Vision Research (2021 online, 2022 printed)

Title: Cross-fixation interactions of orientations suggest high-to-low-level decoding in visual working memory

Authors: Long Luu[1], Mingsha Zhang[2], Misha Tsodyks[3] and Ning Qian[1]

Affiliation: [1]Department of Neuroscience, Zuckerman Institute, and
Department of Physiology & Cellular Biophysics
Columbia University
New York, NY 10027

[2]State Key Laboratory of Cognitive Neuroscience and Learning
IDG/McGovern Institute for Brain Research
Beijing Normal University
Beijing, 100875, China

[3]Simons Center for Systems Biology
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ 08540

Short title: Cross-fixation interactions of orientations

Keywords: visual decoding, perceptual bias, memory noise, retrospective Bayesian

Correspondence: Dr. Ning Qian
3227 Broadway, JLG Rm L5-025
New York, NY, 10027
U.S.A.

nq6@columbia.edu
Phone: 212-853-1105

Cross-fixation interactions of orientations suggest high-to-low-level decoding in visual working memory

Long Luu, Mingsha Zhang, Misha Tsodyks and Ning Qian

Abstract

Sensory encoding (how stimuli evoke sensory responses) is known to progress from low- to high-level features. Decoding (how responses lead to perception) is less understood but is often assumed to follow the same hierarchy. Accordingly, orientation decoding must occur in low-level areas such as V1, without cross-fixation interactions. However, Ding et al (2017) provided evidence against the assumption and proposed that visual decoding may often follow a high-to-low-level hierarchy in working memory, where higher-to-lower-level constraints introduce interactions among lower-level features. If two orientations on opposite sides of the fixation are both task relevant and enter working memory, then they should interact with each other. We indeed found the predicted cross-fixation interactions (repulsion and correlation) between orientations. Control experiments and analyses ruled out alternative explanations such as reporting bias and adaptation across trials on the same side of the fixation. Moreover, we explained the data using Ding et al's retrospective high-to-low-level Bayesian decoding framework.

Introduction

Sensory processing can be framed as involving encoding and decoding (Seriès et al., 2009; Zhaoping, 2014). Encoding reflects how stimuli evoke responses in sensory neurons whereas decoding specifies how the responses eventually lead to perceptual judgments of the stimuli. A large body of research has established beyond doubt that visual encoding progresses from low to high levels, with neurons in later stages of a pathway responding to higher-level features (Felleman and Van, 1991; DiCarlo et al., 2012; Yamins and DiCarlo, 2016; Yamins et al., 2014; Riesenhuber and Poggio, 1999; Serre et al., 2007; Cichy et al., 2016). Decoding, however, is less understood because one has to rely on a decoding model to relate sensory responses to subjective perception. Many decoding models assume, sometimes implicitly, that decoding follows the same low-to-high-level hierarchy of encoding (exceptions discussed below). For example, to discriminate between two line orientations, one first decodes the absolute orientation of each line (a lower-level feature) and then compare the two absolute orientations to determine their relationship (a higher-level feature) (Green et al., 1966; Paradiso, 1988; Seung and Sompolinsky, 1993; Graf et al., 2011; Teich and Qian, 2003). These models essentially assume that sensory responses generate perception (decoding) at about the same time the responses are evoked by stimuli (encoding) so that the decoding and encoding hierarchies are identical (Fig. 1a).

However, Ding et al. (2017) argued that perceptual decoding may often occur after initial sensory responses have entered working memory. This is likely whenever there is a delay between stimulus disappearance and perceptual judgment. Even under natural viewing conditions, because of our small fovea and frequent saccades, visual decoding may happen in working memory where patches of a scene from previous fixations are stored. Although the initial sensory responses to stimulus features (encoding) follow the low-to-high-level hierarchy, once all the relevant features are stored in working memory, their decoding, in principle, could be in any order. By considering invariance, noise tolerance, and behavioral relevance of high- vs. low-level features, Ding et al. (2017) proposed that sensory decoding in working memory should follow a high-to-low-level hierarchy, with the higher-level features producing a prior to constrain the decoding of lower-level features (retrospective Bayesian decoding, Fig. 1b). In particular, higher-level features are more categorical and thus can be stored in noise-resistant point attractors of working memory (Hopfield, 1984). In contrast, lower-level features are more continuous and have to be stored in continuous attractors which are more prone to noise corruption over time (Compte et al., 2000; Itskov et al., 2011). It is therefore advantageous to decode more reliable higher-level features first and use them to constrain the decoding of less reliable lower-level features in noisy working memory.

To test these ideas, Ding et al. (2017) conducted an experiment in which two lines were flashed successively and then subjects reported the absolute orientations of both lines and (implicitly)

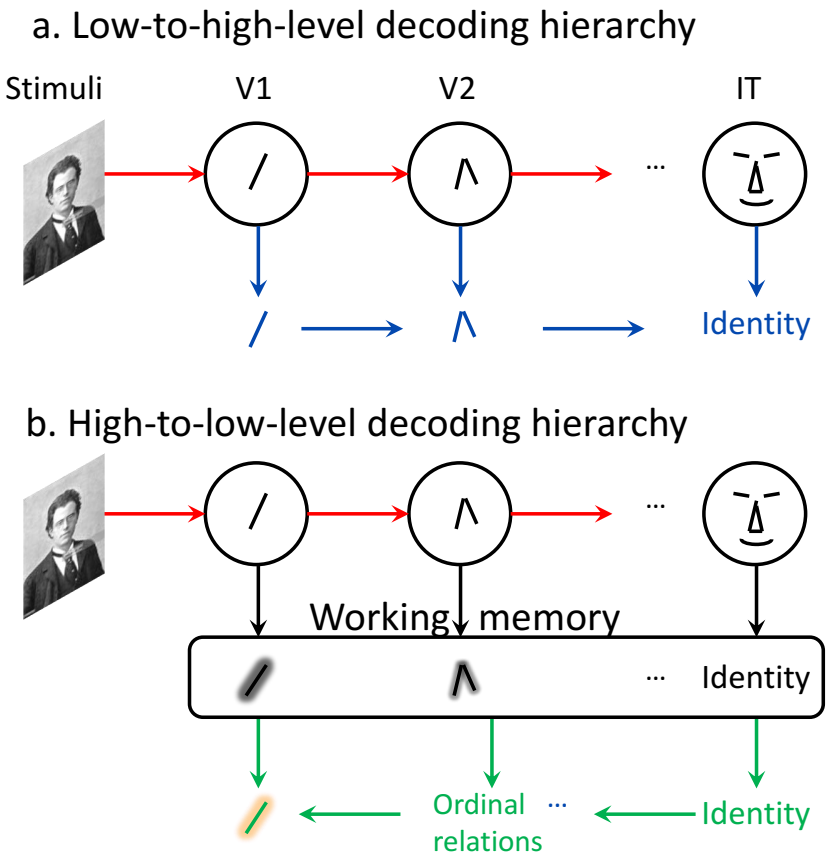

Figure 1: *Opposite decoding hierarchies.* In both panels, the red arrows indicate the well-established encoding hierarchy from low- to high-level features. (**a**) Low-to-high-level decoding of sensory responses (blue arrows). If encoding and decoding occur in sensory neurons at about the same time, then they must follow the same low-to-high-level hierarchy along sensory pathways. (**b**) High-to-low-level decoding in working memory (green arrows). If decoding happens after relevant features enter working memory, then it should progress from high to low levels, with higher-level features constraining lower-level features, because higher-level features are more invariant, reliable, and behaviorally relevant (Ding et al., 2017).

their ordinal relationship (whether the second line is clockwise or counter-clockwise from the first). They found that the two lines interacted perceptually in various ways that can be explained by the high-to-low-level decoding but not by the low-to-high-level decoding. For example, the second line repelled the first line (backward aftereffect) as much as the first line repelled the second line (forward aftereffect). The low-to-high-level decoding cannot explain the backward aftereffect because when the first line was decoded directly from its initial sensory response, the second line had not yet appeared. In contrast, the high-to-low-level decoding is assumed to occur after the encoding of both lines and their relationship have entered working memory where the higher-level ordinal relationship is decoded first, and then constrains the decoding of the lower-level absolute orientations to produce the observed mutual repulsion. The same mechanism accounts for another interaction: the correlation between two reported absolute orientations in a trial.

A surprising prediction of the high-to-low-level decoding scheme is that two stimuli traditionally considered as independent may interact with each other if they are both task relevant and represented in working memory. A specific example is two orientation stimuli, or two translation-motion stimuli, on opposite sides of the fixation. Orientation or translation-motion interactions (such as adaptation aftereffects and simultaneous contrasts) typically require that the stimuli occupy the same or nearby regions on retina (Gibson and Radner, 1937; Meng et al., 2006; Xu et al., 2008). The standard explanation is that these simple features are first decoded in low-level areas such as V1 whose small receptive fields do not include both hemifields to support cross-fixation interactions. However, in such studies, usually only one stimulus, but not the other, is task relevant and stored in working memory. For example, in a standard adaptation paradigm, subjects only report the test stimulus, but not the adaptor. Similarly, the rod-and-frame illusion is usually demonstrated with the frame around the rod, instead of with the frame and rod on opposite sides of fixation, and subjects only report the rod, not the frame (Beh et al., 1971). We thus tested whether two lines could interact cross fixation when *both* lines were task relevant, and indeed found the predicted interactions. Moreover, we found the interactions regardless of whether subjects reported the two lines' orientations one after another continuously or with an interruption between the reports. Finally, we demonstrated that Ding et al. (2017)'s high-to-low-level decoding framework explained the data from both reporting methods.

We note that a wealth of psychophysical results can be re-interpreted as high-to-low-level decoding in working memory although the studies' original interpretations of formally similar models may be different (Luu and Stocker, 2018; Stocker and Simoncelli, 2008; Qiu et al., 2020; Fritsche and de Lange, 2019; Jazayeri and Movshon, 2007; Zamboni et al., 2016; Bronfman et al., 2015; Talluri et al., 2018; Bae and Luck, 2017; Li et al., 2019) (see Ding et al. (2017) for detailed discussions). Another set of studies emphasize high-to-low-level processing (Navon, 1977; Chen, 1982; Ahissar and Hochstein, 2004; Oliva and Torralba, 2006) but they do not separate encoding and de-

coding hierarchies, or consider noisy working memory, or model how higher-level decoding affects lower-level decoding. There are also theories proposing bi-directional interactions along processing pathways (Atkinson and Shiffrin, 1968; Carpenter and Grossberg, 1987; Ullman, 1995; Lee and Mumford, 2003). While these theories address other important issues (such as the ART's solution to the stability-plasticity dilemma), they are not concerned with how noise in working memory may shape the decoding hierarchy (Ding et al., 2017). To our knowledge, previous studies never predicted nor tested cross-fixation interactions of remembered orientations. We therefore believe that by distinguishing between the encoding and decoding hierarchies and specifying the decoding mechanisms as high-to-low-level constraints in working memory, the retrospective Bayesian scheme (Ding et al., 2017) may provide a coherent framework for understanding a range of perceptual phenomena. Preliminary results were published in abstract form (Luu et al., 2020).

