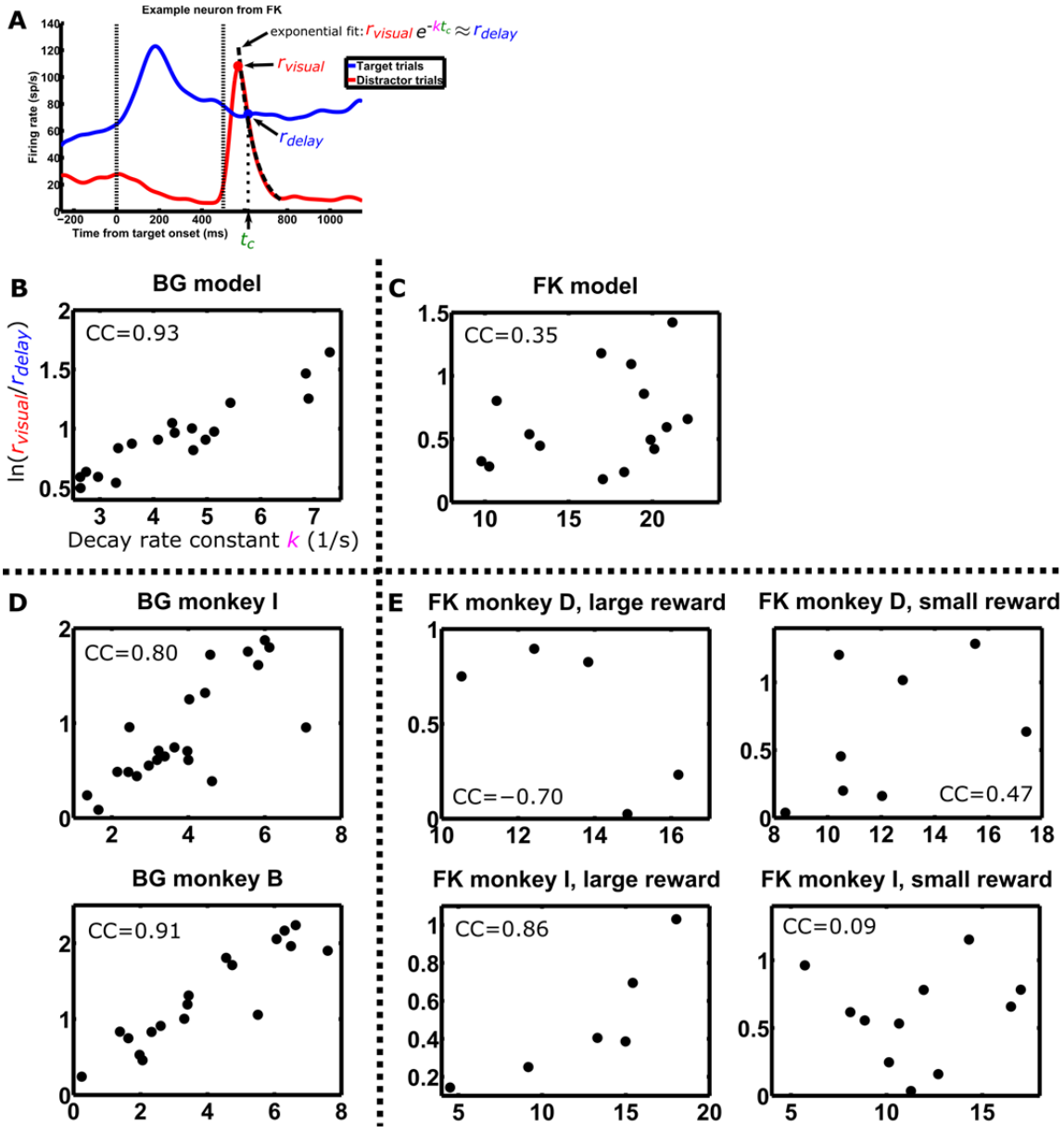


Figure S7, related to the Results section "Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes"

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.

