Center-surround inhibition by expectation: a neuro-computational account

Ling Huang[1,2,6,7], Shiqi Shen[1,2,6], Yueling Sun[1,2,6], Shipei Ou[2], Ruyuan Zhang[3,4], Floris P. de Lange[5] & Xilin Zhang[1,2*]

[1]Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, South China Normal University, Guangzhou, Guangdong 510631, China. [2]School of Psychology, Center for Studies of Psychological Application, and Guangdong Provincial Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, Guangdong 510631, China. [3]Institute of Psychology and Behavioral Science, Antai College of Economics and Management, and Shanghai Mental Health Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 20030, China. [4]Key Laboratory of Brain-Machine Intelligence for Information Behavior- Ministry of Education, Shanghai International Studies University, Shanghai 200083, China. [5]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, 6525 EN Nijmegen, the Netherlands. [6]These authors contributed equally. [7]Current address: Department of Psychology, The Ohio State University, Columbus, OH 43201, USA.

*Correspondence: xlzhang@m.scnu.edu.cn

Xilin Zhang, Prof.

Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education,

School of Psychology, South China Normal University, Guangzhou, Guangdong, China

Tel: + (86)186-8283-9781

Website: https://xlzhanglab.scnu.edu.cn/

Abstract

Expectation is beneficial for adaptive behavior through quickly deducing plausible interpretations of information. The profile and underlying neural computations of this process, however, remain unclear. When participants expected a grating with a specific orientation, we found a center-surround inhibition profile in orientation space, which was independent from attentional modulations by task-relevance. Using computational modeling, we showed that this center-surround inhibition could be reproduced by either a sharpening of tuning curves of expected orientation or a shift of tuning curves of unexpected orientations. Intriguingly, these two computations were further supported by orientation-adjustment and orientation-discrimination experiments. Finally, the ablation studies in convolutional neural networks revealed that predictive coding feedback played a critical role in the center-surround inhibition in expectation. Altogether, our study reveals for the first time that expectation results in both enhancement and suppression, optimizing plausible interpretations during perception by enhancing expected and attenuating similar but irrelevant and potentially interfering representations.

Keywords: expectation; center-surround inhibition; tuning sharpening; tuning shift; deep predictive coding neural network

Introduction

Human behavior is surprisingly efficient and adaptive. Although everyday environment is brimming with noisy and ambiguous information, our cognitive system can quickly and adeptly deduce plausible interpretations of this information by combining it with prior expectations, ultimately facilitating to flexible behavioral arises[1,2]. However, the structured manner (the profile, in other words) regarding how expectation demarcates the anticipated target from various distractors, and underlying neural computations remain largely unclear. This issue is particularly important since such profile is thought to closely reflect neural circuitry[3,4], and therefore, offers us a unique opportunity to give insight into neural circuit level computations of expectation, thereby not only furthering our understanding of how it facilitates perception and behavior to adapt to changing environment, but also addressing a long-standing debate about its underlying neural mechanisms[5,6].

One of the central questions to this debate is about the processing of unexpected stimuli that are sufficiently novel or surprising. The sharpening models (also referred to as Bayesian theories) propose that expectations preferentially suppress neurons tuned toward the unexpected stimuli, resulting in a sharper and more selective population responses[5,7,8]. This sharpening account of expectation is similar to the notion of neuronal resonance[9] and has been supported by neurophysiological[10-14], electro-/magneto-encephalogram[15-19], and functional magnetic resonance imaging (fMRI)[7,20-22] studies. Conversely, the cancelation models (also referred to as

dampening theories) propose a dampening of neural responses reduces redundancy in the sensory system, through suppressing neurons tuned toward the expected stimulus. By canceling the expected information, the brain could highlight the processing and cognitive resources of unexpected information[6,23]. This theory has also drawn wide support from neurophysiological[13,14,24] and brain imaging[25-28] studies. Intriguingly, although these two models explaining how expectations render perception either veridical or informative are seemingly conflicting, both could be incorporated in the framework of predictive coding models[7,9,29-33], which posits that the brain contains distinct neurons/units representing the best guess about the outside world (prediction units) and the discrepancy between these guesses and incoming sensory evidence (prediction error units). Several studies have proposed that the sharpening and cancelation accounts may occur in prediction and error neurons, respectively[6,23,30]. Within this framework, anticipating about what is possible or probable in the forthcoming sensory environment can be cast as a process of hierarchical Bayesian inference, in which the prediction units are more strongly weighted towards the expected rather than unexpected stimuli, while at the same time the prediction error units are selectively biased to surprising inputs. Increased gain on these surprising inputs would lead to high fidelity representations of unexpected stimuli across prediction units. Although, so far there has been no direct evidence for the existence of these two neuron types, and it is unclear how these two mechanisms are reconciled from different neural populations, the predictive coding framework may provide the underlying computational basis for various potential profiles of expectation.

Here, given these two mechanisms making opposite predictions about how expectation changes the neural responses of unexpected stimuli, thereby displaying different profiles of expectation, we speculated that if expectation operates by the sharpening model with suppressing unexpected information, we should observe an inhibitory zone surrounding the focus of expectation, and its profile then should display as a center-surround inhibition (Fig. 1c, left). If, however, expectation operates as suggested by the cancelation model with highlighting unexpected information, the inhibitory zone surrounding the focus of expectation should be eliminated, and the profile should instead display a monotonic gradient (Fig. 1c, right). To adjudicate between these theoretical possibilities, we manipulated the distance between the expected and unexpected stimuli in feature space to measure the profile of expectation in two psychophysical experiments (orientation was task-relevant or task-irrelevant on the orientation and spatial frequency discrimination tasks, respectively, Fig. 1b), both of which supported the sharpening account by showing a classical center-surround inhibition profile in orientation space, with enhanced neural responses to the expected orientation and suppressed neural responses to orientations similar to the expected orientation relative to orientations more distinct from it (Fig. 2). Second, using computational modelling, we showed that the behaviorally observed center-surround inhibition in expectation could be reproduced by either a sharpening of tuning curves of expected orientation (Tuning sharpening account) or a shift of tuning curves of unexpected orientations (Tuning shift account) (Fig. 3a). Third, these neural computations, consisting of both the tuning sharpening and tuning shift

accounts, were further confirmed by orientation-adjustment (Fig. 6) and orientation-discrimination (Fig. 7) experiments. Finally, we found that a deep predictive coding neural network (DPCNN) exhibited a similar center-surround inhibition by expectation profile, both when it was trained to perform an orientation or a spatial frequency task. Most importantly, when we ablated predictive feedback, these center-surround inhibitions were eliminated in the DPCNN (Fig. 8), strongly supporting the framework of predictive coding models in expectation[7,9,29-33]. Altogether, our study reveals for the first time that expectation generates an orientation-specific enhancement and suppression profile that optimizes plausible interpretations during visual perception by boosting expected and attenuating interfering sensory representations.

Results

Profile experiment

The profile experiment consisted of a baseline and main experiment, with the baseline experiment always preceding the main experiment. The two experiments were the same, except for the probability relationship between the auditory cue and the orientation (20°, 30°, 40°, 50°, 60°, and 70°) of the first grating. For the baseline experiment, the auditory cue, comprising either a low- (240 Hz) or high- (540 Hz) frequency tone, predicted the orientation of the first grating with equal validity (16.67%, Fig. 1a, left). In the main experiment, this low- or high-frequency tone auditory cue predicted the orientation (20° or 70°) of the first grating with 75% validity. In the remaining 25% of trials, this orientation was chosen randomly and equally from four non-predicted orientations (30°, 40°, 50°, and 60°, Fig. 1a, right). Thus, for each participant, there were two types of expected conditions: Expect 20° and Expect 70°, and for both conditions, there were five possible distances in orientation space between the expected and test gratings, ranging from Δ0° through Δ40° with a step size of 10°. Note that, the matches between the tone (low- or high-frequency) of auditory cue and the expected orientation (20° or 70°) of the first grating were flipped across participants, and the order was also counterbalanced across participants. Moreover, for each participant, although the tone of auditory cue could not predict 20° or 70° orientation in the baseline experiment, the trials in the baseline experiment with the same tone that was matched with 20° or 70° orientation

in the main experiment, were defined as Baseline 20° (i.e., the baseline of Expect 20°) and Baseline 70° (i.e., the baseline of Expect 70°) conditions, respectively.