Methods

Experimental procedure

Fifteen subjects with normal or corrected-to-normal vision (10 males, 5 females; all naïve) participated in the experiments. All subjects provided informed consent. The experiments were approved by the Institutional Review Board of Columbia University.

General procedure: During the experiments, subjects sat in a darkened room and viewed the stimuli on a large-screen monitor (Samsung QN55Q6F, 55 inch, refresh rate: 120 Hz and resolution: 3840 x 2160 pixels) at a distance of 56 cm. We enforced the viewing distance and head stabilization with a chin rest and head band. All experiments were run in Matlab (Mathworks, Inc.) in combination with PsychoPhysics Toobox (Brainard, 1997). A Dell computer (Intel core i7-8700, 16GB RAM and NVidia GTX 1060 graphics card) controlled the stimulus presentation, and another Dell computer (i5-8400, 8GB RAM) controlled an infrared video-based eye tracker developed in Mingsha Zhang's lab (1000 Hz sampling rate). Subjects' gaze were always monitored during the experiment. There were three experimental conditions run in separate blocks. Before each condition, we gave subjects detailed instruction on the task and let them practice until they were comfortable with their performance. Each stimulus line was 6° by 0.1°.

1-line condition: At the beginning of a trial, subjects had to maintain fixation on a white dot (diameter: 0.27°) at the center of the screen. The trial only started when subjects successfully maintained fixation within a circular window (radius: 3°) around the fixation dot for 1 second. A line then appeared on either the left or right side (counter-balanced and randomized) of the fixation dot, centered at the eccentricity of 8°. The line color was magenta and green for the left

and right side, respectively. The line's orientation was either 49° or 54° from the horizontal in two separate blocks. During the presentation, if subjects' gaze broke the fixation window, a tone (200 Hz, 0.5 second) was played, and the trial was aborted and repeated. After a 1-second duration, the stimulus line disappeared and a beep (400 Hz, 0.2 second) was played to prompt subjects to report the orientation of the stimulus line. To report the line's orientation, subjects first moved the mouse along the perceived orientation. After the mouse motion started, a marker line appeared at the fixation with an orientation along the mouse's moving direction. The marker line had the same color and length as the stimulus line. Subjects then rotated the marker line with the mouse to fine-tune their estimate of the stimulus orientation, and left-clicked to report. They were instructed to take time to be as accurate as possible.

2-line condition: Similar to the 1-line condition, subjects had to successfully maintain fixation for 1 sec before the stimulus presentation. Then, two colored lines were presented on the opposite sides of the fixation dot, each centered at the eccentricity of 8°. Consistent with the 1-line condition, the left line was magenta and the right line was green. The lines' orientations were 49° and 54° that were counter-balanced and randomized across trials. As for the 1-line condition, a trial was aborted and repeated whenever subjects broke the fixation window during the stimulus presentation. This ensured that the two stimulus lines always occupied well-separated retinal locations on opposite sides of the fixation. After 1 second, the stimulus lines disappeared and a tone prompted subjects to first report the orientation of the left line by drawing and adjusting a magenta marker line. After subjects clicked to confirm the estimate of the left line orientation, the marker line changed color from magenta to green and subjects rotated it to the estimate of the right line orientation and clicked again. Note that subjects always reported the left line first and then the right line, to avoid any potential confusion.

2-line-interrupt condition: The experimental procedure was identical to that for the 2-line condition except that after subjects clicked to confirm the report of left line, the magenta marker line disappeared, and subjects had to move the mouse again to draw the green marker line and used it to report the right line's orientation.

Data analysis

Computation and statistical test of repulsion and correlation: To compute the repulsion and correlation of subjects' reports of the two lines, we first flipped (mirrored) all the incorrect trials with respect to the diagonal line (see Results for explanations). Then we computed the mean difference and Pearson correlation between the two reports in a trial. The repulsion was computed by subtracting the mean difference of the baseline, 1-line condition from that of the 2-line or 2-line-interrupt conditions. To test the significance of the observed effects at the group level, we

first obtained the mean values of repulsion and correlation for each individual subject. Then we use Wilcoxon sign rank test on these values. For the statistical test of individual subjects, we used bootstrapping (n = 10,000) to obtain the 95% confidence interval of the mean difference and correlation for each subject. Then we plot the results of the 2-line or 2-line-interrupt conditions versus the 1-line condition. If the confidence interval did not touch the diagonal line, the effect was statistically significant at 0.05 level.

Analysis of cross-trial adaptation at the same site: We quantified how much traditional adaptation across trials on the same site contributed to the observed repulsion effect in the 2-line condition. In the separate n-back analysis, we split the trials into the "same" and "different" sets according to whether stimulus orientations of a given trial and the n-back trial were the same or different. In the cumulative n-back analysis, we split the trials into the "same" and "different" sets according to whether stimulus orientations of a given trial and all the n previous trials were the same or different. This required the n previous trials all had the same orientation, thus halving the number of available data points with each increment of n. For each set, we computed the repulsion by subtracting the mean difference in the 1-line condition from the mean difference in the 2-line condition. To measure how much the traditional adaptation contributed to the observed repulsion, we used the adaptation index: $(R_d - R_s)/(R_d + R_s)$, where R_d and R_s are the repulsion of the "different" and "same" sets, respectively.

Decoding models

Model descriptions

The 1-line condition: We assume that the two orientations are represented independently, each with Gaussian sensory and memory noises, and decoded independently. When stimulus orientation $\theta_i, i = 1, 2$ is presented in a trial, a sensory sample s_i is drawn according to the Gaussian probability density $p(S_i|\theta_i) = N(\theta_i, \sigma_s)$. Then at the report time, a memory sample m_i is drawn according to the Gaussian probability density $p(M_i|s_i) = N(s_i, \sigma_m)$. A Bayesian decoder with a uniform prior generates an estimate of the stimulus orientation at the center-of-mass of the likelihood function, which in this case equals the memory sample m_i.

The 2-line condition with low-to-high-level decoding: According to the low-to-high-level decoding scheme, the two (lower-level) absolute orientations in a 2-line trial are first decoded independently (as in the 1-line case), and the results are then compared to decode the (higher-level) relationship between the orientations. Thus, according to this scheme, the 1-line data predicts the 2-line data. Specifically, for the low-to-high-level decoding, we sampled from the measured 1-line distributions of the two orientations to generate the predicted 2-line joint distribution and its derived proper-

ties (difference distribution, correlation, and repulsion). Note that the low-to-high-level decoding model does not involve working memory (Fig. 1a) but the 1-line data must contain both sensory and memory noise. However, there is no need to separate the noise sources since this model predicts the 2-line data poorly regardless of the noise level: the model cannot explain the cross-fixation interactions (correlation and repulsion) in the 2-line data because it treats the two absolute orientations separately.

The 2-line condition with high-to-low decoding: The model makes the same assumptions about the sensory and memory noise as for the 1-line case to produce a likelihood function for the absolute orientations: $p(m_1, m_2|\theta_1, \theta_2) = p(m_1|\theta_1)p(m_2|\theta_2)$ except that we used a different σ_m for the memory noise because subjects had to memorize two lines instead of one. The decoding procedure (Ding et al., 2017), however, follows the opposite hierarchy of the low-to-high-level scheme above. First, the model uses the sensory sample s_1 and s_2 of the left and right orientations in a trial to decode their ordinal relationship $\hat{O}$, namely whether the left orientation is larger or smaller than the right orientation. Formally, $\hat{O}$ is the option that maximizes the posterior for the ordinal choice O given the sensory samples, $p(O|s_1, s_2)$. Since a priori the two options were equally probable in our experiments, we determine $\hat{O}$ according to whether s_1 is larger or smaller than s_2.

Since the discrete choice $\hat{O}$ can be stored in a noise-resistant, point attractor of the memory system, we assume it is immune to the memory noise (Ding et al., 2017). In contrast, the sensory sample s_1 and s_2 for the continuous, absolute orientations have to be stored in continuous, ring attractors which are prone to memory noise, and at the report time, they become memory samples m_1 and m_2 in a trial. If the ordinal decoding $\hat{O}$ is usually correct, then using it to constrain the likelihood function of m_1 and m_2 can improve the accuracy of the absolute decoding. Specifically, $\hat{O}$ produces a prior, $p(\theta_1, \theta_2|\hat{O})$, which is a step function along the diagonal line in the joint space of the two orientations. The opposite choices of $\hat{O}$ produce the corresponding, opposite step functions. Multiplying this prior to the likelihood function produces the posterior of the absolute orientations:

$$p(\theta_1, \theta_2|m_1, m_2, \hat{O}) \propto p(m_1, m_2|\theta_1, \theta_2)\ p(\theta_1, \theta_2|\hat{O}) \tag{1}$$

The prior erases the part of the likelihood function either above or below the diagonal line that is inconsistent with the ordinal judgment $\hat{O}$. Then the stimulus absolute orientations are decoded as the mean of their posterior:

$$\hat{\theta}_i = \iint \theta_i p(\theta_1, \theta_2|m_1, m_2, \hat{O})\ d\theta_1\ d\theta_2 \tag{2}$$

for $i = 1, 2$.

The 2-line-interrupt condition with high-to-low decoding: The model is identical to the high-to-low decoding model for the 2-line condition up to the posterior $p(\theta_1, \theta_2|m_1, m_2, \hat{O})$. However, only the

left orientation decoded from the posterior is reported before the interruption (the disappearance of the marker line). After the interruption, we assume that the process of redrawing the marker line again for the second report means that a new memory sample (m_1', m_2') is drawn to form a new posterior, $p(\theta_1, \theta_2 | m_1', m_2', \hat{O})$, in the same way as we did for $p(\theta_1, \theta_2 | m_1, m_2, \hat{O})$. This time only the right orientation decoded from the new posterior is reported. We considered two ways for drawing the new memory sample, producing two versions of the model. The first version is to draw m_i' from the Gaussian density $N(m_i, \sigma_m)$; this means that the new memory sample is the old memory sample further corrupted by memory noise. The second version is to draw m_i' from the Gaussian density $N(\hat{\theta}_i, \sigma_m)$ where $\hat{\theta}_i$ are the estimate of the first decoding. This means that the new memory sample is the first decoded orientations further corrupted by memory noise. In both versions, we assume that the new noise has the same σ_m as the memory noise for the first decoding. We believe this is a good approximation because the reaction times of the first and second reports are similar in the 2-line-interrupt condition (see Supplementary Fig. S4b).

Model fitting procedures

To fit the models to subjects' data, we first obtain the distribution of the decoded orientations given the actual orientations, $p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2)$, by marginalizing (integrating over) the latent memory variables and the ordinal judgment variable:

$$p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2) = \sum\nolimits_{\hat{O}} p(\hat{O} | \theta_1, \theta_2) \iint p(\hat{\theta}_1, \hat{\theta}_2 | m_1, m_2, \hat{O})\; p(m_1, m_2 | \theta_1, \theta_2)\; dm_1\; dm_2. \tag{3}$$

For Gaussian noises, this can also be done with Ding et al. (2017)'s analytical formula (their Eqs. 1 and 2) for $\hat{\theta}_1$ and $\hat{\theta}_2$ by sampling m_1, m_2, and $\hat{O}$ for given θ_1 and θ_2. (Note that m_i was called r_i, and $\sigma_s^2 + \sigma_m^2 = \sigma_i^2$ in Eqs. 1 and 2 of Ding et al. (2017), and the two opposite choices of $\hat{O}$ correspond to swapping $\hat{\theta}_1$ and $\hat{\theta}_2$ in the two equations.)

