Appendix S1: More example V4 cells

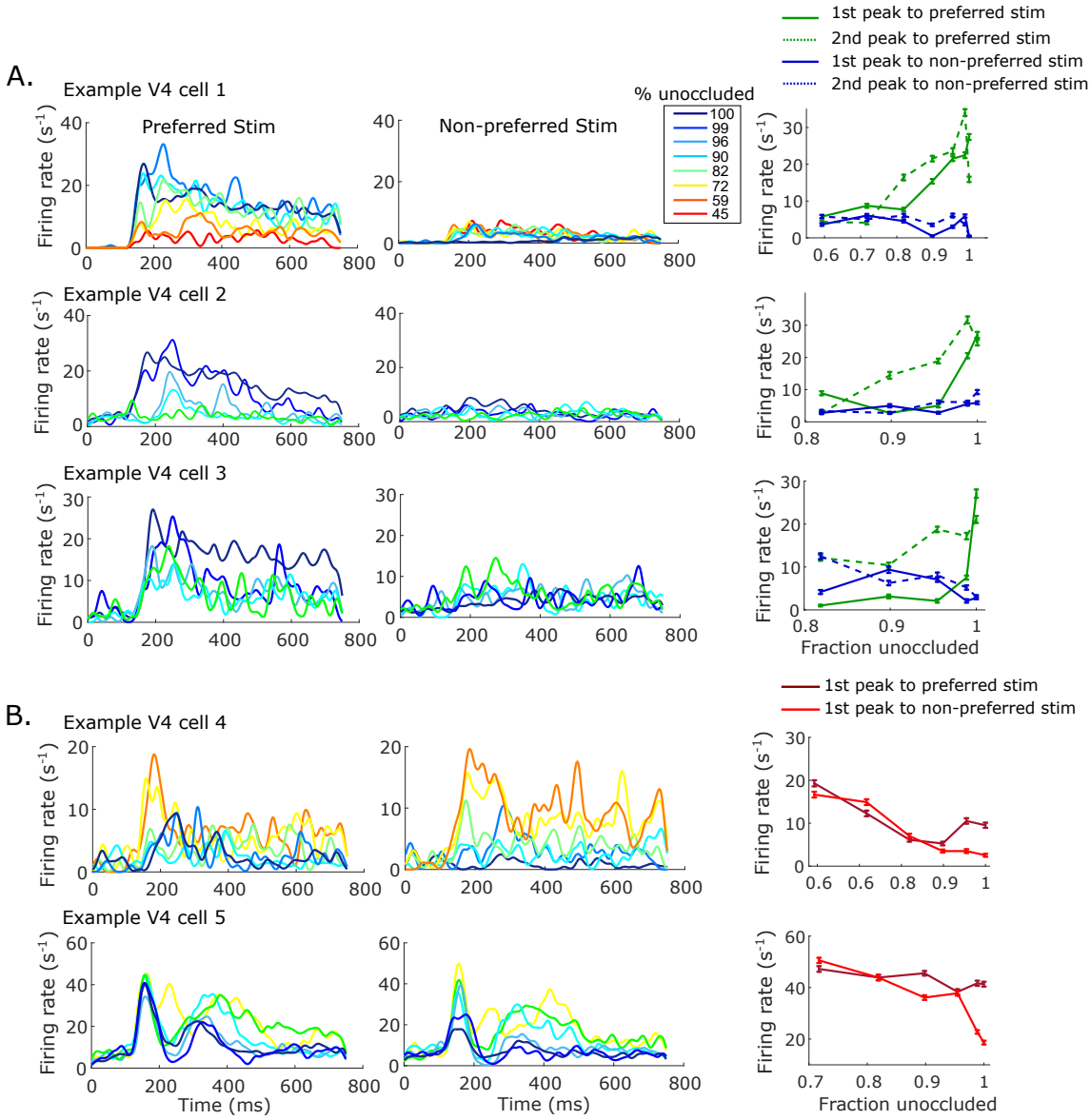

Figure S1: *Recordings from example V4 cells during the discrimination task. (A) Example V4 cells whose responses to shapes decrease with added occlusion. Their responses to preferred (left panels) and non-preferred (center panels) shape stimuli under varying degrees of occlusion are shown. The averaged firing rates during the initial peak (solid) and the second peak (dotted) of responses to the preferred (green) and non-preferred (blue) shape stimuli are plotted in the right panels. (B) Example V4 cells showing strong responses to occlusion. Their preference to occlusion was maintained regardless of whether occluders were presented with preferred (left) or non-preferred*

(center) shape stimuli. Shape selectivity decreases with added occlusion in these cells. The averaged firing rates during the initial transient responses to the preferred (dark red) and the non-preferred (light red) shape stimuli are plotted in the right panels.