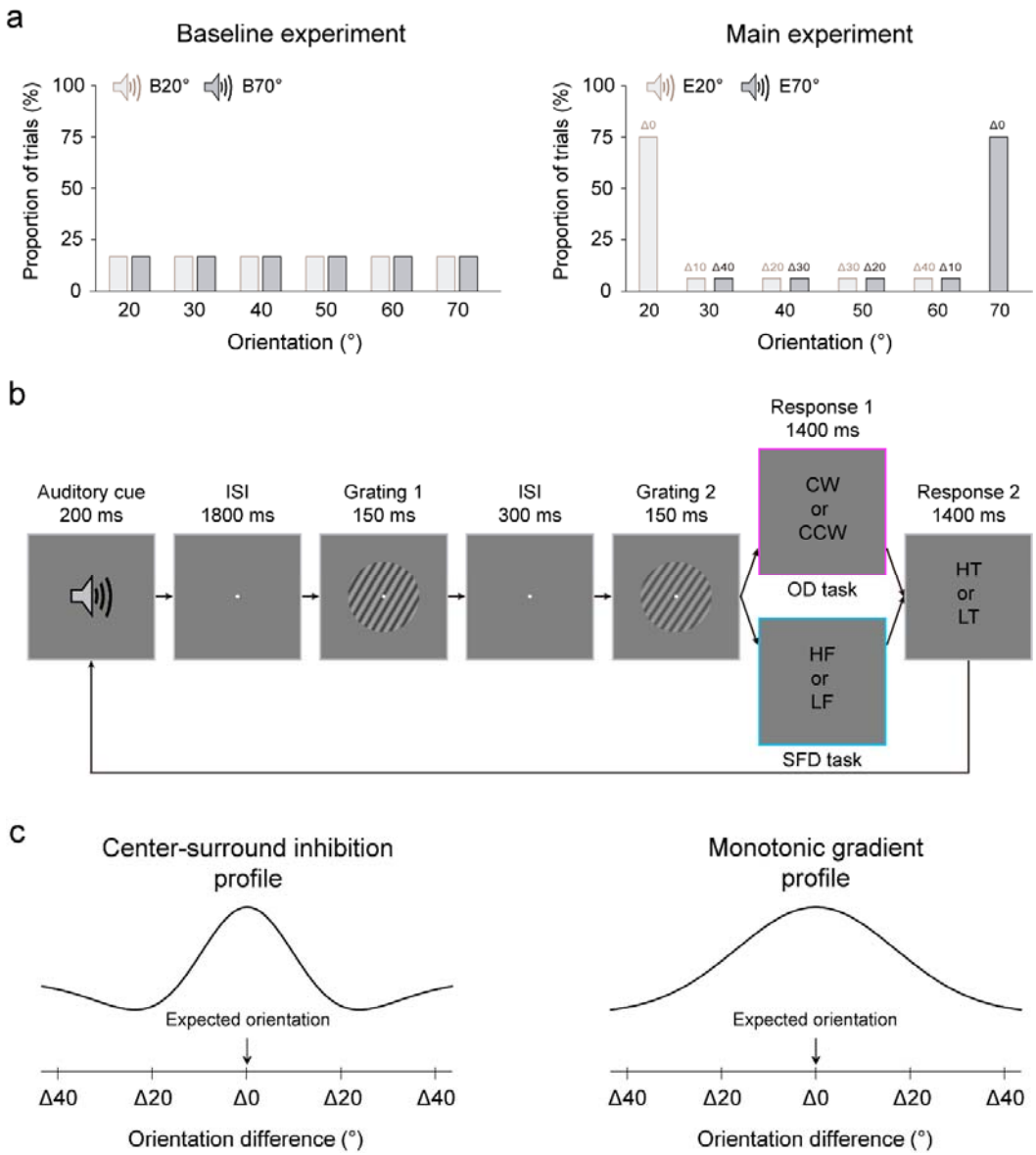

Fig. 1 | Stimuli and protocols of the profile experiment. a, Left: the auditory cue, comprising either a low- or high- frequency tone, predicted the orientation of the first grating with equal validity in the baseline experiment. B20°: Baseline 20°; B70°: Baseline 70°. Right: in the main experiment, the low- or high-frequency tone predicted 20° or 70° (expected) orientation of the first grating with 75% validity. In

the remaining 25% of trials, this orientation was chosen randomly and equally from four non-predicted orientations (30°, 40°, 50°, and 60°). There were two types of expected conditions: Expect 20° (E20°) and Expect 70° (E70°), and for both conditions, there were 5 possible distances in orientation space between the expected and test gratings, ranging from Δ0° through Δ40° with a step size of 10°. **b,** In both baseline and main experiments, each trial began with an auditory cue, followed by an 1800 ms fixation interval. Then, two consecutive gratings were each presented for 150-ms and separated by a 300-ms blank interval. Participants were first asked to make a 2AFC judgment of either the orientation (clockwise or anticlockwise) or the spatial frequency (lower or higher) of the second grating relative to the first on orientation discrimination (OD, purple) and spatial frequency discrimination (SFD, blue) tasks, respectively. Then, participants were asked to make another 2AFC judgment on the tone of auditory cue, either low or high. CW: clockwise; CCW: counterclockwise; HF: higher frequency; LF: lower frequency; HT: high tone; LT: low tone. **c,** Left: expectation operates by the sharpening model with suppressing unexpected information, under this configuration, the profile of expectation could display as a center-surround inhibition, with an inhibitory zone surrounding the focus of expectation. Right: expectation operates by the cancelation model with highlighting unexpected information, under this configuration, the profile of expectation could display as a monotonic gradient, without the inhibitory zone.

Both the baseline and main experiments consisted of two tasks: the orientation discrimination (OD) task and spatial frequency discrimination (SFD) task; with the two tasks occurring on different days; the order of the two tasks was counterbalanced across participants. Differently, the baseline experiment consisted of 4 blocks (two for OD task and the other two for SFD task), and each block had 2 QUEST staircases[34] for each of six orientations (20°, 30°, 40°, 50°, 60°, and 70°). The main experiment

consisted of 2 blocks (one for OD task and the other one for SFD task), and each block had 24 QUEST staircases for the expected orientations (20° and 70°) and 2 QUEST staircases for each of unexpected orientations (30°, 40°, 50°, and 60°). Each QUEST staircase comprised 40 trials and each trial began with an auditory cue, followed by a fixation interval. Then, two gratings were presented sequentially and participants were asked to make a two-alternative forced-choice (2AFC) judgment of either the orientation (clockwise or anticlockwise, where orientation was task-relevant) or the spatial frequency (lower or higher, where orientation was task-irrelevant) of the second grating relative to the first, on the OD and SFD tasks, respectively (Fig. 1b). The second grating differed trial by trial from the first in either orientation ($\Delta\theta$°) or spatial frequency ($\Delta\lambda$ cycles/°) on the OD and SFD tasks, respectively. The QUEST staircase was used to control the varied $\Delta\theta$° or $\Delta\lambda$ cycles/° adaptively for estimating participants' discrimination thresholds (75% correct). At the end of each trial, participants needed to report the tone (either low or high) of the auditory cue. For either OD or SFD tasks, there was no significant difference in accuracy of this reporting across different conditions in either baseline or main experiments (Supplementary Fig. 1).

In the baseline experiment, participants' mean discrimination thresholds in Baseline 20° and Baseline 70° conditions were submitted to a one-way repeated-measures ANOVA with orientation (20°, 30°, 40°, 50°, 60°, and 70°) as a within-participants factor. Results showed that the main effect of orientation was not

significant in either OD (Baseline 20°: F(5,115) = 0.955, p = 0.431, η_p^2 = 0.040; Baseline 70°: F(5,115) = 1.314, p = 0.274, η_p^2 = 0.054) or SFD (Baseline 20°: F(5,115) = 1.163, p = 0.331, η_p^2 = 0.048; Baseline 70°: F(5,115) = 1.593, p = 0.184, η_p^2 = 0.065) tasks (Fig. 2a), indicating that there was no significant difference in participant performance among six orientations. In other words, the tone of auditory cue in the baseline experiment was uninformative about the orientation of gratings. On both OD and SFD tasks and both two expected conditions (Expect 20° and Expect 70°), for each distance (Δ0°-Δ40°), we computed a discrimination sensitivity (*DS*) to quantify how much the discrimination threshold (*DT*) changed between baseline ($DT_{baseline}$) and main (DT_{main}) experiments: $DS = DT_{baseline} - DT_{main}$. Because the *DS* from Expect 20° and Expect 70° showed as a similar pattern, they were pooled together for further analysis (unless otherwise stated, we present average data from two expected conditions). The averaged *DSs* were submitted to a one-way repeated-measures ANOVA with the distance (Δ0°-Δ40°) as a within-participants factor. Results showed that, the main effect of distance was significant in both OD (F(4,92) = 3.739, p = 0.010, η_p^2 = 0.140, Fig. 2c, top) and SFD (F(4,92) = 2.822, p = 0.042, η_p^2 = 0.109, Fig. 2c, bottom) tasks. To directly address the potential inhibitory zone surrounding the focus of expectation, we compared the *DS*s between Δ20° and Δ0°, and between Δ20° and Δ40° on each task. *Post hoc* paired *t* tests revealed that, for both tasks, the *DSs* of Δ20° were significantly lower than those of both Δ0° (OD task: t(23) = -4.263, p < 0.001, Cohen's *d* = 0.870; SFD task: t(23) = -4.679, p < 0.001, Cohen's *d* = 0.955) and Δ40° (OD task: t(23) = -2.214, p = 0.037, Cohen's *d* = 0.452; SFD task: t(23) =

-2.694, p = 0.013, Cohen's *d* = 0.550), indicating a classical center-surround inhibition in expectation with the enhanced discriminability to the expected orientation (Δ0°) and decreased discriminability to orientations (Δ20°) similar to the expected orientation relative to orientations (Δ40°) more distinct from it. Intriguingly, this center-surround inhibition profile of expectation was independent of attentional modulations by task-relevance of the orientation.

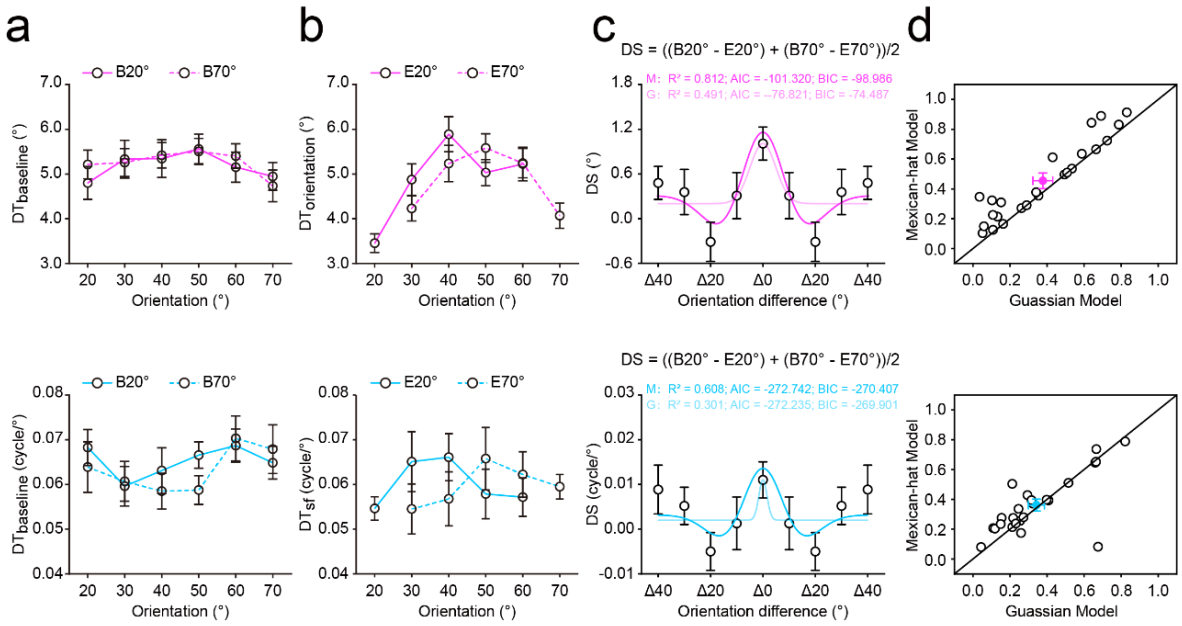

Fig. 2 | Results of the profile experiment. The discrimination thresholds of OD (top) and SFD (bottom) tasks during baseline (**a**) and main (**b**) experiments. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70°. **c,** The averaged discrimination sensitivity (*DS*) of each distance on OD (top) and SFD (bottom) tasks, and the best fitting Gaussian and Mexican-hat functions to these DSs across distances. G, Gaussian model; M, Mexican-hat model. **d,** R^2 of the best fitting Gaussian and Mexican-hat functions for individual participants in OD (top) and SFD (bottom) tasks. For both tasks, most dots located in the upper-left zone, demonstrating that the Mexican-hat model was favored over the Gaussian model. Open symbols represent

the data from each participant and filled colored dots represented the mean across participants. Error bars indicate 1 SEM calculated across participants.