We then use $p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2)$ to obtain the difference distribution $p(\hat{\theta}_2 - \hat{\theta}_1 | \theta_1, \theta_2)$. We jointly fit the model to the 1-line and 2-line data pooled over all subjects by maximizing the likelihood of data with respect to the model parameters using Nelder-Mead algorithm. The model has 3 parameters: the sensory noise σ_s and the separate memory noises σ_m for the 1-line and the 2-line conditions. For the 2-line data, we fit the difference distribution instead of the joint distribution because the joint distribution is 2D and we do not have a large amount of data to fit it robustly. Moreover, fitting the difference distribution can already capture the characteristic bimodal pattern of the joint distribution.

Model prediction for the 2-line-interrupt condition

Given the fit parameters for the 1-line and 2-line conditions, we predict the 2 line-interrupt condition without new free parameters using the two high-to-low-level decoding steps described above.

Results

Cross-fixation interactions of orientations in working memory

The first experiment was similar to that of Ding et al. (2017) but instead of presenting two lines sequentially at the fixation, we presented them simultaneously on opposite sides of the fixation point (Fig. 2b), for 1 sec. The lines were 6° by 0.1°, and oriented 49° and 54° from horizontal, respectively. The two orientations were counter-balanced and randomized for the two sides over 50 trials of a block. The center-to-center distance between the lines was 16°. An infrared eye tracker (see Methods) was employed to monitor eye position online, and each trial started after subjects acquired fixation for 1 sec. The fixation window was a circle of 3° radius, and trials with broken fixation during stimulus presentation were aborted and repeated. After the lines disappeared, subjects first reported the left line's orientation by drawing a marker line with a mouse according to the perceived orientation, adjusting it to match the perceived orientation as closely as possible, and clicking a button. They then continued to rotate the marker line to match the right line's orientation as closely as possible and clicked to report. As in Ding et al. (2017), the continuation from the first to the second report let subjects implicitly indicate the lines' ordinal relationship (the second experiment below interrupted this continuation). After an inter-trial-interval of 300-600 ms, the next trial started. To minimize potential mix-up of the two stimulus lines, we always colored the left and right lines magenta and green, respectively, and changed the marker line color from magenta to green after the first click (Fig. 2b).

In addition to the above 2-line condition, we also ran the corresponding baseline, 1-line condition, in which only one line (either 49° or 54° in separate 50-trial blocks) was presented either on the left or on the right of the fixation (counter-balanced and randomized) and subjects reported its orientation as they did for the first line in the 2-line condition (Fig. 2a).

We collected data from 15 subjects (all naive). We first describe the distributions of the reported absolute orientations of the individual lines. In the 1-line condition, the absolute distributions (Fig. 2c) are roughly centered at the lines' actual orientations (49° and 54°). The difference between the means of the two distributions is 5.3°, close to the actual 5° difference. In the 2-line condition, the absolute distributions (Fig. 2d) for the two lines are further apart compared with those of the 1-line condition, with an 8.6° difference between the means, indicating a perceptual repulsion between the lines. The repulsion is statistically significant ($p = 0.00006$, Wilcoxon sign rank test).

There was considerable variability of the reported absolute orientations. Because the stimulus lines were flashed on the periphery and subjects reported well after the stimuli disappeared, the variability must reflect both sensory and memory noises. The variance of absolute distributions in

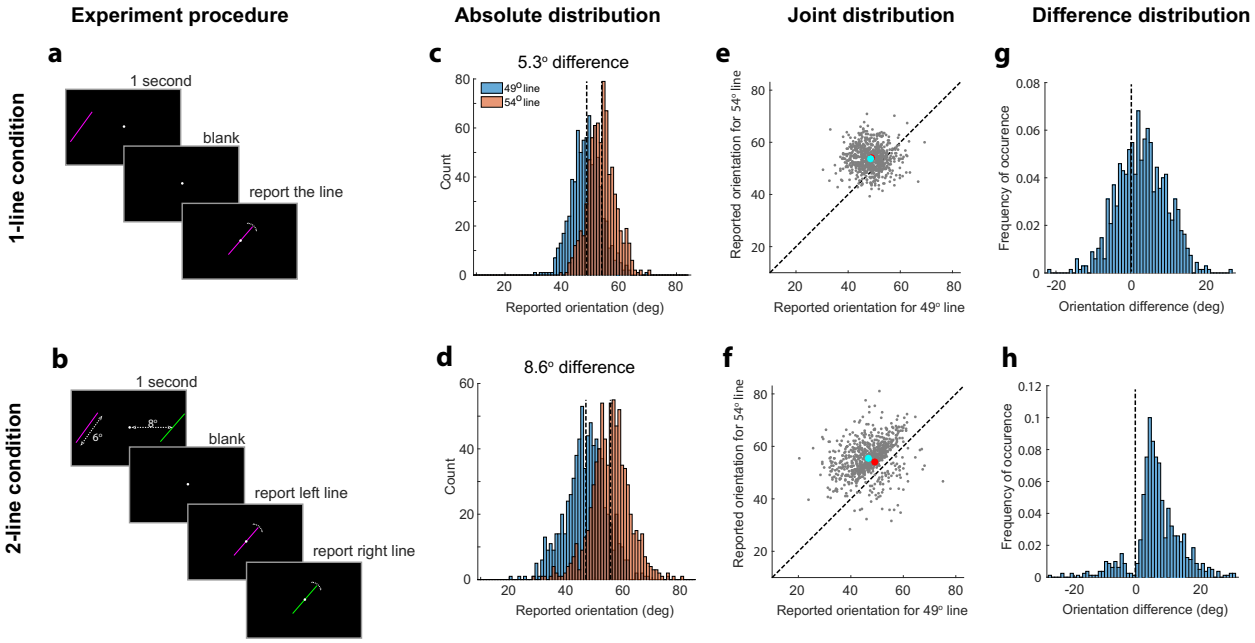

Figure 2: *The 1-line and 2-line conditions with data pooled from all 15 naive subjects.* (**a**) Trial sequence for the 1-line condition. (**b**) Trial sequence for the 2-line condition. For both conditions, during the blank after the stimulus disappearance, subjects drew a marker line for reporting. See text and Methods for details. (**c, d**) Reported distributions of the stimulus lines' absolute orientations for the 1-line and 2-line conditions, respectively. For each condition, the distributions for the 49° and 54° lines are in blue and orange, respectively. The dashed vertical lines indicate the means of the distributions. The difference between the means was greater in the 2-line condition than that in the 1-line condition, indicating repulsion. (**e**) Simulated joint distribution of the 2-line condition predicted from the 1-line data according to the low-to-high-level decoding scheme. (**f**) The measured joint distribution of the 2-line condition. The red dot indicates the true stimulus orientations, and the cyan dot indicates the means of the reports. The measured distribution showed a correlation between the two reports in a trial and a bimodal pattern with shifts away from the diagonal line whereas the joint distribution predicted by the low-to-high-level decoding did not. (**g-h**) The difference distributions (the 54° line minus the 49° line), obtained from the simulated and measured joint distributions in panels e and f, respectively. They are equivalent to projecting the joint distributions along the negative diagonal axis.

the 2-line condition was also greater than that of the 1-line condition (p = 0.00006, Wilcoxon sign rank test, see Fig. S1). Since the stimulus orientations and duration were exactly the same for the two conditions, the greater variance in the 2-line condition was likely due to increased memory noise because subjects had to hold two lines in working memory instead of one line.

We next examined the joint distribution of the two reported orientations in a trial of the 2-line condition. Fig. 2f plots the report for the 54° line against that for the 49° line. The distribution was elongated along the diagonal, indicating a positive correlation between the two reports in a trial (p = 0.00006, Wilcoxon sign rank test). The data points above and below the diagonal line were from the trials with correct and incorrect ordinal judgments, respectively. There was a gap between these two sets of trials as they shifted away from the diagonal (the decision boundary for the ordinal judgments), rendering the joint distribution bimodal. By subtracting the 49° report from the 54° report in a trial, we obtained the difference distribution (Fig. 2h), which was also bimodal. The difference distribution is equivalent to projecting the joint distribution along the negative diagonal axis, and the correct and incorrect trials are on the left and right sides of 0, respectively.

These results were quite similar to those of Ding et al. (2017). Importantly, Ding et al. presented the two lines successively at the same spatial location whereas here we presented them simultaneously on opposite sides of the fixation. This suggests that the two lines interacted similarly in working memory regardless of whether they were presented at the same or very different locations. Also similar to Ding et al.'s findings, the results of the 2-line condition cannot be explained by the low-to-high-level decoding scheme, which assumes that V1 cells in opposite hemispheres first decode the two lines' absolute orientations separately, which are then compared to determine their relationship. Obviously this decoding scheme cannot reproduce the observed interactions between the lines. We simulated this scheme's predicted joint distribution in Fig. 2e by randomly sampling pairs of orientations from the 49° and 54° distributions of the 1-line condition. This joint distribution is unimodal, and centered on, and evenly distributed around, the physical stimulus orientations, without the correlation, gap, and repulsion in the 2-line data. The predicted difference distribution is also unimodal, symmetrically centered on the actual difference between the two lines' orientations (5°), again unlike the measured difference distribution of the 2-line condition.

Although we pooled all subjects' data above, the interactions between the lines in the 2-line condition (repulsion and correlation) were consistently observed across all subjects (see Supplementary Fig. S2 for the joint distributions of all individual subjects). We computed each subject's repulsion and correlation in the 2-line condition, and compared with those predicted from the low-to-high-level decoding scheme applied to the 1-line data. Since the repulsion and correlation occurred separately for the correct and incorrect trials (trials above and below the diagonal line in Fig. 2f), we flipped (mirrored) the incorrect trials with respect to the diagonal line before the computation, and applied the same procedure to the simulated joint distributions from the 1-line

data. (Without the flipping, we would underestimate the repulsion and correlation, particularly for subjects with a large number of incorrect trials, because the repulsion values in the two opposite directions away from the diagonal would cancel each other, and the separation, along the negative diagonal, of the two positive-diagonal elongations would reduce the actual correlation.) Fig. 3a shows that all 15 subjects reported greater orientation difference in the 2-line condition than in the 1-line condition. We computed the 95% confidence interval using bootstrapping for each subject, and found that for 12 out of the 15 subjects, the 95% confidence interval did not touch the diagonal line in Fig. 3a. Therefore, the repulsion in the 2-line condition is significant for the majority of subjects individually. Fig. 3b shows that the correlation in the 2-line condition was greater than that in the 1-line condition. Again, the 95% confidence interval for each subject calculated with bootstrapping indicates that 12 out of the 15 subjects showed significant correlation individually.

Finally, Fig. 3c shows that the ordinal discrimination performance in the 2-line condition was better than that predicted by the 1-line data according to the low-to-high-level decoding (mean accuracy: 90% vs. 77%). The difference is significant at the group level ($p = 0.025$, Wilcoxon sign rank test). This can also be seen in the joint and difference distributions in Fig. 2 which show a larger portion of correct trials in the 2-line condition compared to the 1-line condition. At the individual level, 11 out of the 15 subjects showed the same trend (Fig. 3c) although only 5 subjects reached significance based on the bootstrapping test.