Appendix S2: Population average responses

In this appendix, the model is extended to include populations of neurons with slight heterogeneity. In our main model, each unit in the model V4 and PFC is considered as a population of neurons with similar tuning properties. For example, V4 unit 1 represents a population of V4 neurons that respond preferentially to shape A, and V4 unit 3 is interpreted as a population of V4 neurons responding strongly to some salient features of the occluders. With each unit representing a neuronal population, the optimal response inferred by minimizing the cost function depicts the average response of each neuronal population. Since the cost function in Eq.14 increases linearly with added neuronal units that share the same properties with the existing populations, representing a neuronal population as a single unit seems a reasonable simplification.

We now test this simplification explicitly. We performed further numerical simulations with slightly heterogeneous group of neurons for each neuronal unit in V4 and PFC. The heterogeneity is introduced to the V4 neurons by assigning $\boldsymbol{\mu}$, the mean vector of the feedforward sensory input-driven response distribution, from a normal distribution with a unit standard deviation for each neuron within the population. Therefore, the bottom-up sensory input drives neurons within the same group to converge to slightly different optimal responses. In addition, the initial connection weights and the initial firing rates of the neurons are also slightly heterogeneous, chosen from normal distributions with the means at the initial values used in previous simulations, and the standard deviations of 0.1 for initial weights and 0.5 for initial firing rates. The PFC neurons in each population, therefore, also show weakly heterogeneous optimal representations as a result. Each neuronal population is composed of 10 slightly heterogeneous neurons.

Moreover, each PFC neuron sends the prediction signals to one neuron from each of the three V4 populations, and each V4 neuron receives the feedback that is a weighted sum of two PFC neurons, one from each PFC population. Therefore, the convergence ratio from V4 to PFC is preserved as in the previous simulations where populations are represented as single units. We have tested a couple other convergence ratio (eg, two neurons per each V4 population connected to a single PFC neuron) and found that they produce the qualitatively similar results.

The connection weight matrix $\mathbf{u}$ is learned during the preliminary phase, and the optimal responses of the V4 and PFC neurons with the learned weights $\mathbf{u}$ are obtained by minimizing the cost function E, using the same method as in the previous simulations with single unit representation. Fig. S2B shows the averaged inferred responses (dots) and the standard deviation (bars) within the population, of the shape A-selective V4 neurons (green) and the shape B-selective V4 neurons (blue) before (solid line) and after (dotted line) the feedback predictions, as a function of unoccluded area. Fig. S2C illustrates the same results in a state space view, for the responses before (left) and after (right) the feedback. The inferred responses of the neurons in the shape A-selective (V4 unit 1) and the shape B-selective (V4 unit 2) populations, predicted by the common PFC neurons, are projected onto the 2D space of the shape A and shape B-selective population responses. The level of occlusion is indicated by the colorbar, and the yellow line represents the population average responses. The population responses shown in Fig. S2B, C match the results from the single-unit representation model in Fig. 4 and Fig. 6; the shape discriminability increases during the delayed responses when the feedback predictions are included. Although not shown here, the population average

responses of the PFC populations and the occluder-selective V4 population also agree with the previous results in Fig. 4.

Treating each population as a single unit as done in Fig. 1 is therefore a reasonable simplification of the model, which expedites computation while maintaining the core mechanisms of the model. Furthermore, there may be recurrent connections among the neurons within the same group, which reduce the variances among these neurons and further validate representation of these neurons as a single unit.