Subsequently, to further assess the shape of this expectation pattern, we fitted a monotonic model and a non-monotonic model to the average *DS*s across distances (Δ0°-Δ40°) on both OD and SFD tasks. The monotonic and nonmonotonic models were implemented as the Gaussian and Mexican-hat functions, respectively[35,36]. To compare these two models to our data, we first computed the Akaike information criterion (AIC)[37] and Bayesian information criterion (BIC)[38] with the assumption of a normal error distribution. Then, we calculated the Likelihood ratio (LR) and Bayes factor (BF) of the Mexican-hat model over the Gaussian model based on AIC[39] and BIC[40] approximation, respectively. Results showed that, in both tasks, the LR/BFs were larger than 1 (OD task: LR/BF = 2.088×10^5 ; SFD task: LR/BF = 1.288) and therefore strongly favored the Mexican-hat model over the Gaussian model (Fig. 2c). Notably, we also conducted similar model comparisons for each participant's data and found that the Mexican-hat model was favored over the Gaussian model in 23 and 17 of 24 participants, for OD and SFD tasks, respectively (Fig. 2d). Together, these results constituted strong evidence for the center-surround inhibition profile of expectation, and further indicated its independence of attentional modulations by task-relevance of the orientation.

Computational models of the center-surround inhibition in expectation

Our results demonstrated the classical center-surround inhibition profile in expectation, yet it remains unclear what type of neural computations could account for this profile. We proposed that this profile could be explained by either Tuning sharpening (Fig. 3a, left) or Tuning shift (Fig. 3a, right) models. The Tuning sharpening model postulates that expectation sharpens the tuning of individual neurons (thick curves) of the expected orientation, which results in a center-surround population response profile (black curve) centered at the expected orientation. The Tuning shift model postulates that expectation attracts the tuning of individual neurons (thick curves) from unexpected orientations towards the expected orientation, which also results in a center-surround population response profile. Note that, in our study, the shift towards to 20° was (arbitrarily) considered to be the negative value, whereas the shift towards to 70° was thus the positive value, and unless otherwise stated, we present the average shift, i.e., *mean shift = (shift towards to 70°- shift towards to 20°)/2*, across conditions hereafter. For both OD and SFD tasks, to compare these two models, we fitted both the Tuning sharpening and Tuning shift models (sum of idealized channel tuning functions) to the population response profiles (the smooth negative values of discrimination thresholds) during baseline and main experiments, and measured their root mean squared deviation (RMSD) metric[41]. RMSD takes the number of model parameters into account, and a smaller RMSD indicates better model fitness. Our results showed that both models exhibited robust fits to our data (Fig. 3b and Fig. 3e), as indicated by high R^2 values and comparably low RMSDs in both OD (Fig. 4a) and SFD (Fig. 4b) tasks. Similarly, we computed a

discrimination sensitivity (*DS*) to quantify how much the fitted discrimination threshold (*FDT*) changed between baseline ($FDT_{baseline}$) and main (FDT_{main}) experiments: $DS = FDT_{baseline} - FDT_{main}$. For both models, a similar center-surround inhibition profile of the *DS* was found on both OD (Fig. 3c) and SFD (Fig. 3f) tasks. Further model comparisons for each participant's data confirmed that the Mexican-hat model was favored over the Gaussian model on both OD (22 and 19 of 24 participants for Tuning sharpening and Tuning shift models, respectively, Fig. 3d) and SFD (14 and 17 of 24 participants for Tuning sharpening and Tuning shift models, respectively, Fig. 3g) tasks. These results imply that the center-surround inhibition in expectation could be reproduced by either Tuning sharpening or Tuning shift models.

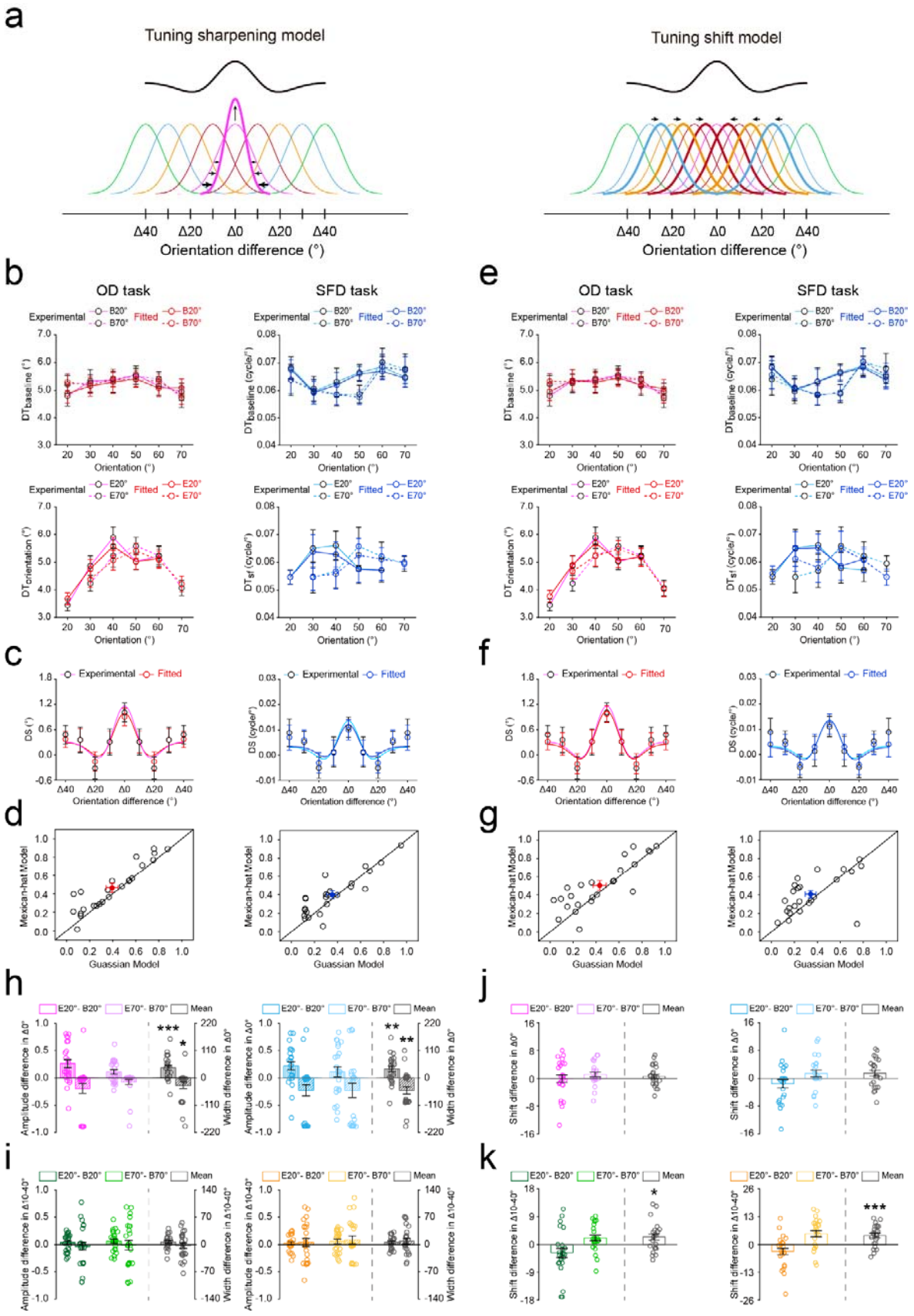
a
Tuning sharpening model
Tuning shift model
Δ40 Δ20 Δ0 Δ20 Δ40
Orientation difference (°)
b
e
OD task
SFD task
Experimental
Fitted
B20°
B70°
E20°
E70°
Orientation (°)
c
f
Orientation difference (°)
d
g
Mexican-hat Model
Guassian Model
h
j
E20°- B20°
E70°- B70°
Mean
Amplitude difference in Δ0°
Width difference in Δ0°
Shift difference in Δ0°
i
k
Amplitude difference in Δ10-40°
Width difference in Δ10-40°
Shift difference in Δ10-40°

Fig. 3 | Results of computational modeling. a, Illustration of the Tuning sharpening model (left) and the Tuning shift model (right). The Tuning sharpening model postulates that expectation sharpens the tuning of individual neurons (thick curves) towards the expected orientation, which results in a center-surround population response profile (black curve) centered at the expected orientation. The Tuning shift model postulates that expectation attracts the tuning of individual neurons (thick curves) from unexpected orientation towards the expected orientation, which also results in a center-surround population response profile. **b,** The fitted discrimination thresholds on OD (left) and SFD (right) tasks in the baseline (top) and main (bottom) experiments. **c,** The averaged DSs using Tuning sharpening model on OD (left) and SFD (right) tasks. **d,** R^2 of the best fitting Gaussian and Mexican-hat functions for individual participants based on the fitted DSs using Tuning sharpening model on OD (left) and SFD (right) tasks. Open symbols represent the data from each participant and filled colored dots represented the mean across participants. **e-g,** The results from Tuning shift model, see caption for (**b-d**) for a description of each type of graph. The amplitude *A* (vertical stripes) and width σ (diagonal stripes) differences between the baseline and main experiments using Tuning sharpening model in Δ0° (**h**) and Δ10°-Δ40° (**i**) conditions, on OD (left) and SFD (right) tasks. The location *x0* differences between the baseline and main experiments using Tuning shift model in Δ0° (**j**) and Δ10°-Δ40° (**k**) conditions, on OD (left) and SFD (right) tasks. Open symbols represent the data from each participant and error bars indicate 1 SEM calculated across participants. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70° (*$p < 0.05$; **$p < 0.005$; ***$p < 0.001$).