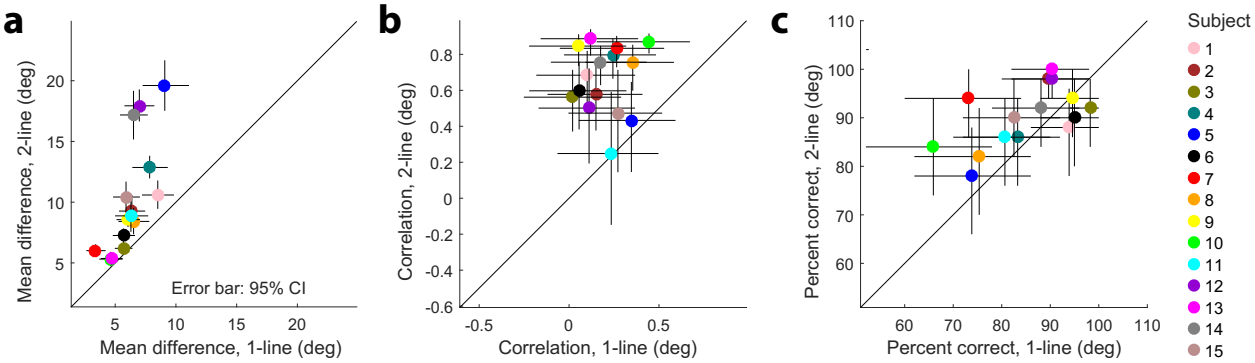

Figure 3: *Comparison of the 1-line and 2-line conditions for individual subjects.* Each color represents one subject. (**a**) The mean absolute difference between the reports for the 49° and 54° lines. (**b**) The correlation coefficient between the two reports in a trial. The correlation for the 1-line condition was based on the prediction of the low-to-high-level decoding. (**c**) The percentage correct of ordinal discrimination between the two lines. The percentage correct for the 1-line condition was based on the prediction of the low-to-high-level decoding. All error bars were 95% confidence intervals obtained by bootstrapping 10,000 times.

The above results suggest that when stimulus orientations are decoded in working memory, they interact with each other even when presented on opposite sides of the fixation. However,

there are two potential confounds and we address them below.

Orientation interactions under a different report method

The first potential confound is that in the 2-line condition above, subjects rotated the marker line continuously from the first report to the second report, and this continuity might introduce interactions artificially. For example, subjects might over-rotate to ensure that the two reports were different even though the instructions emphasized accuracy. This, however, was unlikely because the actual 5° orientation difference was well above the orientation discrimination threshold of around 1° at fovea where the marker line was placed. To directly address any potential problems of the continuous report method above, we did a control experiment by running the same group of subjects on the 2-line condition with an interruption between the two reports (2-line-interrupt condition). It was identical to the above 2-line condition except that after subjects clicked to report the left orientation, the marker line immediately disappeared and subjects had to move the mouse to redraw the marker line according to their perceived right orientation, adjusted it to match the perception as closely as possible, and then clicked (Fig. 4a). This method was very similar to that used by Bae and Luck (2017) but they presented stimuli at fovea and did not measure cross-fixation interactions. We planned both reporting methods before the data collection and randomized the order of the 2-line and 2-line-interrupt conditions across subjects.

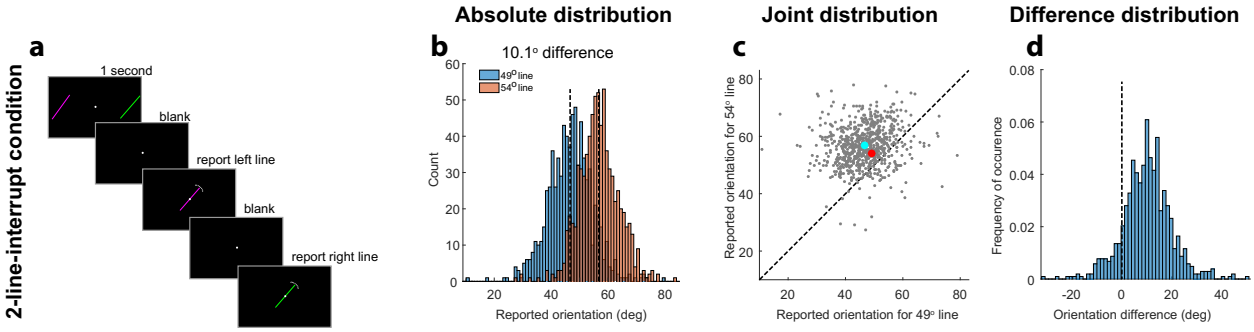

Figure 4: *The 2-line-interrupt condition with data pooled from all 15 naive subjects.* The plot format is identical to that of Fig. 2. (**a**) Trial sequence for the 2-line-interrupt condition. During each blank, subjects drew a marker line for reporting. The second blank interrupted the continuity of the two reports. See text and Methods for details. (**b**) Reported distributions of the stimulus lines' absolute orientations, showing even larger repulsion between the 49° and 54° lines than that for the 2-line condition (Fig. 2d). (**c**) The joint distribution, showing much reduced correlation and bimodality compared with the 2-line condition (Fig. 2f). (**d**) The distribution of the difference between the two reported orientations, again showing a much reduced bimodality compared with the 2-line condition (Fig. 2h).

The results pooled across all subjects are shown in Fig. 4. The distributions of the reported absolute orientations (Fig. 4b) showed a significant repulsion compared with the 1-line condition ($p = 0.0003$, Wilcoxon sign rank test). In fact, the repulsion in the 2-line-interrupt condition (mean orientation difference 10.1°) was even larger than that in the 2-line condition (mean orientation difference 8.6°). However, the interrupt report method changed the joint distribution of the two reports in a trial (Fig. 4c). Although the joint distribution shifted away from the diagonal, there was no clear gap between the correct and incorrect trials along the diagonal, and the difference distribution was unimodal (Fig. 4d). The correlation between the two reports in a trial was much reduced though still significant ($p = 0.035$, Wilcoxon sign rank test).

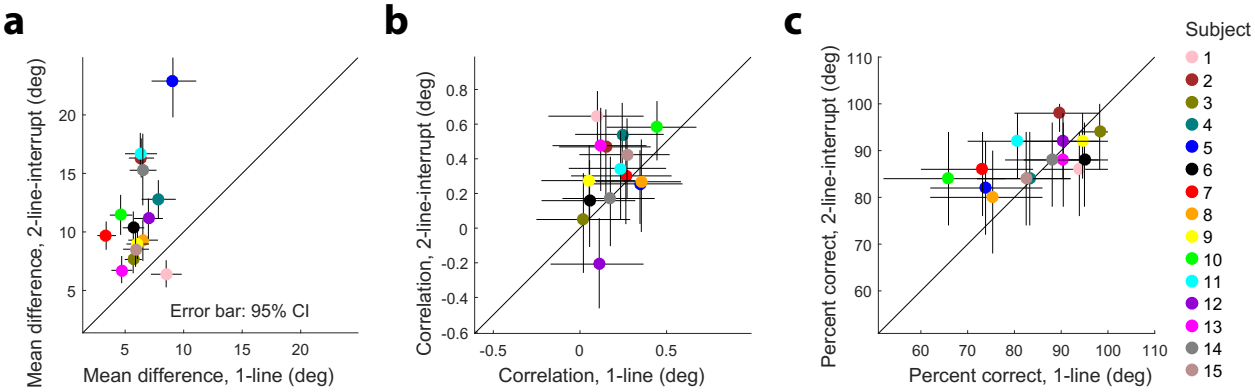

Figure 5: *Comparison of the 1-line and 2-line-interrupt conditions for individual subjects.* The plot format is identical to that of Fig. 3. (**a**) The mean absolute difference between the reports for the 49° and 54° lines. (**b**) The correlation coefficient between the two reports in a trial. The correlation for the 1-line condition is based on the prediction of the low-to-high-level decoding in Fig. 2e. (**c**) The percentage correct of ordinal discrimination between the two lines. Note that the subjects did not explicitly perform the ordinal discrimination task so the percent correct was inferred from their reported absolute orientations of the stimuli. All error bars were 95% confidence intervals obtained by bootstrapping 10,000 times.

We also analyzed the 2-line-interrupt data for each subject individually, as we did for the 2-line condition. We found that 14 out of 15 subjects showed a significant repulsion (Fig. 5a), and the repulsion magnitudes were generally greater than those for the 2-line condition (cf. Fig. 3a). The correlations were weaker than those for the 2-line condition (cf. Figs. 5b and 3b). This can also be seen from shapes of individual subjects' joint distributions of the 2-line-interrupt condition in Supplementary Fig. S3. Although some subjects showed similar joint distributions for the two report methods, others showed little elongation along the diagonal under the interrupted report method. Finally, Fig. 5c shows the subjects' ordinal discrimination performances; unlike the original 2-line condition, the mean was not significantly better than that predicted by the 1-line data ($p = 0.23$). This is perhaps not surprising because the interruption must make it difficult (and unnecessary) for

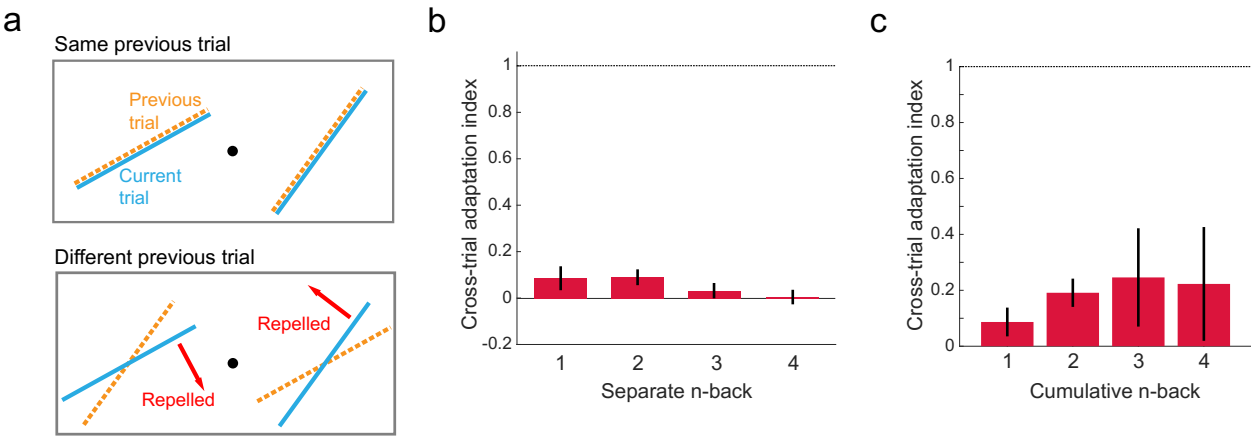

Figure 6: *Cross-trial adaptation at the same site cannot explain the observed repulsion* (**a**) The orientations of the current (blue) and a previous (yellow) trial can be either the same (top) or different (bottom). The "different" case could produce cross-trial adaptation aftereffect whereas the "same" case could not. (**b**) n-back cross-trial adaptation index for the 2-line condition, with $n =$ 1, 2, 3 and 4 separately. The index values of 0 and 1 indicate that cross-trial adaptation explains none and all of the measured repulsion, respectively. (**c**) Cumulative n-back cross-trial adaptation index for the 2-line condition, with $n =$ 1, 2, 3 and 4. All error bars represent $\pm$ 1 SEM. They grow with n in panel c because the number of available data points is halved for each increment of cumulative n.

the subjects to indicate the ordinal relationship through the two absolute reports. In other words, the ordinal discrimination performances calculated from the 2-line-interrupt data did not reflect the subjects' actual ordinal discrimination performances.