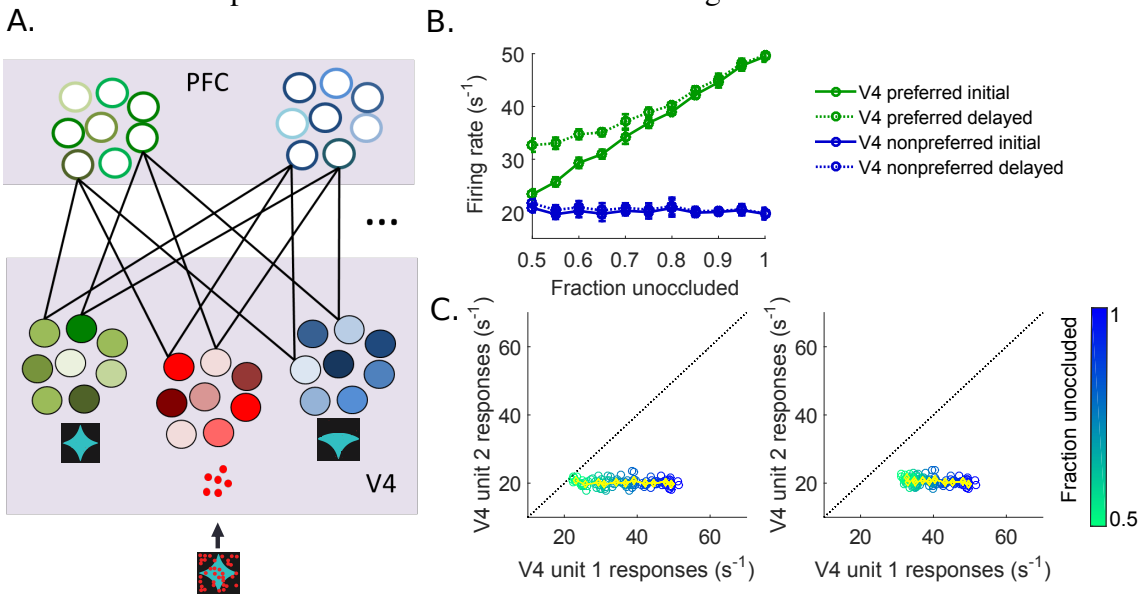

Figure S2: *Simulation with slightly heterogeneous neurons within each population. (A) Model schematic. For visualization, a smaller number of neurons per population and a subset of connections are shown. The actual model includes 10 neurons with similar tuning properties per population. Each neuron in PFC is connected to three V4 neurons from each V4 population, and each V4 neuron is connected to two PFC from each of the two PFC populations. The neurons of green shades prefer shape A and correspond to V4 unit 1, those of blue shades prefer shape B (V4 unit 2), and the neurons of red shades are selective for occluder properties (V4 unit 3). Varied shades of colors for neurons within each population represent slight heterogeneity. (B) Inferred responses of the shape-selective V4 neurons before (solid) and after (dotted) the top-down prediction. The green lines represent the optimal responses of the V4 population selective for the test*

shape A and the blue lines are those of the non-preferred V4 population that responds preferentially to shape B, as in Fig. 4D. The lines and the error bars show the averaged responses and the standard deviations across the population of 10 neurons, respectively. (C) The inferred neuronal responses of the 10 sets of V4 neurons, each of which is predicted by a common PFC neuron across degrees of occlusion, are projected onto the state space of V4 unit 1 (preferred) and unit 2 (non-preferred) responses. Yellow line represents the averaged inferred responses. Responses to high occlusion are colored green; responses to low occlusion are blue. The left and the right panels show the responses before and after the top-down prediction, respectively.