For each model and each task, to directly compare the tuning curve changes of both the expected (Δ0°) and unexpected orientations (Δ10°-Δ40°) with and without expectation, we calculated the parameters changes of tuning functions (amplitude *A*, location *x0*, and width σ) for hypothesized channels between baseline and main

experiments (Supplementary Fig. 2). For the Tuning sharpening model, the tuning width of each channel's tuning function is parameterized by ⍰, while all tuning functions are evenly distributed with 10° spacing on the *x*-axis and the areas under the curves (response energy) are identical. Conversely, for the Tuning shift model, the location of each channel's tuning function is parameterized by *x0*, while they all share the same tuning amplitude and width. For both models, parameters were varied to obtain the minimal sum of squared errors between the population response profile and the model prediction, which is the sum of all channels' tuning responses. For the expected orientation (Δ0°) of Tuning sharpening model, results showed that the amplitude change was significantly higher than zero on both OD ($t(23) = 4.198$, $p < 0.001$, Cohen's $d = 0.857$) and SFD ($t(23) = 3.247$, $p = 0.004$, Cohen's $d = 0.663$) tasks (Fig. 3h, vertical stripes); the width change was significantly lower than zero on both OD ($t(23) = -2.235$, $p = 0.035$, Cohen's $d = 0.456$) and SFD ($t(23) = -3.313$, $p = 0.003$, Cohen's $d = 0.676$) tasks (Fig. 3h, diagonal stripes). For unexpected orientations (Δ10°-Δ40°), however, the amplitude and width changes were not significant with zero on either OD (amplitude change: $t(23) = 1.948$, $p = 0.064$, Cohen's $d = 0.397$; width change: $t(23) = -0.412$, $p = 0.684$, Cohen's $d = 0.084$) or SFD (amplitude change: $t(23) = 1.708$, $p = 0.101$, Cohen's $d = 0.349$; width change: $t(23) = 1.273$, $p = 0.216$, Cohen's $d = 0.260$) tasks (Fig. 3i). For the Tuning shift model, results showed that the location shift was significantly different than zero for unexpected orientations (Δ10°-Δ40°, OD task: $t(23) = 2.547$, $p = 0.018$, Cohen's $d = 0.520$; SFD task: $t(23) = 4.099$, $p < 0.001$, Cohen's $d = 0.837$ (Fig. 3k), but not for the

expected orientation (Δ0°, OD task: t(23) =0.993, p = 0.331, Cohen's d = 0.203; SFD task: t(23) = 1.750, p = 0.093, Cohen's d = 0.357 (Fig. 3j). These results further confirm the Tuning sharpening and Tuning shift computations for the center-surround inhibition in expectation.

In addition, across participants, we further used the non-parametric Wilcoxon signed-rank test to compare both the R^2 and RMSD between two models for Baseline 20°, Baseline 70°, Expect 20°, and Expect 70° conditions during each task. Results showed that there was no significant difference between two models in Baseline 20° (OD task: R^2: z = 1.372, p = 0.170, effect size: r = 0.280; RMSD: z = 1.200, p = 0.230, effect size: r = 0.245; SFD task: R^2: z = 0.857, p = 0.391, effect size: r = 0.175; RMSD: z = 0.829, p = 0.407, effect size: r = 0.169), Baseline 70° (OD task: R^2: z = 0.371, p = 0.710, effect size: r = 0.076; RMSD: z = 0.029, p = 0.977, effect size: r = 0.006; SFD task: R^2: z = 1.657, p = 0.097, effect size: r = 0.338; RMSD: z = 0.686, p = 0.493, effect size: r = 0.140), Expect 20° (OD task: R^2: z = 0.686, p = 0.493, effect size: r = 0.140; RMSD: z = 1.600, p = 0.110, effect size: r = 0.327; SFD task: R^2: z = 1.257, p = 0.209, effect size: r = 0.257; RMSD: z = 1.600, p = 0.110, effect size: r = 0.327), or Expect 70° (OD task: R^2: z = 1.486, p = 0.137, effect size: r = 0.303; RMSD: z = 1.686, p = 0.092, effect size: r = 0.344; SFD task: R^2: z = 0.514, p = 0.607, effect size: r = 0.105; RMSD: z = 0.143, p = 0.886, effect size: r = 0.029) conditions (Fig. 4). These results further imply that Tuning sharpening and Tuning shift models may equally contribute to center-surround inhibition in expectation.

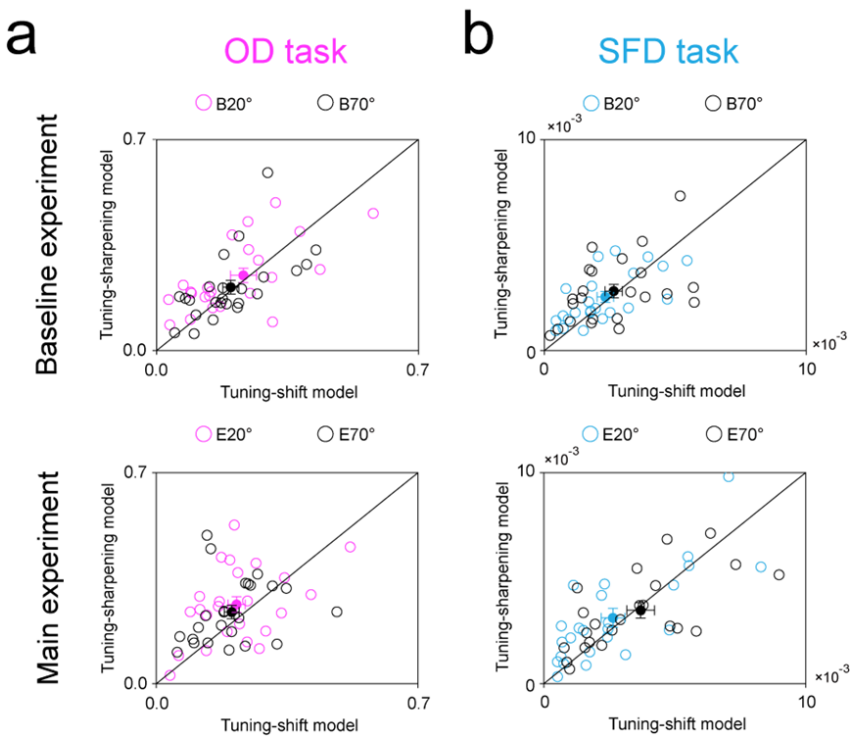

Fig. 4 | RMSDs of Tuning sharpening and Tuning shift models. RMSDs of Tuning sharpening and Tuning shift models during the baseline (top) and main (bottom) experiments, on OD (**a**) and SFD (**b**) tasks. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70°. Open symbols represent the data from each participant and filled colored dots represented the mean across participants. Error bars indicate 1 SEM calculated across participants.

Orientation adjustment experiment

To further explore the co-existence of both Tuning sharpening and Tuning shift computations in center-surround inhibition profile of expectation, participants were asked to perform a classic orientation adjustment experiment. The protocol of orientation adjustment experiment was similar with that of the profile experiment, except for two aspects. First, there were 4 possible (20°, 40°, 50°, and 70°)

orientations for the first grating: 20°/70° (Δ0° deviated from the expected orientation) and 40°/50° (Δ20°/Δ30° deviated from the expected orientation). Second, in both baseline and main experiments, the second grating was set as a random orientation within the range of 0° to 90°, and participants were required to rotate the orientation of the second grating to match the first (Fig. 5a). Similar to the profile experiment, no significant difference was found in tone report accuracies across distances (Supplementary Fig. 3). For both expected (Δ0°) and unexpected (Δ20°/Δ30°) orientations, we calculated the adjusted orientation difference between the baseline and main experiments. Results showed the adjusted difference was significantly higher than zero for unexpected orientations (0.735 ± 0.308: $t(19) = 2.387$, $p = 0.028$, Cohen's $d = 0.534$, Fig. 6a, right), but not for the expected orientation (0.143 ± 0.523: $t(19) = 0.274$, $p = 0.787$, Cohen's $d = 0.061$, Fig. 6a, left), suggesting a significant bias in the unexpected orientation representation towards the expected orientation.

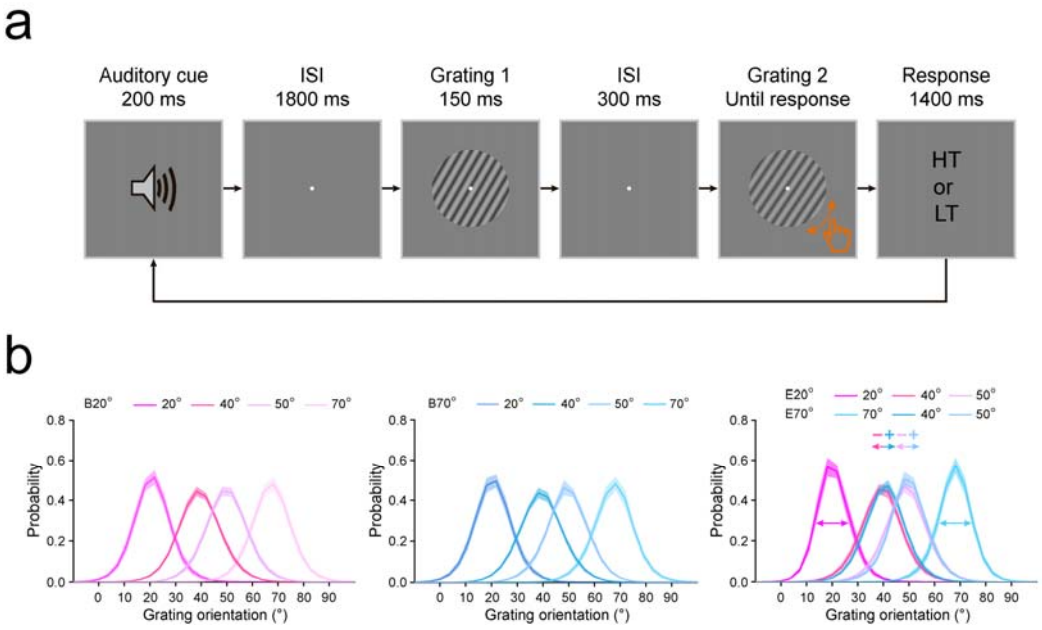

Fig. 5 | Protocol and error distributions of the orientation adjustment experiment. **a,** The protocol of orientation adjustment experiment was similar with that of the profile experiment, except for two aspects. First, there were 4 possible (20°, 40°, 50°, and 70°) orientations for the first grating: 20°/70° (Δ0° deviated from the expected orientation) and 40°/50° (Δ20°/Δ30° deviated from the expected orientation). Second, in both baseline and main experiments, the second grating was set as a random orientation within the range of 0° to 90°, and participants were required to rotate the orientation of the second grating to match the first. HT: high tone; LT: low tone. **b,** Three-component mixture model to the adjusted errors from baseline (left) and main (middle) experiments. In current study, the shift towards to 20° was (arbitrarily) considered to be the negative value ('-'), whereas the shift towards to 70° was thus the positive value ('+'). The mean shift was calculated as: *mean shift = (shift towards to 70°- shift towards to 20°)/2*. The shaded error bars indicate 1 SEM calculated across participants. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70°.