In sum, interrupting the continuity of the two reports in a trial did not eliminate the cross-fixation interactions of orientations. Both the repulsion and correlation remained significant at the group level. Although the correlation was much weaker, the repulsion appeared stronger. We will explain these data and their differences in a modeling section later.

The repulsion cannot be explained by adaptation across trials at the same site

Another potential confound of the 2-line condition is that the observed repulsion might be explained by traditional adaptation aftereffects across trials on the same side of the fixation. Specifically, at a given site, if the stimulus orientation in the current trial was different from that in a previous trial, subjects' perceived orientation in the current trial could be repelled away from the orientation of the previous trial (Fig. 6a). However, if the stimulus orientations for the two trials were identical, then there would be no adaptation-induced repulsion (Gibson and Radner, 1937). We first note that such cross-trial adaptation aftereffects must be small because of the long intervals between

stimuli of successive trials (around 8 sec for the 2-line condition) compared to the stimulus duration (1 sec). It might be further reduced by the attractive, serial effect (Fischer and Whitney, 2014). Nevertheless, we analyzed this possibility in great detail. First, we split each subject's 2-line data into the "same" and "different" sets according to whether the stimulus orientation in a trial and that n trials back were identical or not, for n = 1, 2, 3, and 4. We quantified the n-back cross-trial adaptation effect by calculating the index $(R_d - R_s)/(R_d + R_s)$, where R_d and R_s are the repulsion of the "different" and "same" sets, respectively. If the repulsion all came from the n-back cross-trial adaptation, instead of from cross-fixation interactions, then R_s would be 0, and the index would be 1. Conversely, if the repulsion all came from cross-fixation interactions, then R_d and R_s would be identical, and the index would be 0. The results are shown in Fig. 6b. We found that as expected, the contribution of the cross-trial adaptation to the repulsion was small even for $n = 1$ and disappeared for $n = 4$. The sum of the indices across n is around 0.2, well below 1, and thus cannot account for the observed repulsion. Second, we investigated the possibility that different n-back adaptation effects might sum superlinearly to explain the repulsion. We therefore determined the cumulative n-back adaptation effect directly, instead of summing the separate n-back effects. To this end, we defined the "same" and "different" sets according to whether the stimulus orientation of a trial were identical to, or different from, those of all n previous trials (which had to have the same orientation). The results in Fig. 6c show that the n-back cumulative effect had the index saturated around 0.25 for $n = 3$, again well below 1. The error bar grew with n because when n increased by 1, the available data was halved. We conclude that traditional adaptation aftereffects across trials at the same site cannot explain the repulsion in the 2-line condition.

The first and second reports in a trial showed similar repulsion

Ding et al. (2017) presented two lines in a trial sequentially (and subjects reported them sequentially); this allowed them to measure both the forward aftereffect (how much the first line repelled the second line) and the backward aftereffect (how much the second line repelled the first line). They found that the two aftereffects were similar for a given subject. As they noted, this result contradicts standard adaptation theories whose sequential considerations of sensory responses predict only the forward aftereffect, and prompted them to propose high-to-low-level decoding in working memory. In the current study, we presented two lines in a trial simultaneously so the forward and backward aftereffects were not defined. Nevertheless, subjects had to report the two lines sequentially, and we analyzed whether the first and second reports of a line were similar or not. For both the 2-line and 2-line-interrupt conditions, we calculated the means of the first and second reports for the 49° line separately, and did the same for the 54° line. Using the means of the 49° and 54° lines of the 1-line condition as the baselines, we determined the repulsion values for each line when it was reported first and second. The results (Fig. 7) indicate that the

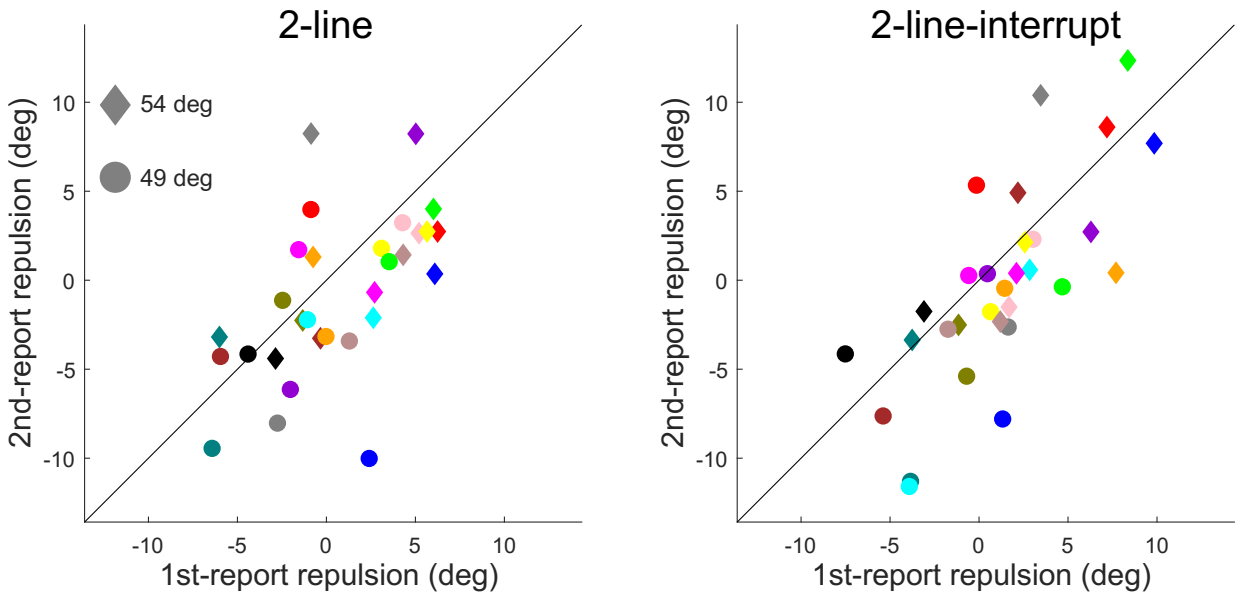

Figure 7: *First and second reports showed similar repulsion.* The left and right panels are for the 2-line and 2-line-interrupt conditions, respectively. In each panel, the second-report repulsion is plotted against the first-report repulsion across subjects. Each subject had two data points, one for the 49° line (round dot) and the other for the 54° line (diamond).

first and second reports showed similar repulsion, analogous to the similar backward and forward aftereffects in Ding et al. (2017).

High-to-low-level Bayesian decoding explains the data from both report methods

The cross-fixation interactions of orientations established above, in the form of repulsion and correlation, cannot be explained by the low-to-high-level decoding scheme (Figs. 2-5). We thus adopted Ding et al. (2017)'s high-to-low-level decoding scheme to account for the data. The main hypothesis is that in a 2-line trial, subjects (implicitly) judged the lines' ordinal relationship and used this higher-level information to constrain the decoding of the lower-level, absolute orientations of the lines (Ding et al., 2017). To explain the differences between the two report methods, we applied the scheme twice to take into account the interruption in the second report method, as detailed below.

We started with the 1-line condition; we simply assumed that subjects made a noisy sensory measurement of the stimulus line's orientation in a trial. Then the sensory sample was corrupted by memory noise to produce a memory sample. We assumed both the sensory and memory noises are Gaussian. With a uniform prior on orientation, the posterior was the same as the likelihood function which was a Gaussian around the memory sample. Consequently, the decoded estimate

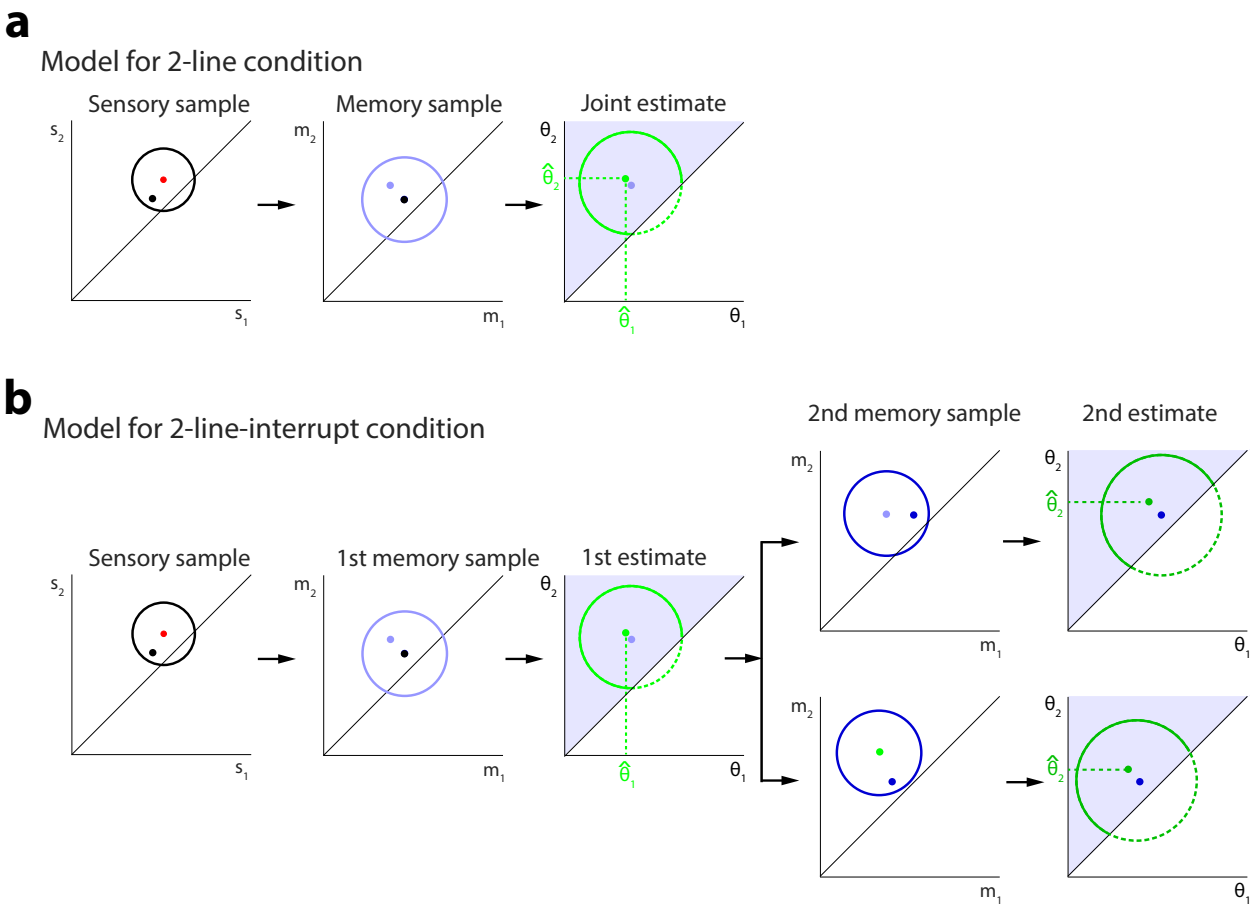

Figure 8: *High-to-low-level Bayesian decoding scheme.* (**a**) Model for the 2-line condition. First panel: a sensory sample (black dot) is drawn from the sensory distribution of the two lines (black circle) centered on the stimulus orientations (red dot). Second panel: a memory sample (blue dot) is drawn from the memory distribution (blue circle) centered on the sensory sample (black dot). Third panel: The posterior distribution (solid green arc above diagonal) is obtained by multiplying the likelihood function (green circle) centered on the memory sample (blue dot) and a Bayesian prior (shaded step function along the diagonal) from the ordinal judgment. The posterior mean is the decoded estimate of the two orientations (green dot). (**b**) Model for the 2-line-interrupt condition. It is similar to the 2-line model above except that the memory decoding process is repeated, one before and the other after the interruption, and each process reports only one of the two estimated orientations. The second decoding process is represented by the darker blue and green colors. The distribution (dark blue circle) for the second memory sample (dark blue dot) can be centered either on the first memory sample (top row) or on the first estimate (bottom row), resulting in two versions of the model.