Appendix S3: Gradient descent dynamics of neuronal firing rates

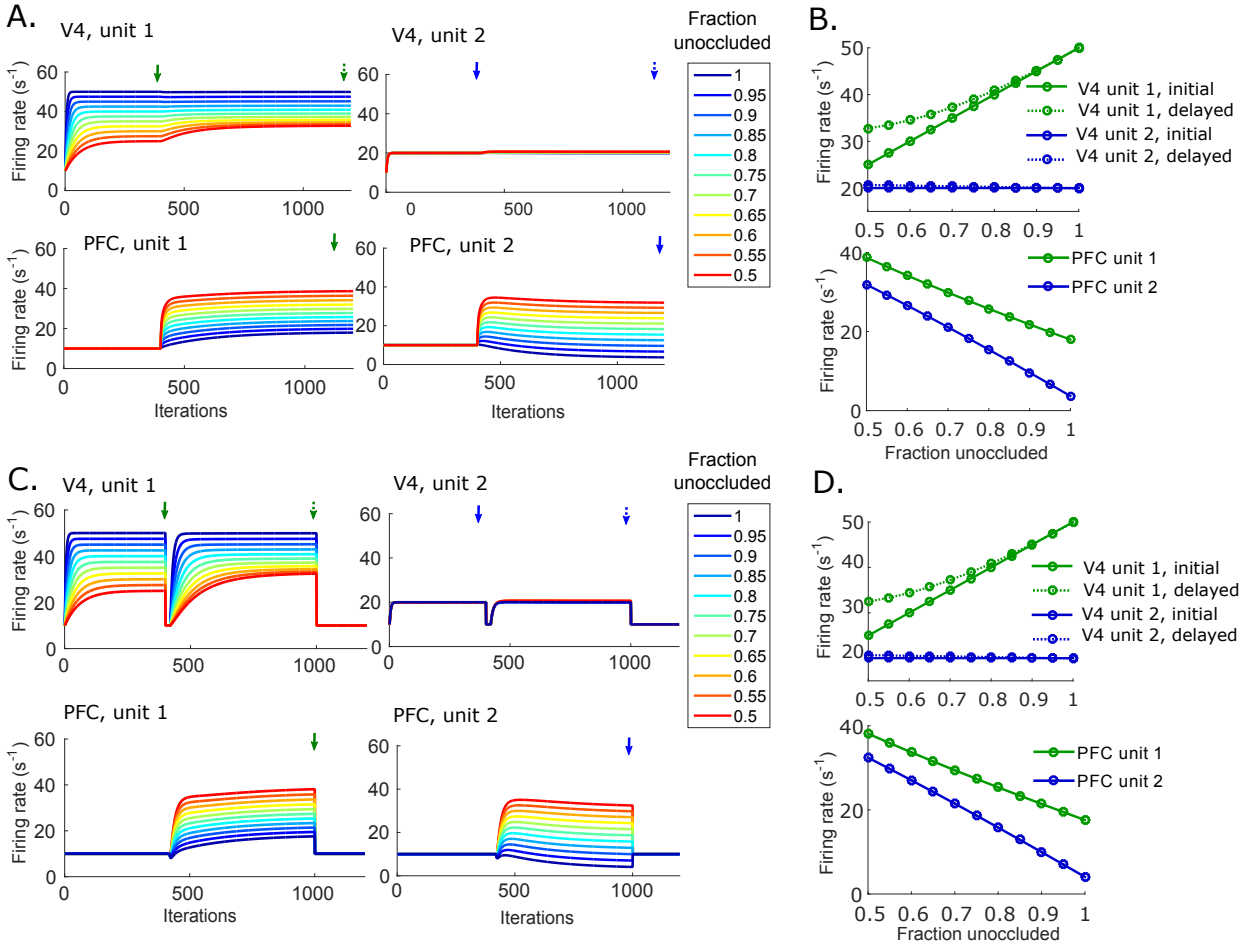

Figure S3: *Gradient descent dynamics of neuronal firing rates. (A) The V4 firing rates* $\mathbf{r_{v4}}$ *and the PFC firing rates* $\mathbf{r_{pfc}}$ *during the gradient descent. The optimization function switches from* E_1 *to* E_2*– in other words, the feedback inputs from PFC start being included in the optimization after 400 iterations. The initial peak responses and the delayed peak responses are measured at the end of each optimization process on* E_1 *and* E_2*, respectively (indicated by arrows). (B) The optimal responses of each neuronal unit found by the gradient descent in (A). (C),(D) Same as in (A),(B), except that a brief suppression is included before the gradient descent on* E_2 *starts.*

Appendix S4: Connection weights learned with partially occluded shapes

In this appendix, we make predictions on shape discriminability when the synaptic weights store templates of partially occluded shapes instead of unoccluded shapes. This represents the case where the animal has memory of partially occluded shapes rather than being exposed to unoccluded shapes.

In the simulations of the previous sections, the connection weight matrix $\mathbf{u}$ is learned based on presentations of the pair of unoccluded shapes, in order to mimic the experimental procedure where the animals discriminated a pair of unoccluded shapes at the beginning of each trial. Here we test our model with the weight matrix learned from partially occluded shapes. We train the weight matrix on the pair of the shapes under 30% and 50% occlusion (Fig. S4 B,C) and compare the results to the simulation with the weights learned based on unoccluded shapes (Fig. S4 A).