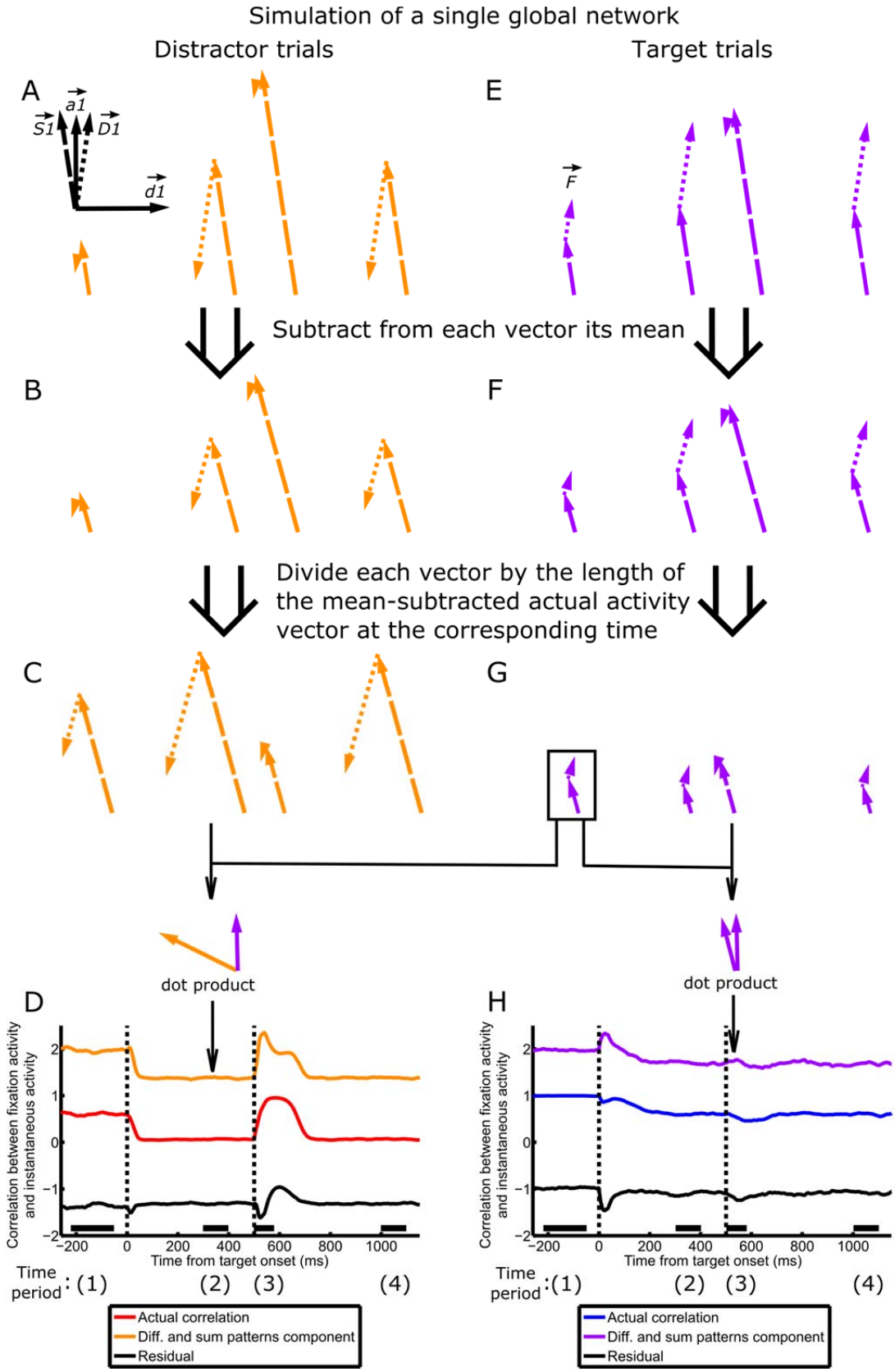

Simulation of a single global network
Distractor trials
Target trials
A
a1
S1
D1
d1
E
F
Subtract from each vector its mean
B
F
Divide each vector by the length of the mean-subtracted actual activity vector at the corresponding time
C
G
dot product
dot product
D
H
Correlation between fixation activity 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.