was identical to the memory sample. These estimates were used in the above simulations of the low-to-high-level decoder that used the 1-line data to predict the 2-line data. As noted above, the low-to-high-level predictions did not match the data.

In the high-to-low-level decoding model for the 2-line condition (Fig. 8a), we also started with drawing sensory and memory samples (black and light blue dots, respectively) for a trial according to the sensory and memory noise distributions (black and light blue circles, respectively). The key difference was that the prior was not uniform but determined by subjects' ordinal judgment based on the sensory measurements. For instance, if the ordinal judgment was that the 54° orientation was greater than the 49° orientation, then the prior was non-zero only above the diagonal line in the joint space (the shaded region in the last panel of Fig. 8a). As a result, combining the likelihood function (green circle) and this step-function prior led to a posterior distribution (solid green arc) whose center of mass (green dot), the decoded estimate, was shifted away from the diagonal. Note that here we modeled sensory and memory noises separately instead of grouping them together as in (Ding et al., 2017). The reason was that here the stimulus lines were presented simultaneously so that subjects could make ordinal judgments based solely on the sensory evidence. As explained in Ding et al. (2017), the binary, ordinal judgments were assumed to be resistant to memory noise. In contrast, the continuous, absolute orientations of the lines were degraded by the memory noise.

For the 2-line-interrupt condition, we used the same high-to-low-level decoding scheme as for the 2-line condition but we assumed that there were two decoding processes (Fig. 8b), one before, and the other after, the interruption (the disappearance of the marker line). Specifically, the first decoding process was identical to that for the 2-line condition. However, although both absolute orientations were decoded, only the left orientation was reported before the interruption. With the redrawing of the marker line after the interruption, we assumed a repeat of the decoding process but this time only the right orientation was reported. The second memory sample could be based on the first memory sample but further corrupted by the memory noise (Fig. 8b, first row). Alternatively, it could be based on the first estimate, also further corrupted by the memory noise (Fig. 8b, second row). We considered both versions of the model. We let the additional memory noise for the second decoding be the same as that for the first decoding because the reaction times of the two reports were similar in the 2-line-interrupt condition (see Supplementary Fig. S4).

The free parameters were the standard deviations for the sensory and memory noises (see Methods). We first jointly fit the parameters by maximizing the likelihood of the data of the 1-line and 2-line conditions. The resulting model matched the data well (Fig. 9, the first two columns). Notably, the model reproduced the characteristic repulsion, correlation, and the bimodel pattern in the 2-line joint distribution (Fig. 9, second column). We then used the fit parameters to generate

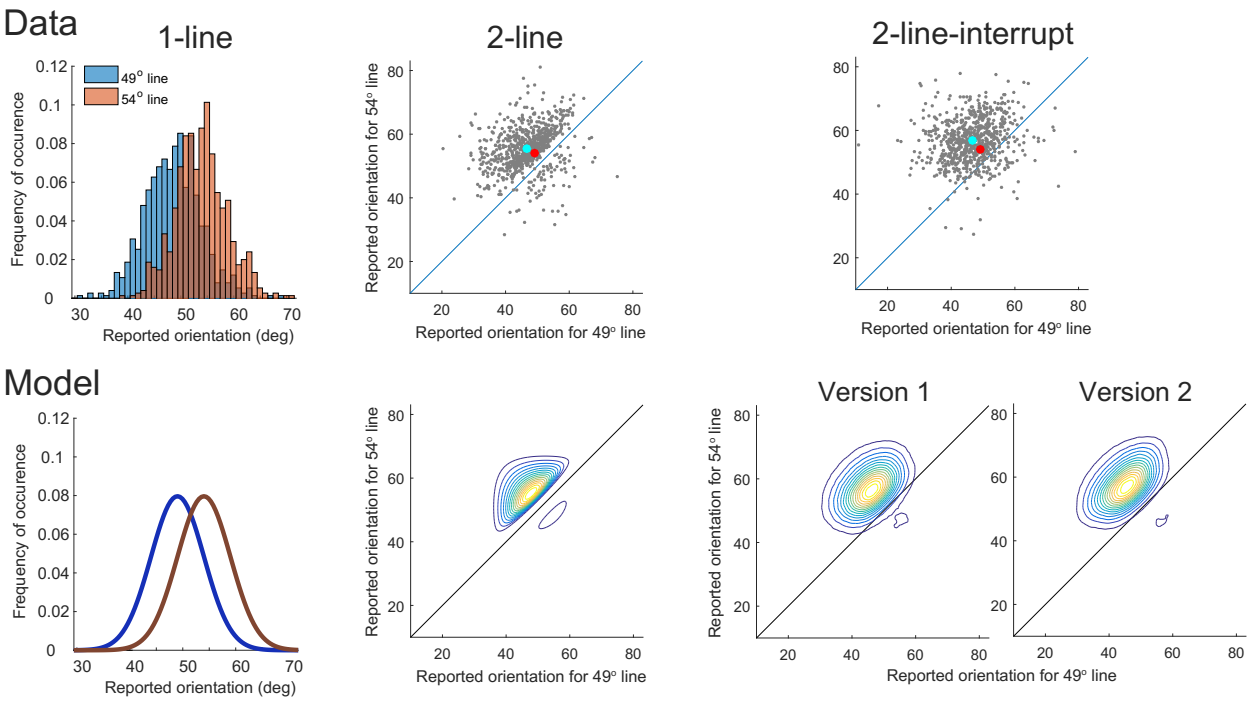

Figure 9: *Model fit of the 1-line and 2-line data, and prediction of the 2-line-interrupt data.* The first row shows the data (pooled over all subjects) and the second row shows the model fit or prediction. The first column shows the absolute distributions of the 1-line condition. The second column shows the joint distribution of the 2-line condition. The third column shows the joint distribution of the 2-line-interrupt condition, with two different model versions.

parameter-free predictions for the 2-line-interrupt condition. Both model versions for the 2-line-interrupt conditions (Fig. 9, last column) fit the data similarly well. We also compared the measured and the modeled difference distributions in Fig. 10, again showing good agreements.

Discussion

In this study, we tested a prediction of Ding et al. (2017)'s theory positing that visual decoding often occurs in working memory where it progresses from high- to low-level features, with higher-level features, which are more invariant, reliable, and behaviorally relevant, constraining the decoding of lower-level features. Since the high-to-low-level constraints introduce interactions between lower-level features, the theory predicts that low-level features that are traditionally considered as independent may interact with each other when they are decoded in working memory.

In our experiment, the lower- and higher-level features were the absolute orientations of two lines (on opposite sides of the fixation) and their ordinal relationship, respectively. Their *encoding* likely follows the standard low-to-high-level hierarchy of sensory responses (Hubel and Wiesel, 1968;

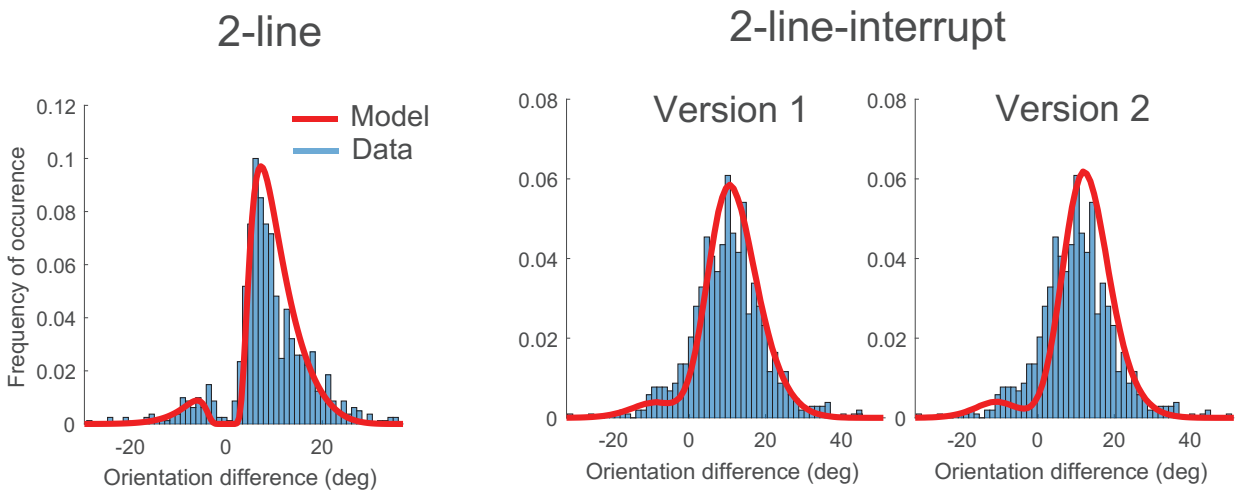

Figure 10: *Model fit and prediction of the difference distributions.* The first panel shows the fit (red curve) to the 2-line difference distribution (blue histogram). The last two panels shows the two model versions' predictions (red curves) of the 2-line-interrupt data (blue histogram).

Riesenhuber and Poggio, 1999; Anzai et al., 2007). The traditional view is that their *decoding* follows the same hierarchy, and the absolute orientations are decoded in an early visual area with small receptive fields and thus should be mutually independent. In contrast, according to Ding et al. (2017), the encoded absolute orientations and their ordinal relationship enter working memory after the disappearance of the stimuli. During the delay before the reports, the stored binary ordinal relationship is noise resistant whereas the continuous absolute orientations are corrupted by noise over time. By the report time, the brain decodes the reliable ordinal relationship first, and uses it to constrain the decoding of the unreliable absolute orientations, producing interactions between the absolute orientations. Using an eye-tracker to ensure fixation, we indeed found the predicted cross-fixation interactions of the orientations in the form of repulsion and correlation. Control experiments and analyses ruled out alternative explanations such as reporting-method bias and cross-trial adaptation aftereffects on the same side of the fixation. Finally, we showed that Ding et al. (2017)'s retrospective Bayesian decoding model well fit the 2-line data, and without new free parameters, predicted the 2-line-interrupt data. Unlike many Bayesian models that adjust priors to fit the data, in our simulations, the prior is a step function fully determined by the ordinal judgment and only the likelihood function has free parameters.