Subsequently, we employed a three-component mixture model[42,43] to the adjusted errors from both baseline and main experiments (Fig. 5b). This allowed us to estimate representation precision, including the mean shift (*mu*) and standard deviation (*s.d.*) of the von Mises distribution (positive values indicating rightward shift and higher values indicating lower precision, respectively), along with assessing the probability of stimulus guessing (*g*). Using the difference between the main and baseline experiments (Supplementary Fig. 4), we also found that the orientation representation significantly shifted for unexpected orientations (0.752 ± 0.303: t(19) = 2.481, p = 0.023, Cohen's d = 0.555, Fig. 6b, right), but not for the expected orientation (0.214 ± 0.493: t(19) = 0.434, p = 0.669, Cohen's d = 0.097, Fig. 6b, left). Conversely,

participants exhibited higher orientation representation precision than baseline in the expected orientation (-0.973 ± 0.271: $t(19) = -3.597$, $p = 0.002$, Cohen's $d = 0.804$, Fig. 6c, left), but not in unexpected orientations (-0.390 ± 0.212: $t(19) = -1.837$, $p = 0.082$, Cohen's $d = 0.411$, Fig. 6c, right). Finally, we found no significant difference with zero in the rate of guessing in either expected (-0.0047 ± 0.0554: $t(19) = -0.849$, $p = 0.406$, Cohen's $d = 0.190$, Fig. 6d, left) or unexpected (-0.0023 ± 0.0148: $t(19) = -1.551$, $p = 0.138$, Cohen's $d = 0.347$, Fig. 6d, right) orientations. These results provide converging evidence supporting our hypothesis that both Tuning sharpening and Tuning shift contribute to the center-surround inhibition profile of expectation.

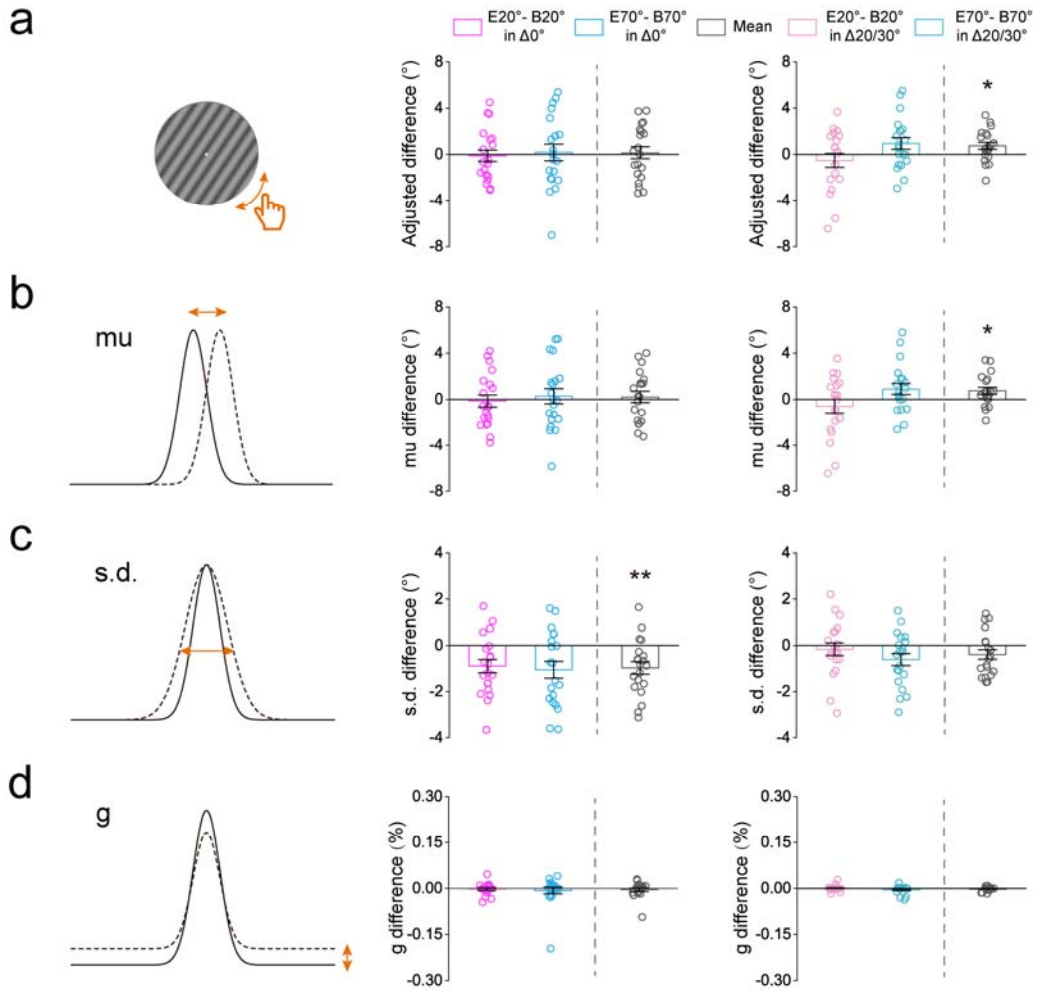

Fig. 6 | Results of the orientation adjustment experiment. a, The adjusted orientation difference between the baseline and main experiments in both expected (20°/70°, i.e. Δ0°, middle) and unexpected (40°/50°, i.e. Δ20°/Δ30°, right) conditions. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70°. **b-d,** The parameter estimates difference between the baseline and main experiments in both expected (Δ0°, middle) and unexpected (Δ20°/Δ30°, right) orientations. The parameter estimates were obtained by fitting a three-component mixture model to adjusted errors in different conditions. **b,** *mu* reflects the response distribution shift away from the presented grating orientation. **c,** *s.d.* reflects precision of responses (with higher values indicating worse precision). **d,** *g* estimates the probability that the participant produced a random response (i.e., the guess). Open symbols represent the

data from each participant and error bars indicate 1 SEM calculated across participants (*p < 0.05; **p < 0.005; ***p < 0.001).

Orientation discrimination experiment

Note that behavioral benefits in our orientation adjustment task could be due to improvements in either perceptual or decisional processes, as the expectation cue held information about both the most likely stimulus and the most likely correct response[5,7,15,16]. To remove this link between stimulus and response expectations and thereby avoid potential response biases induced by the cue we designed an additional orientation discrimination experiment. The protocol of this orientation discrimination experiment was very similar with that of the orientation adjustment experiment, except for two aspects (Fig. 7a). First, there were 3 possible (20°, 45°, and 70°) orientations for the first grating: 20°/70° (Δ0° deviated from the expected orientation) and 45° (Δ25° deviated from the expected orientation). Second, in both baseline and main experiments, the second grating were 1°, 3°, 5°, 7°, and 9° deviated from the first grating, either clockwise or counterclockwise. Participants were asked to make a 2AFC judgment of the orientation of the second grating relative to the first, either clockwise or anticlockwise. Fig. 7b and Fig. 7c shows the psychometric functions for each condition. We plotted the percentage of trials in which participants indicated the orientation of second grating that was anticlockwise or clockwise to the first for 20° (Baseline 20° and Expect 20°) and 70° (Baseline 70° and Expect 70°) conditions, respectively, as a function of the actual orientation difference between the two

gratings. For each participant and each condition, the psychometric values at 10 orientation differences were fitted to a cumulative Gaussian function and we interpolated the data to find the slope (orientation uncertainty) and PSE (point of subjective equality, which is the shift here) as an index for Tuning sharpening and Tuning shift models, respectively.