During the preliminary phase, the gradient descent with respect to the connection weights starts from the initial weight matrix

$$\mathbf{u} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Using -1 instead of 0.1 for $u_{2,1}$ and $u_{1,2}$ produces the qualitatively same result. When trained with 30% occlusion, the weight matrix converges to

$$\mathbf{u} = \begin{bmatrix} 1.39 & 0.48 \\ 0.56 & 1.47 \\ 2.13 & 2.13 \end{bmatrix},$$

and with 50% occlusion, the weight matrix converges to

$$\mathbf{u} = \begin{bmatrix} 1.27 & 0.36 \\ 0.29 & 1.19 \\ 3.00 & 3.00 \end{bmatrix}.$$

On the other hand, when unoccluded shapes are presented during the training, the weight matrix converges to

$$\mathbf{u} = \begin{bmatrix} 1.78 & 0.85 \\ 0.89 & 1.83 \\ 0.95 & 0.95 \end{bmatrix}.$$

When the weight matrix is trained on partially occluded shapes instead of unoccluded shapes, the connection weights to the occluder-selective V4 population converge to larger values over the course of the preliminary training phase. Having the weights learned, the responses of the test-shape preferred V4 unit (V4 unit 1) are plotted across degrees of occlusion, before (solid line) and after (dotted line) the feedback (Fig. S4, left column). We also plotted the total sum of the squared error signals from all three V4 units, namely, the unweighted second term of the cost function E, $(\mathbf{r_{v4}} - \mathbf{u} \cdot \mathbf{r_{pfc}})^T (\mathbf{r_{v4}} - \mathbf{u} \cdot \mathbf{r_{pfc}})$ (Fig. S4, right column). When trained on unoccluded shapes, the squared total error is minimum at zero occlusion. When trained on partially occluded shapes with 30% and 50% occlusion, the squared total error is lowest approximately at the respective occlusion levels (Fig. S4, right column).

The stronger connection weights between the PFC units and the occluder-selective V4 unit 3, that emerge from training on partially occluded shapes, change the response pattern of the preferred V4. Due to the stronger weights, the PFC responses are relatively

lower overall. Then, the delayed responses of shape A-preferred V4 unit 1 induced by PFC predictions are moved to lower values when the stimulus has no or low degrees of occlusion. As the occlusion level increases, the standard deviation σ_1 of the preferred V4 increases, weakening the bottom-up influence which suppresses the V4 responses under occlusion. As a result, under high occlusion, the optimal representation of the preferred V4 unit 1 responses depends more on the top-down PFC prediction reflecting the occluder-selective V4 response pattern.

When the training is based on a pair of unoccluded shapes, the responses of the preferred V4 unit is never lower with the feedback than without the feedback across all occlusion levels, and thus, the feedback enhances the responses under higher degrees of occlusion. On the other hand, when trained on shapes with 30% occlusion, the delayed responses are lower than the initial responses under low degrees of occlusion. For occlusion levels higher than $\sim 25-30\%$ occlusion, the delayed responses are higher than the initial responses. When trained on 50% occlusion, the delayed V4 responses are always lower than the initial responses in the occlusion range of $0-50\%$. However, the differences between the initial and the delayed responses of the test shape-preferred V4 unit 1 are very small; the initial and the delayed responses are almost identical with the parameter set used here. In addition, across the range of occlusion levels, the errors between the top-down predictions and the inferred V4 activities are smaller when the weights are trained on partially occluded shapes (Fig. S4, right column).