We used a continuous and an interrupt report method for the 2-line and 2-line-interrupt conditions, respectively. The continuous report method was nearly identical that of Ding et al. (2017), and the 2-line data here resembled those of Ding et al. (2017) showing repulsion and correlation between the two reported orientations in a trial. The interrupt report method was nearly identical to that of Bae and Luck (2017), and our 2-line-interrupt data were similar to those of Bae and Luck (2017), showing repulsion but reduced correlation. Importantly, however, we placed the

two orientations on opposite sides of the fixation whereas both Ding et al. (2017) and Bae and Luck (2017) placed them (successively) at the fixation. Therefore, the current study demonstrated cross-fixation interactions of orientations whereas the two previous studies were not designed to do so. On the other hand, the three studies collectively indicate that when two orientations are both task relevant and decoded in working memory, they interact with each other regardless of whether they appear on the same or different retinal locations. In addition to retinal locations, these studies also differ in stimulus shape, size, eccentricity, and duration, the magnitude of orientation difference, and simultaneous vs. sequential presentations. The fact that they still produced similar results suggests that stimulus interactions in working memory are a robust phenomenon.

Both frontal/parietal areas and various sensory cortices have been implicated in working memory (Pasternak and Greenlee, 2005). Since working memory does not necessarily require sustained neuronal firing after stimulus disappearance (Mongillo et al., 2008), it could in principle reside even in low-level sensory areas. However, the working memory area for orientation decoding in our experiments is likely a high-level area that does not maintain fine retinotopy but instead, let relevant features from different locations affect each other. A related finding is the transfer of perceptual learning between well-separated retinal locations under certain training procedures (Xiao et al., 2008). For example, contrast training at one location transferred to another location that only received orientation training. Although there are key differences between short-term working memory and long-term perceptual learning, these studies, and that of Ding et al. (2017), suggest that perceptual decoding of low-level features could occur in high-level brain areas where the binding or integration of the features may produce various interactions among them across space and time. Alternatively, low-level features might be stored retinotopically in low-level sensory areas which are modulated by high-level feedback connections to produce perceptual interactions (Pasternak and Greenlee, 2005). In either case, high-level processing must be involved in the decoding of low-level features.

Sensory processing and working memory are often treated as separate topics in the literature. Our theory, however, explicitly integrates them by proposing that decoding of perceptual judgments may happen in working memory. It is this integration that provides a key reason that decoding should proceed from high- to low-level features (Ding et al., 2017), which then leads to our prediction of cross-fixation interactions of orientations (see Introduction). Since sensory processing includes both encoding and decoding, we consider decoding in working memory as part of sensory processing. Alternatively, one may argue that working memory should not be included in sensory processing. Accordingly, our framework becomes the following: sensory processing proceeds from low- to high-level features, and high-to-low-level decoding of perceptual judgments in working memory should just be viewed as a memory process, not part of the sensory process. We note that this is mostly an issue of definition that does not change our reasoning on why de-

coding in working memory should proceed from high- to low-level features or how higher-level features should constrain the decoding of lower-level features.

Binary ordinal judgments could also be viewed as perceptual decisions. So an equivalent interpretation of our model is that the perceptual decision on the ordinal relationship provides a prior to constrain the decoding of the absolute orientations. What is important, however, is not the different choices of terminology, but the common theme that the higher-level ordinal relationship between two lines affects the decoding of the lower-level absolute orientations of the individual lines. Therefore, the decision interpretation is consistent with our high-to-low-level decoding hierarchy. On the other hand, without the consideration of different noise tolerance of low- vs. high-level features in working memory, the decision interpretation alone misses a key reason of why the high-to-low-level decoding scheme is desirable (Ding et al., 2017). Also note that the binary, ordinal decision was not a separate task imposed on the subjects. For the 2-line-interrupt condition, the ordinal decision was not even implicitly required. Our study is therefore different from typical task-dependence studies where the tasks in question are usually required. Additionally, unlike our theory, task dependence alone does not provide a reason on why decoding should proceed from high- to low-level features in working memory or how higher-level features should constrain the decoding of lower-level features.

As noted above, according to Ding et al. (2017), interactions between lower-level features in working memory stems from higher-level constraints on lower-level decoding. In our experiments, the lower-level features were the individual, absolute orientations of the two lines, and the higher-level feature was their ordinal relationship. For the continuous report method (2-line condition), subjects implicitly indicated their ordinal choice when rotating the marker line continuously from the first report to the second report of the absolute orientations. In contrast, for the interrupt report method (2-line-interrupt condition), because the marker line disappeared after the first report, subjects could not use the continuous rotation to indicate their ordinal choice. The fact the the 2-line-interrupt data can be explained by the same high-to-low-level decoding scheme (applied twice but without new free parameters) suggests that the ordinal relationship was still decoded first, which then constrained the absolute decoding, even when its reporting was not required. The reason, we believe, is that the ordinal relationship is more invariant against viewing conditions, more reliable against memory noise, and more behaviorally useful, than the absolute orientations so that the brain may automatically prioritize its decoding. When the ordinal relationship is decoded correctly, it can then help improve the decoding of less reliable, absolute orientations through the high-to-low-level constraint (Ding et al., 2017). High-to-low-level decoding in noisy working memory could be a general principle for understanding perception.

Acknowledgment

Supported by NSF grant 1754211, AFOSR grant FA9550-15-1-0439, and Irving Weinstein Foundation Inc.

References

Peggy Seriès, Alan A Stocker, and Eero P Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural Computation*, 21(12):3271–3304, 2009.

Li Zhaoping. *Understanding vision: theory, models, and data*. Oxford University Press, USA, 2014.

Daniel J Felleman and DC Essen Van. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, 1991.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

MA Paradiso. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological cybernetics*, 58(1):35–49, 1988.

H Sebastian Seung and Haim Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, 1993.

Arnulf BA Graf, Adam Kohn, Mehrdad Jazayeri, and J Anthony Movshon. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature neuroscience*, 14(2):239, 2011.

Andrew F Teich and Ning Qian. Learning and adaptation in a recurrent model of v1 orientation selectivity. *Journal of Neurophysiology*, 89(4):2086–2100, 2003.

Stephanie Ding, Christopher J Cueva, Misha Tsodyks, and Ning Qian. Visual perception as retrospective bayesian decoding from high-to low-level features. *Proceedings of the National Academy of Sciences*, 114(43):E9115–E9124, 2017.

John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923, 2000.

Vladimir Itskov, David Hansel, and Misha Tsodyks. Short-term facilitation may stabilize parametric working memory trace. *Frontiers in computational neuroscience*, 5:40, 2011.

James J Gibson and Minnie Radner. Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of experimental psychology*, 20(5):453, 1937.

Xin Meng, Pietro Mazzoni, and Ning Qian. Cross-fixation transfer of motion aftereffects with expansion motion. *Vision research*, 46(21):3681–3689, 2006.

Hong Xu, Peter Dayan, Richard M Lipkin, and Ning Qian. Adaptation across the cortical hierarchy: Low-level curve adaptation affects high-level facial-expression judgments. *Journal of Neuroscience*, 28(13):3374–3383, 2008.

Helen C Beh, Peter M Wenderoth, and AT Purcell. The angular function of a rod-and-frame illusion. *Perception & Psychophysics*, 9(4):353–355, 1971.

L. Luu and A.A. Stocker. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, In press. 2018.

Alan A Stocker and Eero P Simoncelli. A bayesian model of conditioned perception. In *Advances in neural information processing systems*, pages 1409–1416, 2008.

Cheng Qiu, Long Luu, and Alan A Stocker. Benefits of commitment in hierarchical inference. *Psychological review*, 127(4):622, 2020.

Matthias Fritsche and Floris P de Lange. Reference repulsion is not a perceptual illusion. *Cognition*, 184:107–118, 2019.

M. Jazayeri and J.A. Movshon. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446:912ff, April 2007.

E. Zamboni, T. Ledgeway, P.V. McGraw, and D. Schluppeck. Do perceptual biases emerge early or late in visual processing? decision-biases in motion perception. *Proc. of Royal Society of London B*, 283(1833), 2016. doi: 10.1098/rspb.2016.0263.

Z. Z. Bronfman, N. Brezis, R. Moran, K. Tsetsos, T. Donner, and M. Usher. Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1810), 2015. doi: 10.1098/rspb.2015.0228.

Bharath Chandra Talluri, Anne E Urai, Konstantinos Tsetsos, Marius Usher, and Tobias H Donner. Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, 28(19):3128–3135, 2018.

Gi-Yeul Bae and Steven J Luck. Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, 79(8):2376–2395, 2017.

Qinglin Li, Andrew Isaac Meso, Nikos K Logothetis, and Georgios A Keliris. Scene regularity interacts with individual biases to modulate perceptual stability. *Frontiers in neuroscience*, 13, 2019.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.

Lin Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982.

Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–464, 2004.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.

Gail A Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1): 54–115, 1987.

Shimon Ullman. Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral cortex*, 5(1):1–11, 1995.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

Long Luu, Mingsha Zhang, Misha Tsodyks, and Ning Qian. Cross-fixation interactions of orientations suggest that orientation decoding occurs in a high-level area of visual working memory. *Journal of Vision*, 20(11):216–216, 2020.

David H Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature neuroscience*, 17(5):738–743, 2014.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area v2 encode combinations of orientations. *Nature neuroscience*, 10(10):1313–1321, 2007.

Tatiana Pasternak and Mark W Greenlee. Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6(2):97–107, 2005.

Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008.

Lu-Qi Xiao, Jun-Yun Zhang, Rui Wang, Stanley A Klein, Dennis M Levi, and Cong Yu. Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, 18(24):1922–1926, 2008.