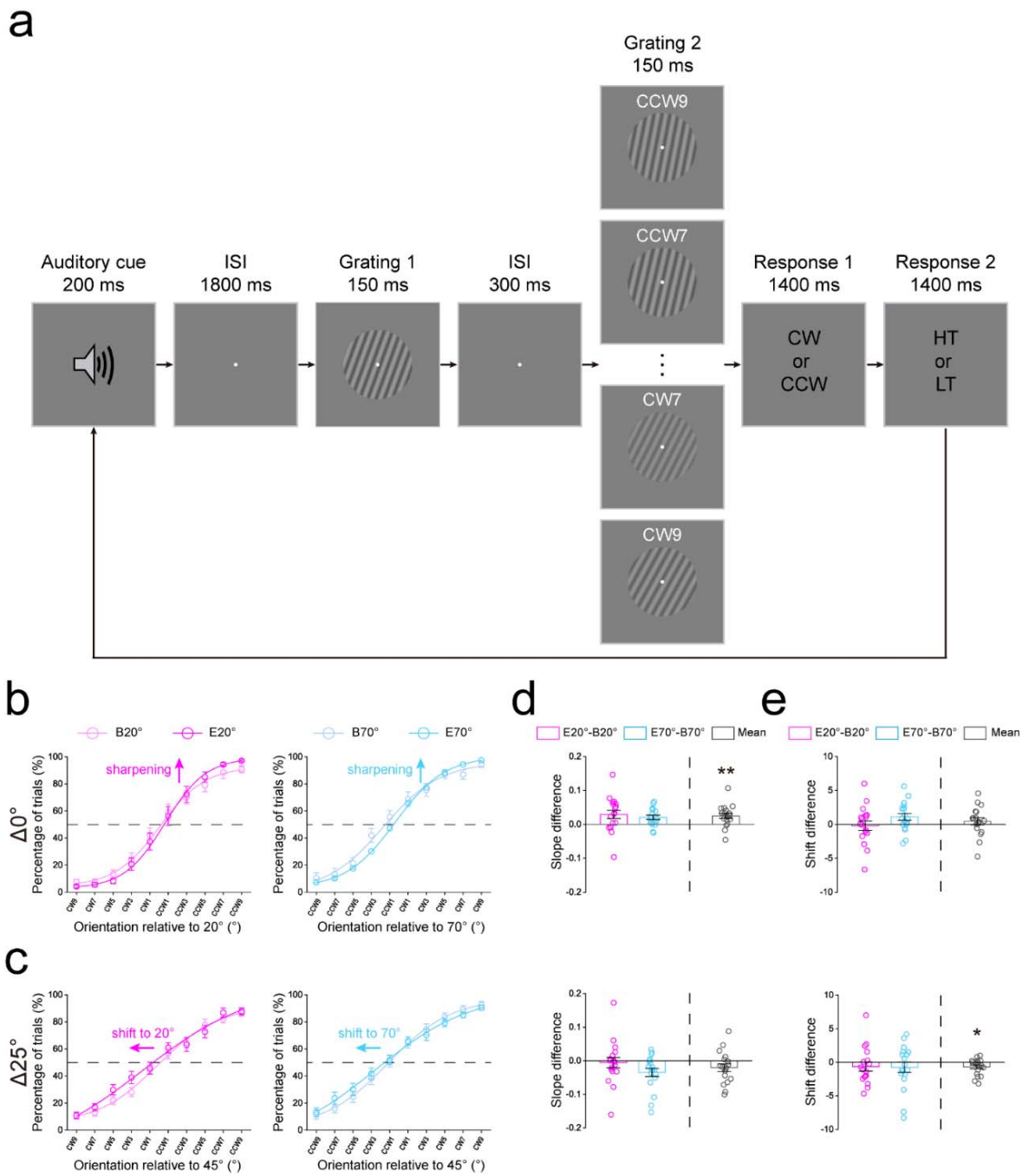

Fig. 7 | Protocol and results of the orientation discrimination experiment. a, The protocol of orientation discrimination experiment was similar with that of the orientation adjustment experiment, except for two aspects. First, there were 3 possible (20°, 45°, and 70°) orientations for the first grating: 20°/70° (Δ0° deviated from the expected orientation) and 45° (Δ25° deviated from the expected orientation). Second,

in both baseline and main experiments, the second grating were 1°, 3°, 5°, 7°, and 9° deviated from the first grating, either clockwise (CW) or counterclockwise (CCW). Participants were asked to make a 2AFC judgment of the orientation of the second grating relative to the first, either clockwise or anticlockwise. HT: high tone; LT: low tone. Psychometric functions showing orientation judgements in each condition for Δ0° (**b**) and Δ25° (**c**). Data points averaged across participants were fit using a cumulative normal function. The abscissa refers to ten orientation differences between the first and second gratings. The ordinate refers to the percentage of trials in which participants indicated the orientation of second grating that was anticlockwise or clockwise to the first for expected 20° (left) and 70° (right) conditions, respectively. The slope (an index for the Tuning sharpening model, **d**) and shift (an index for the Tuning shift model, **e**) differences between the baseline and main experiments for expected 20° and 70° conditions. Negative: shift to left; Positive: shift to right. Open symbols represent the data from each participant and error bars indicate 1 SEM calculated across participants. B20°: Baseline 20°; B70°: Baseline 70°; E20°: Expect 20°; E70°: Expect 70° (*$p < 0.05$; **$p < 0.005$).

Similar to the orientation adjustment experiment, no significant difference was found in tone report accuracies across distances (Supplementary Fig. 5). For both expected (Δ0°) and unexpected (Δ25°) orientations, we calculated the slope and shift difference between the baseline and main experiments. Results showed the slope difference was significantly higher than zero for expected orientations (0.0250 ± 0.0075: $t(17) = 3.324$, $p = 0.004$, Cohen's $d = 0.627$, Fig. 7d, top), but not for the unexpected orientation (-0.0204 ± 0.0113: $t(17) = -1.812$, $p = 0.088$, Cohen's $d = 0.402$, Fig. 7d, bottom). Conversely, the shift difference was significantly lower than zero for the unexpected orientation (-0.696 ± 0.287: $t(17) = -2.423$, $p = 0.027$,

Cohen's d = 0.507, Fig. 7e, bottom), but not for expected orientations (0.449 ± 0.509: t(17) = 0.881, p = 0.391, Cohen's d = 0.452, Fig. 7e, top). These results indicated that the expectation not only sharpened the tuning curves of neurons for the expected orientation but also attracted the tuning curves of neurons for unexpected orientations, further confirming Tuning sharpening and Tuning shift models, respectively, in the center-surround inhibition of expectation.

Artificial neural networks for the center-surround inhibition in expectation

Finally, we trained a deep predictive coding neural network (DPCNN), modified from Predify[44,45] to perform both the OD and SFD tasks. For both tasks, the DPCNN consisted of 6 feedforward encoding layers (*e1-e6*), 5 generative feedback decoding layers (*d1-d5*), and 3 fully connected (fc) layers (Fig. 8a). The reconstruction error (*E1-E5*) is computed and used for the proposed predictive coding updates[31], denoting by *P.C.* loops. Note that the updating is only applied to *e1-e5* and for the last layer *e6*, there is no feedback. Before the layer *e0*, we obtained the pixel difference between the target and reference images, which was then superimposed on the channels of the reference image. This superimposed image was set as the input of the network and remained constant over timesteps. Besides, we used a feedforward encoding layer (i.e., *e0*) to match the number of the channels between superimposed feature maps and the pre-trained DPCNN. During the training, the last layer of network was trained to capture the difference between the target and reference and finally obtain the

classification by softmax, to model decision making in our 2AFC paradigm (Fig. 1c), in which participants were asked to make a 2AFC judgment of the orientation (either clockwise or anticlockwise) or the spatial frequency (either lower or higher) of the second grating (target) relative to the first (reference) in OD and SFD tasks, respectively. For both tasks, the DPCNN was independently and randomly trained 12 times, and for each distance (Δ0°-Δ40°), the training effect was defined as the accuracy difference ($ACC_{difference}$) between the pre- ($ACC_{baseline}$) and post- ($ACC_{trained}$) training. Similar to our psychophysical results, on both tasks, the LR/BFs were much larger than 1 (OD task: LR/BF = 2.045×10^5, Fig. 8d, left; SFD task: LR/BF = 5.5929, Fig. 8h, left) and therefore strongly favored the Mexican-hat model over the Gaussian model. The model comparison based on fitting individual data advocated that the Mexican-hat model was favored over the Gaussian model in 10 and 11 of 12 training data on OD (Fig. 8e, left) and SFD (Fig. 8i, left) tasks, respectively. Besides, across individual data, a non-parametric Wilcoxon signed-rank test was conducted to compare the R^2 of two models, and results significantly advocated the Mexican-hat model over the Gaussian model on both OD (z = 2.197, p = 0.028, effect size: r = 0.634) and SFD (z = 2.981, p = 0.003, effect size: r = 0.861) tasks. These results suggest that our DPCNN can emerge the similar center-surround inhibition by expectation on both the orientation and spatial frequency trainings.

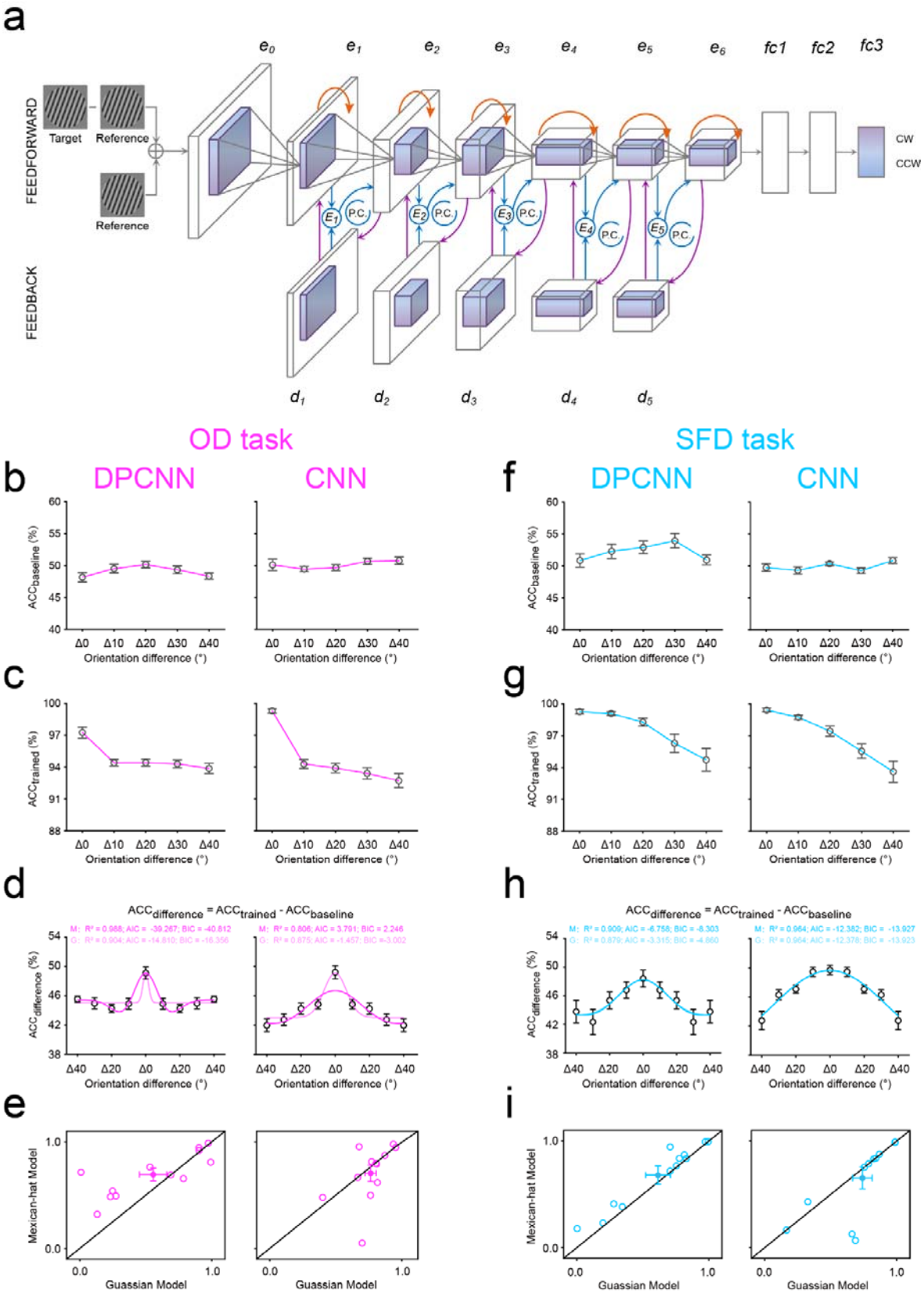
a
FEEDFORWARD
FEEDBACK
Target
Reference
Reference
e0
e1
e2
e3
e4
e5
e6
fc1
fc2
fc3
CW
CCW
E1
E2
E3
E4
E5
P.C.
d1
d2
d3
d4
d5
OD task
SFD task
b
DPCNN
CNN
f
DPCNN
CNN
ACCbaseline (%)
Orientation difference (°)
c
g
ACCtrained (%)
d
h
ACCdifference = ACCtrained - ACCbaseline
ACCdifference (%)
e
i
Mexican-hat Model
Guassian Model