The deviation from the initial responses (solid line, Fig. S4) is on average smaller for the simulations with weights trained on partially occluded shapes. Since the variance σ_3' is smaller than σ_1' and σ_2', the prediction $\mathbf{u} \cdot \mathbf{r_{pfc}}$ tends to follow the response patterns of

the occluder-selective V4 unit which increase with added occlusion. When the weights are trained on unoccluded shapes, the connection from a PFC unit to the V4 unit 1 with the same shape preference is the strongest, while its connection to the occluder-selective V4 unit 3 is weaker. Then, the increase in the PFC responses with added occlusion is relatively large, compensating the effects of the small weights $u_{3,1}$ and $u_{3,2}$. The large increase in PFC responses induced by added occlusion can then evoke a larger deviation in the test shape-selective V4 unit 1 from its initial responses. When the weights are trained on partially occluded shapes, compared to the case with training on unoccluded shapes, the weights to the test shape-selective V4 unit 1 are reduced by a little, and the weights to the occluder-selective V4 unit 3 increase significantly. Then, the PFC responses do not increase as much as the occlusion level increase (the preferred PFC unit responses decrease slightly when trained on 50% occlusion or stay constant when trained on 30% occlusion, while the other PFC unit exhibits increasing responses as occlusion increases; data not shown), and thus exert milder effects on the shape-selective V4 units.

In brief, when the connection weights of the network are trained on partially occluded shapes, the feedback from PFC does not improve the shape discriminability. Our model predicts that seeing the unoccluded shapes and learning them prior to the occluded shape-discrimination task may be a necessary step to benefit from the delayed enhancement of shape discriminability induced by the feedback predictions. Testing this hypothesis will be an interesting future experimental study.

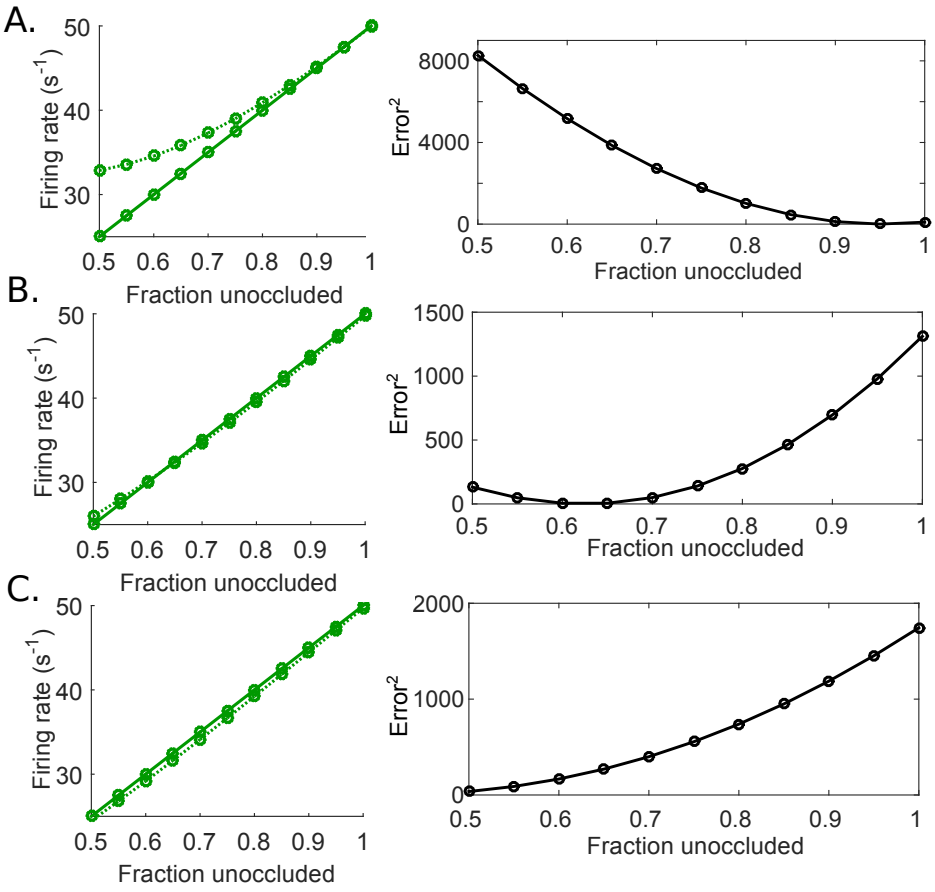

Figure S4: *Model simulations when the connection weights are learned from training on partially occluded shapes. The initial (solid) and the delayed (dotted) responses of the test shape-selective V4 unit 1 (left column), and the squared total errors between the top-down predictions and the inferred responses of the V4 units (right column), when the connection weight matrix is trained with repeated presentations of (A) unoccluded, (B)* 30% *occluded, (C)* 50% *occluded shapes chosen from either shape A or B.*