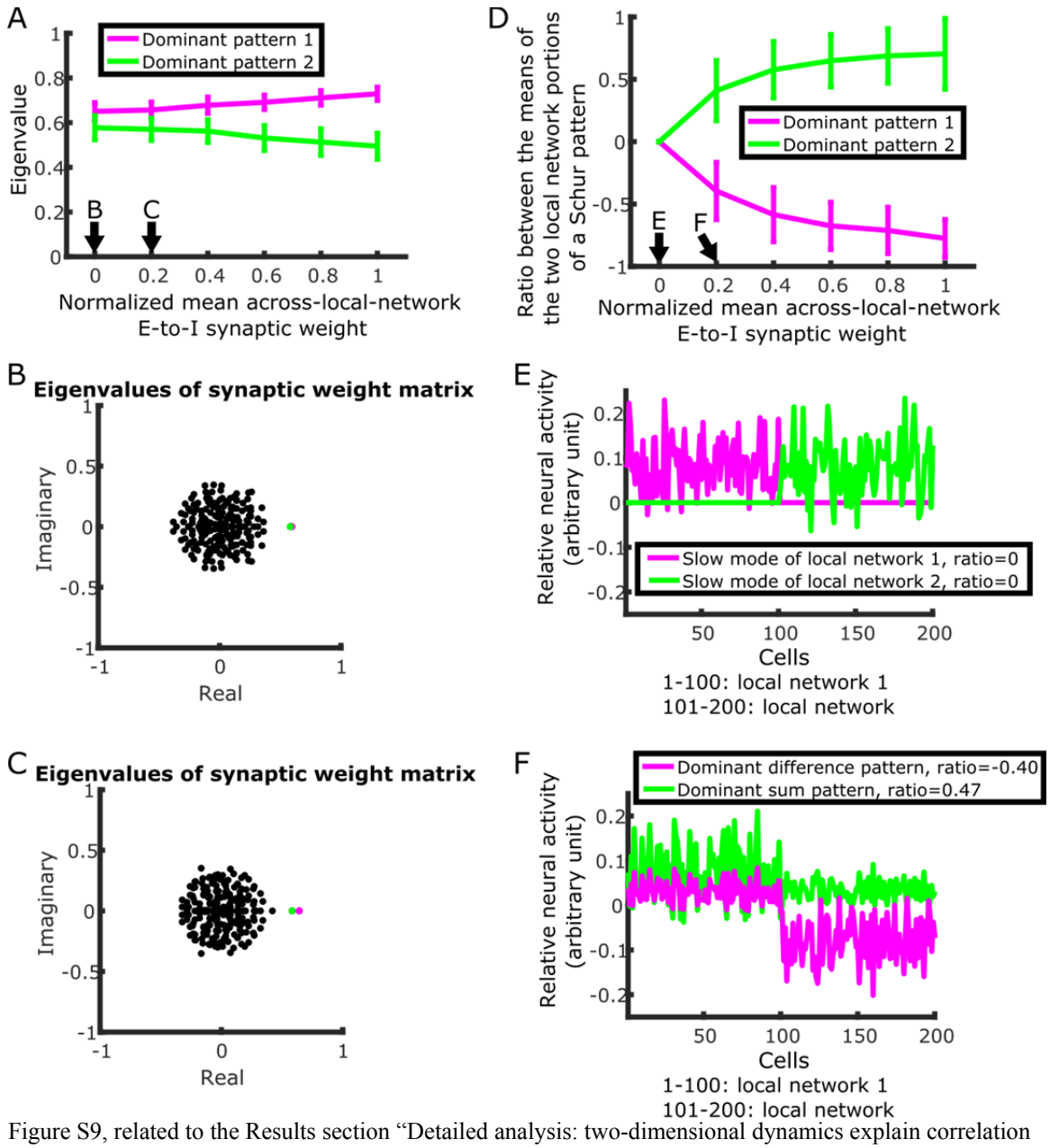


Figure S9, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"

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.