Supplementary information

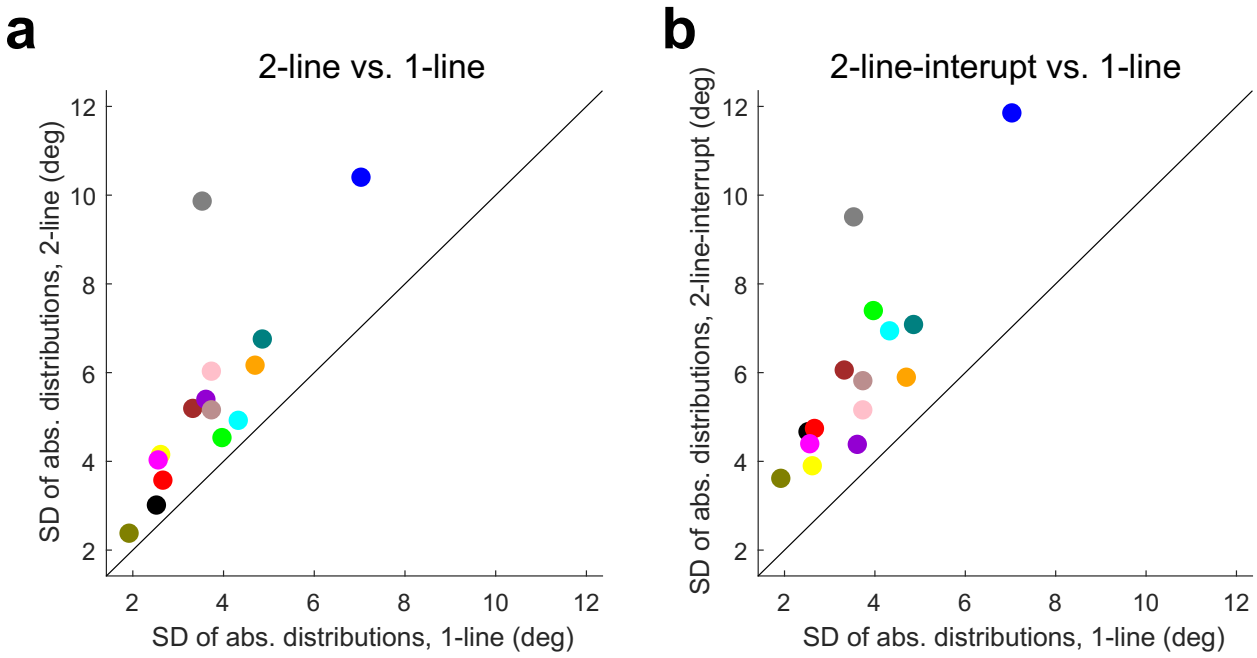

Figure S1: *More memory noise in the 2-line and 2-line-interrupt conditions than in the 1-line condition.* For each condition, the SD is the square root of a subject's mean of the variances for the $49°$ and $54°$ absolute distributions. (**a**) The 2-line condition vs. the 1-line condition. (**b**) The 2-line-interrupt condition vs. 1-line condition.

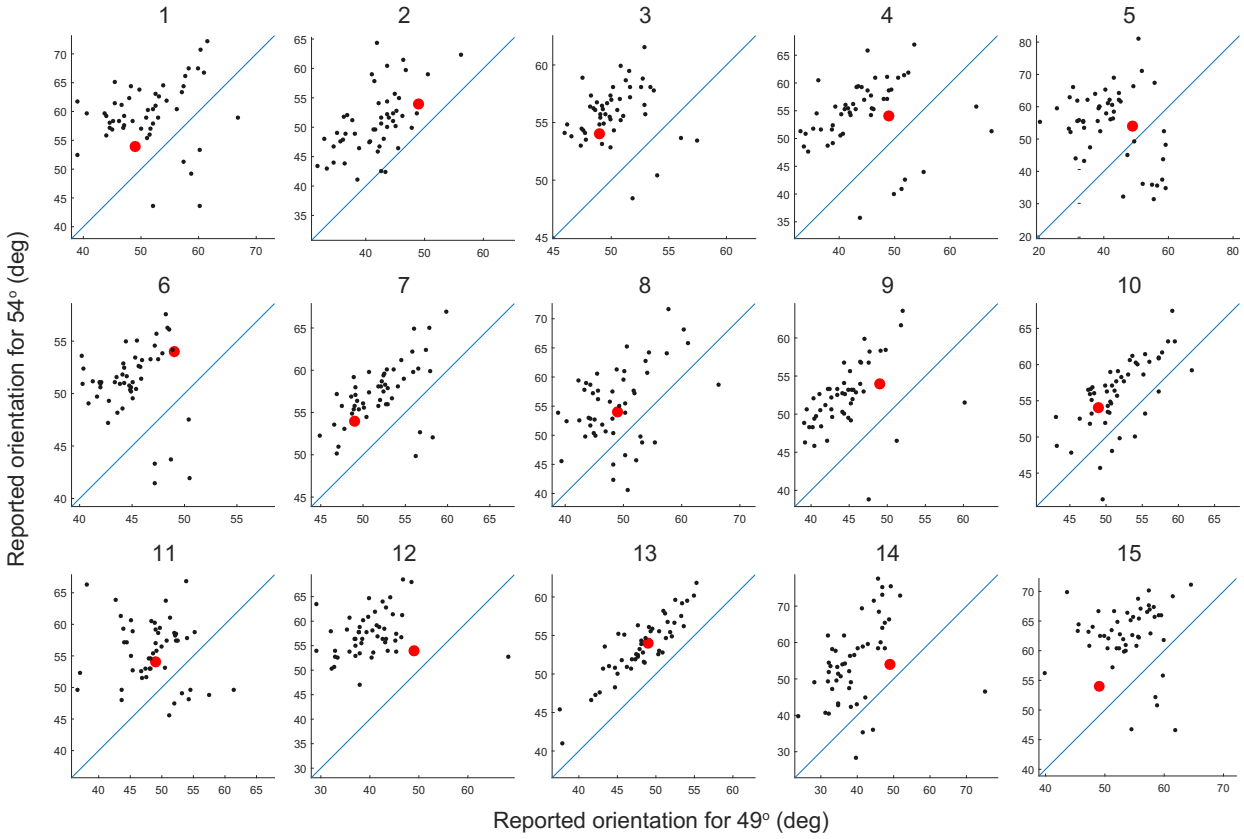

Figure S2: *Joint distributions of individual subjects in the 2-line condition.*

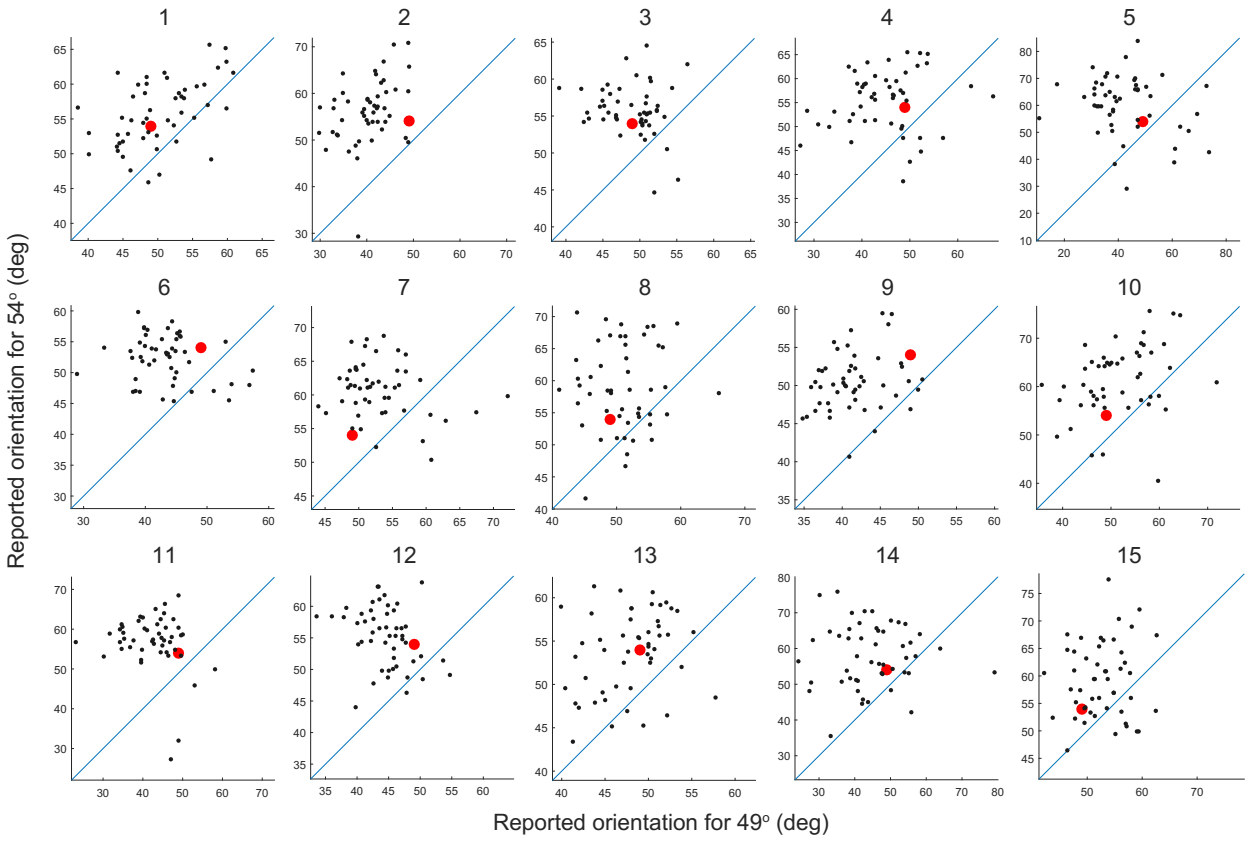

Figure S3: *Joint distributions of individual subjects in the 2-line-interrupt condition.*

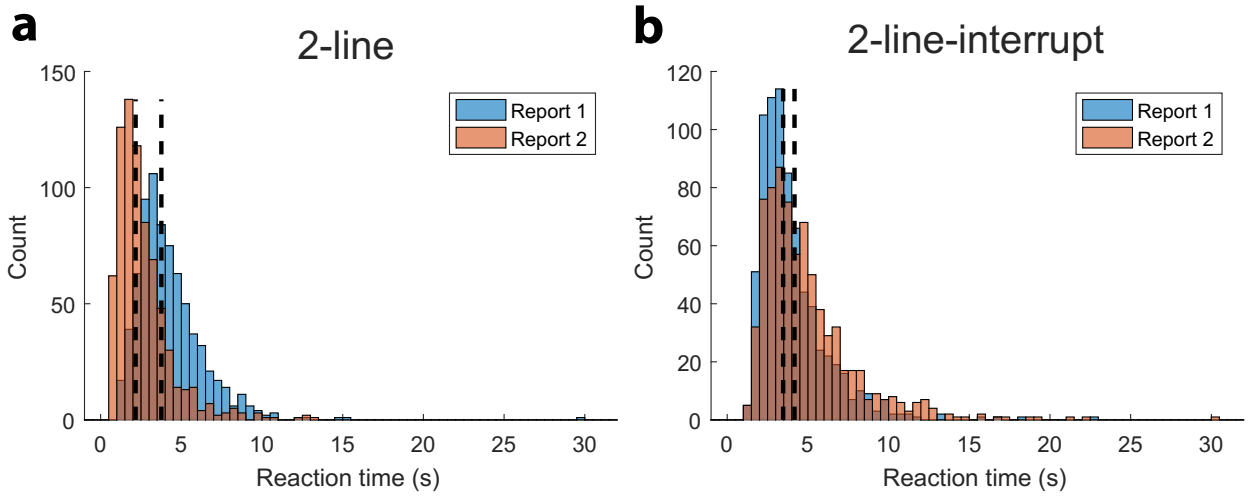

Figure S4: *Reaction time distributions in the 2-line (**a**) and 2-line-interrupt (**b**) conditions.* Data are pooled from all subjects. In each panel, the reaction time (RT) distributions for the first and second reports in a trial are shown in blue and orange, respectively. The vertical lines indicate the means. The mean RT difference between the two reports in a trial were 1.6 and -0.7 sec for the 2-line and 2-line-interrupt conditions, respectively.