Fig. 8 | Results of artificial neural networks. a, Model structure and stimulus examples for deep predictive coding neural network (DPCNN) and standard feedforward CNN, on both OD (purple) and SFD (blue) tasks. DPCNN consisted of 6 feedforward encoding layers (*e1-e6*), 5 generative feedback decoding layers (*d1-d5*), and 3 fully connected (fc) layers. The reconstruction error (*E1-E5*) is computed and used for the proposed predictive coding updates, denoting by *P.C.* loops. The CNN is the same with DPCNN but removing feedback predictive coding iterations. The accuracy of each distance during the pre- (**b**) and post- (**c**) training for DPCNN (left) and CNN (right), on the OD task. **d,** The training effect (i.e., the ACC difference between pre- and post-training) of each distance in DPCNN (left) and CNN (right), and the best fitting Mexican-hat and Gaussian functions to these training effects across distances, on the OD task. M, Mexican-hat model; G, Gaussian model. **e,** R^2 of the best fitting Mexican-hat and Gaussian functions from individual data in DPCNN (left) and CNN (right) on the OD task. Open symbols represent individual data and filled colored dots represented the mean across data. Error bars indicate 1 SEM calculated across data. **f-i,** The results from the SFD task, see caption for (**b-e**) for a description of each type of graph.

Additionally, to further determine the contribution of predictive (reconstructive) feedback to center-surround inhibition in expectation, we performed ablation studies, in which we trained the same network but removing feedback predictive coding iterations (a standard feedforward CNN, i.e., a modified network of AlexNet[46]). As expected, on both tasks, removing feedback leads to the disappearance of center-surround inhibition in expectation. Across individual data, there was no significant difference in the R^2 between Mexican-hat and Gaussian models on either OD (non-parametric Wilcoxon signed-rank test: z = -0.314, p = 0.754, effect size: r = 0.091, Fig. 8e, right) or SFD (non-parametric Wilcoxon signed-rank test: z = -0.356, p

= 0.722, effect size: r = 0.103, Fig. 8i, right) tasks. These results further confirm that the predictive coding feedback plays a critical role in producing the center-surround inhibition in expectation.

Discussion

The present results provide support for an attentional modulation-independent center-surround inhibition profile of expectation and further reveal its underlying neural computations. Specifically, on both OD and SFD tasks, the finest-grained discrimination performance, indexed by the lowest thresholds, of the expected orientation confirmed the previous notion that expectation had a facilitatory effect on various perceptions[7, 47-52]. Whereas the coarser-grained discrimination performance, indexed by the higher thresholds, of orientations very similar to the expected orientation relative to orientations more distinct from the expected orientation, demonstrated a classical inhibitory zone surrounding the focus of expectation (i.e., the center-surround inhibition profile, Fig. 2). One could argue that this profile was derived from top-down attention rather than expectation. Compared to unexpected gratings with much low validity, the expected grating with very high validity in our study, presumably, had more degree of top-down attention that has been proven to display the similar center-surround inhibition profile in orientation space by previous studies[53-56] and computational models[57-59]. In other words, our study may not examine a center-surround inhibition profile of expectation, but instead of top-down attention. It is important to note that, in our study, for each grating, participants performed the same discrimination task at threshold, measured by the QUEST staircase procedure (75% correct)[34], which could maximally (although not completely) control the difference in top-down attention among distances. More importantly, our observed

center-surround inhibition profile of expectation in orientation space was independent of attentional modulations by the task-relevance of orientation (i.e., OD and SFD tasks, Fig. 2), consistent with previous findings[7,8,60-65], showing an independency between attention and expectation. If the center-surround inhibition profile was derived from attention rather than expectation, then we should have not observed it on SFD task, in which the orientation was never task-relevant. Participants did not need to direct attention to this task-irrelevant feature, and therefore yielding none of profiles in orientation space.

The center-surround inhibition profile of expectation evident in our study is consistent with what has been observed for spatial attention[66-74], feature-based attention[53-56,75-78], working memory[79-81], and visual perceptual learning[35], in various feature spaces. This suggests that center-surround inhibition could be a unifying principle underlying a diversity of visual representations, as previously proposed by the selective tuning model[57-59], however, the extent of the inhibitory zone varied largely across these domains and features. For example, within the orientation space, the inhibitory zone was about 20°, 45°, and 54° for expectation evident here, feature-based attention[21], and visual perceptual learning[35], respectively; within the feature-based attention, it was about 30° and 45° in color[77] and motion direction[53] spaces, respectively. Although speculative, these findings hint at the exciting possibility that the width of an inhibitory surround may vary flexibly with both cognitive and task demands, ultimately facilitating our perception and behavior to

adapt to changing environment. Mechanistically, the center-surround inhibition profile can be optimal to locally resolve competition between inputs that overlap in their neural representations, specifically attenuating the interference from nearby irrelevant and confusable representations that would be presumably within the same cortical map, and therefore at the largest risk to confuse the current processing. Given the presence of a well-defined map-based organization of the cerebral cortex[82-87], the center-surround inhibition would be beneficial across all features and therefore serves as a canonical neural computation that sharpens various cognitive processing across different domains.

Strikingly, we found that the center-surround inhibition profile of expectation observed behaviorally can be accounted by sharpening of tuning curves of neurons of the expected orientation, as revealed by the computational model (Fig. 3), orientation adjustment (Fig. 5 and Fig. 6), and orientation discrimination (Fig. 7) experiments. These changes – sharpening of tuning curves – are not only in line with the sharpening hypothesis of expectation developed by previous neurophysiological[10-14], electro-/magneto-encephalogram[15-19], and fMRI[7,20-22] studies that has invoked the tuning sharpening as the neural basis of expectation-related effects (e.g., the sharpening of tuning curves facilitates fine orientation discrimination by increasing the activity difference between similar orientations), but also extend this hypothesis by identifying the same neural computation in its center-surround inhibition profile. More importantly, we further found that this profile of expectation can be accounted

by the tuning shift computation that neurons of unexpected orientations shift their spectral tuning toward the expected orientation (Fig. 3, Fig. 6, and Fig. 7). In accordance with expectation, several other cognitive processing has also been shown to shift neuronal tuning curves or receptive fields toward the target, such as spatial[88-94] and feature-based[95-98] attention, as well as visual search[58,99-103] and perceptual learning[104-106]. Interestedly, several brain imaging studies have reported that expectation alters the baseline[16,107,108] or gain[32,109-111] of neurons in visual areas, consistent with a classical hypothesis, i.e., the labeled-line theories of visual information processing[112-116], which posits that neurons in sensory cortex act as labeled lines with fixed tuning properties that encode input features consistently, regardless of task demands. However, this theory does not account for either tuning curve sharpening or tuning curve shifts of sensory neurons induced by expectation in our study. These changes we observed in the spectral tuning profiles of sensory neurons, conversely, are not only strongly supported by the matched filter hypothesis that neurons could act as matched filters and reshape or shift their tuning to match the target exactly[114], but also compatible with both proposals from computational models[58,100] and Kalman filtering schemes for the signal detection[103].

Although our study succeeded in linking the center-surround inhibition profile of expectation directly with the response of sensory neurons whose tuning properties make them optimal for demarcating the expected information from various unexpected information, we cannot deny a potential contribution from other cognitive

processes, such as the decision making. Indeed, previous studies have indicated that expectations primarily influence decisions by modulating post-perceptual stages of information processing[8,117-119] or modulate interactions between lower sensory and higher decision areas[109,120]. In addition, these changes in the spectral tuning profiles of sensory neurons evident here derive mainly from psychophysics and computational models. To fully understand how changes in sensory responses contribute to both expectation and its center-surround inhibition profile, further work is needed using neurophysiological techniques or ultra-high field fMRI to explore the locus of events responsible for expectation-induced changes, the identity of neurons that undergo these changes, their patterns of connections, their interactions with higher decision processing, and underlying synaptic bases, especially for our observed shifts in unexpected orientation tunings.

In addition, the emerged center-surround inhibition of expectation in the pretrained DPCNN, is not only in line with previous studies and theories that interprets expectation within the predictive coding framework[7,9,29-33], but also adds strong evidence supporting artificial neural networks' potential to perform various human-like representations, such as visual perceptual learning[35,121,122] and hierarchical coding[123], face processing[124], contour integration[125], and the perception of illusory contours[45]. More importantly, our ablation studies further confirm a critical role of the predictive coding feedback in producing the center-surround inhibition in expectation (Fig. 8). Although our similarities between artificial neural networks and humans

were mostly qualitative, the artificial neural network can provide new ways of studying expectation from behavior to physiology, serving as a test bed for various theories and assisting in generating predictions for physiological studies.

In sum, our study provides, to the best of our knowledge, the first evidence for a center-surround inhibition profile of expectation and how it is supported by not only changes in the tuning curves of sensory neurons but also the predictive coding framework, leading the way towards diversifying models or theories and takes a significant step in unraveling the neuronal computations underlying expectation, or, more generally, top-down processing.