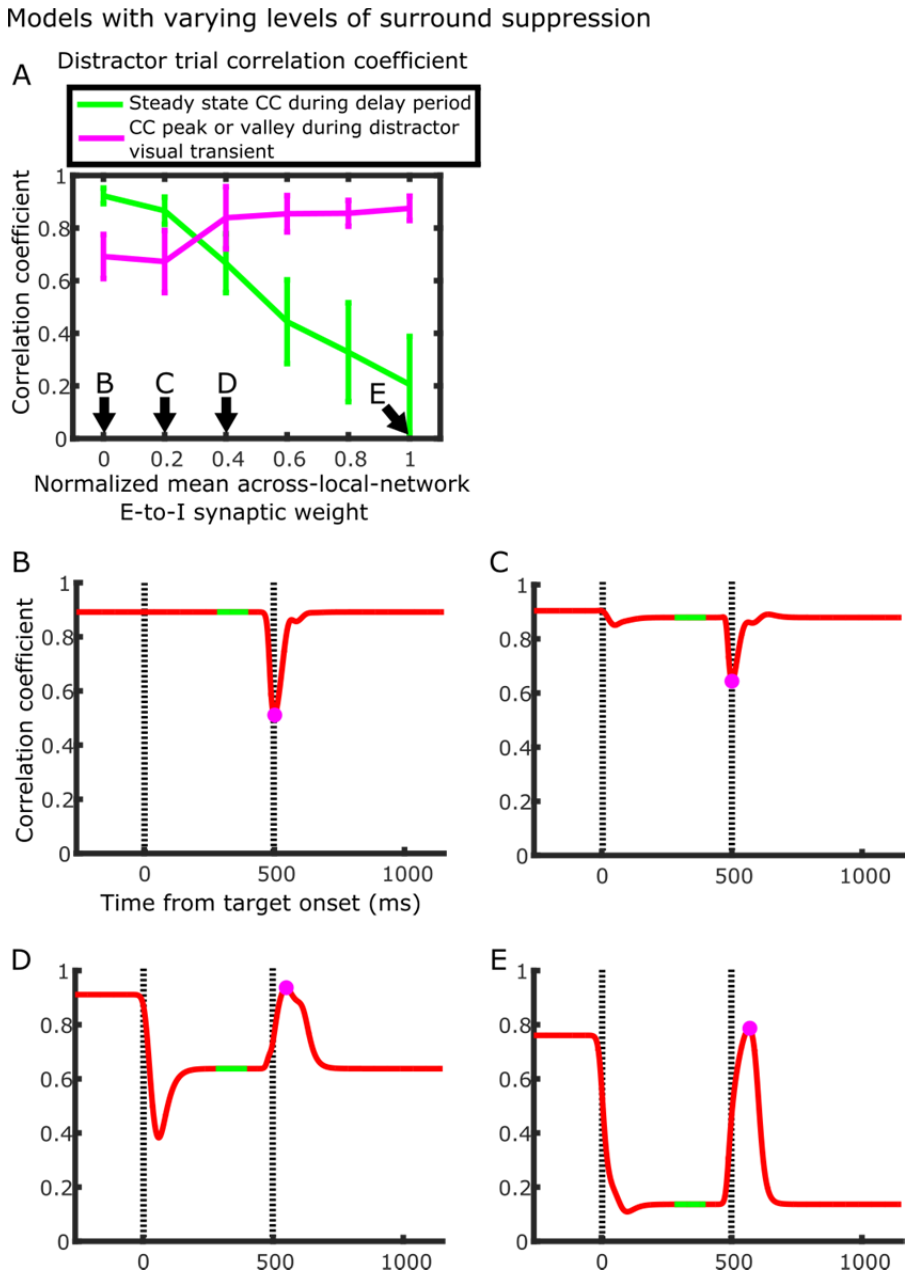


Figure S10, related to the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns"

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, and computed standard errors from the correlations calculated from each bootstrap sample population.

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, τ = 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,

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

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

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

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

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 2*N*-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:

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

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):

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

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.