Methods

Participants

A total of 24 healthy human adults (16 females, 19–26 years old) were involved in the study. All of them participated in the profile experiments, 20 and 18 of them participated in the orientation adjustment and orientation discrimination experiments, respectively. All participants with normal or corrected-to-normal vision, gave written informed consent, were right-handed, and were naïve to the purpose of the experiments. The procedures and protocols were approved by the human participants review committee of School of Psychology at South China Normal University

Apparatus

The experiments were conducted in a dark, acoustically shielded room. Visual stimuli were displayed on an IIYAMA color graphic monitor (model: HM204DT; refresh rate: 60 Hz; resolution: 1,280 × 1,024; size: 22 inches) at a viewing distance of 57 cm. Participants' head position was stabilized using a chin rest.

Experimental stimuli

Visual stimuli were two consecutive sinusoidal grating stimuli (1.0 contrast, random phase, radius 10°), which were generated using MATLAB (MathWorks, Natick, MA) in conjunction with the Psychophysics Toolbox[126], and displayed centrally on the gray background (11.196 cd/m^2). A white fixation point (radius 0.278°) was always

presented at the center of the screen throughout the experiment. The auditory cue consisted of two pure tones (240 Hz and 540 Hz), presented over earphones.

Experimental design and statistical analysis

Profile experiment

Experimental design. The profile experiment consisted of baseline and main experiments and the baseline experiment always preceded the main experiment. The two experiments were the same, except for the predicting probability relationship between the auditory cue and the orientation (20°, 30°, 40°, 50°, 60°, and 70°) of the first grating. For the baseline experiment, the auditory cue, comprising either a low- (240 Hz) or high- (540 Hz) frequency tone, predicted the orientation of the first grating with equal validity (16.67%, Fig. 1a, left). In the main experiment, this low- or high-frequency tone auditory cue predicted the orientation (20° or 70°) of the first grating with 75% validity. In the remaining 25% of trials, this orientation was chosen randomly and equally from four non-predicted orientations (30°, 40°, 50°, and 60°, Fig. 1a, right). Thus, for each participant, there were two types of expected conditions: Expect 20° and Expect 70°, and for both conditions, there were five possible distances in orientation space between the expected and test gratings, ranging from Δ0° through Δ40° with a step size of 10°. Note that, the matches between the tone (low- or high-frequency) of auditory cue and the expected orientation (20° or 70°) of the first grating were flipped across participants, and the order was also counterbalanced across participants. For each participant, although the tone of auditory cue could not

predict 20° or 70° orientation in the baseline experiment, whose trials with the same tone that was matched with 20° or 70° orientation in the main experiment, were defined as Baseline 20° (i.e., the baseline of Expect 20°) and Baseline 70° (i.e., the baseline of Expect 70°) conditions, respectively.

Both the baseline and main experiments consisted of two tasks: the orientation discrimination (OD) task and spatial frequency discrimination (SFD) task; with the two tasks occurring on different days; the order of the two tasks was counterbalanced across participants. Differently, the baseline experiment consisted of 4 blocks (two for OD task and the other two for SFD task), and each block had 2 QUEST staircases[34] for each of six orientations (20°, 30°, 40°, 50°, 60°, and 70°). The main experiment consisted of 2 blocks (one for OD task and the other one for SFD task), and each block had 24 QUEST staircases for the expected orientations (20° and 70°) and 2 QUEST staircases for each of unexpected orientations (30°, 40°, 50°, and 60°). Each QUEST staircase comprised 40 trials, and on each trial, a low- (240 Hz) or a high- (540 Hz) frequency tone (i.e., the auditory cue) was randomly and equally presented for 200 ms, followed by an 1800 ms fixation interval. Then, two consecutive gratings were each presented for 150-ms and separated by a 300-ms blank interval (Fig. 1b). Participants were first asked to make a two-alternative forced-choice (2AFC) judgment of either the orientation (clockwise or anticlockwise, where orientation was task-relevant) or the spatial frequency (lower or higher, where orientation was task-irrelevant) of the second grating relative to the first, on the OD and SFD tasks,

respectively. Then, participants were required to make another 2AFC judgment on tone of the auditory cue, either low or high. In the baseline experiment, for both OD and SFD tasks, the orientation of the first grating was chosen randomly and equally from 20°, 30°, 40°, 50°, 60°, and 70°, while its spatial frequency was fixed at 0.9 cycles/°. The second grating differed slightly from the first in terms of both orientation and spatial frequency. Differently, for OD task, its orientation difference ($\Delta\theta$°, where orientation was task-relevant) varied trial by trial and was controlled by the QUEST staircase to estimate participants' OD thresholds (75% correct), while its spatial frequency difference was set at 0.06 cycle/°; for SFD task, its spatial frequency difference ($\Delta\lambda$ cycles/°, where orientation was task-irrelevant) varied trial by trial and was controlled by the QUEST staircase to estimate participants' SFD thresholds (75% correct), while its orientation difference was set at 4.8° based on pretest data. Similarly, for each participant, the discrimination threshold obtained during the baseline experiment was used to set the undiscriminated feature difference (i.e., the spatial frequency and orientation for OD and SFD tasks, respectively) during the main experiment, to make the stimuli as similar as possible in both contexts.

Model fitting and comparison. In both OD and SFD tasks, for two expected conditions and each distance (i.e., Δ0° - Δ40°), we computed a discrimination sensitivity (*DS*) to quantify how much the discrimination threshold (*DT*) changed between baseline ($DT_{baseline}$) and main (DT_{main}) experiments: $DS = DT_{baseline} - DT_{main}$. Because the *DS* from two expected conditions (Expect 20° and Expect 70°) showed as

a similar pattern, they were pooled together for further analysis (unless otherwise stated, we present average data from two expected conditions). During both tasks, for each participant, a monotonic model and a non-monotonic model to the averaged *DS* were fitted. The monotonic and non-monotonic models were implemented as the Gaussian and Mexican-hat (i.e., a negative second derivative of a Gaussian function) functions[35,36], respectively, as follows:

$$\text{Gaussian function}: \ y = y0 + \frac{2A}{W\sqrt{2\pi}} e^{-2(\frac{x}{w})^2}$$

$$\text{Mexican-hat function}: \ y = \frac{2H}{\sqrt{3m}\pi^{\frac{1}{4}}} e^{-\frac{x^2}{2m^2}} \left(1 - \frac{x^2}{m^2}\right) + y1$$

where *y* is the measured *DS*, *x* is the distance; *w*, *A*, and *y0* are the three parameters controlling the shape of the Gaussian function; *m*, *H*, and *y1* are three parameters controlling the shape of the Mexican-hat function. To compare these two models to our data, we first computed the Akaike information criterion (*AIC*)[37] and Bayesian information criterion (*BIC*)[38], with the assumption of a normal error distribution as follow:

$$AIC = Nln\left(\frac{RSS}{N}\right) + 2K + \frac{2K(K+1)}{N-K-1}$$

$$BIC = Nln\left(\frac{RSS}{N}\right) + Kln(N)$$

where *N* is the number of observations, *K* is the number of free parameters, and *RSS* is residual sum of squares[127]. Then, we further calculated the likelihood ratio (*LR*) and Bayes factor (*BF*) of the non-monotonic models (Mexican-hat) over monotonic model

(Gaussian) based on *AIC*[39] and *BIC*[40] approximation, respectively, as follows;

$$LR = e^{\left(\frac{AIC_G - AIC_M}{2}\right)}$$

$$BF = e^{\left(\frac{BIC_G - BIC_M}{2}\right)}$$

where AIC_G and BIC_G are for the Gaussian model and AIC_M and BIC_M are for Mexican-hat models.

Computational models of the center-surround inhibition in expectation

Prior to initiating model fitting, for both OD and SFD tasks, we first transformed the negative values of thresholds during baseline and main experiments into smooth population response profiles using linear interpolation, respectively. Subsequently, we fitted two candidate models, namely Tuning sharpening model and Tuning shift model (Fig. 3a), to these population response profiles for each participant. In both models, the idealized tuning function for each channel was defined by the Gaussian functions:

$$f(x) = A * e^{-\frac{(x - x0)^2}{\sigma^2}}$$

where *x* is the grating orientation, *A* is the amplitude of tuning function, *x0* is the location, and σ is the width. Six and five tuning channels were hypothesized for data in baseline and main experiments, respectively. For the Tuning sharpening model, the tuning width of each channel's tuning function is parameterized by σ, while all tuning functions are evenly distributed with 10° spacing on the x-axis and the areas under the

curves (response energy) are identical. Conversely, for the Tuning shift model, the location of each channel's tuning function is parameterized by *x0*, while they all share the same tuning amplitude and width. The parameter *x0* was constrained within ± 5° of the grating orientation limits, ranging from 15° to 75° during the baseline experiment, 15° to 65° and 25° to 75° for expected 20° and expected 70° conditions, respectively, during the main experiment. The parameters σ was set within the range of 0.01 to 200 to ensure the comparable goodness of fit. For both models, parameters were varied to obtain the minimal sum of squared errors between the population response profile and the model prediction, which is the sum of all channels' tuning responses. To statistically compare the two models, for both orientation and SF discrimination tasks, we computed the root mean squared deviation (RMSD)[41] of the two fitted models for each participant during baseline and main experiments:

$$\mathrm{RMSD} = \sqrt{\frac{SSE}{(N-K)}}$$

where *SSE* is the sum of squared errors. *N* is the number of data points (i.e., 51 and 61), and *K* is the number of model parameters.

Orientation adjustment experiment

Experimental design. The protocol of orientation adjustment experiment was similar with that of the profile experiment, except for two aspects. First, there were 4 possible (20°, 40°, 50°, and 70°) orientations for the first grating: 20°/70° (Δ0° deviated from

the expected orientation) and 40°/50° (Δ20°/Δ30° deviated from the expected orientation). Second, in both baseline and main experiments, the second grating was set as a random orientation within the range of 0° to 90°, and participants were required to rotate the orientation of the second grating to match the first (Fig. 5a). Each participant completed 8 blocks of 48 trials in the baseline experiment, and 16 blocks of 48 trials in the main experiment.

Modeling response error. Response error was measured as the angular difference between the orientation of the first grating and the adjusted orientation of the second grating, such that errors ranged from 0˚ (a perfect response) to ± 90˚ (a maximally imprecise response). To evaluate performance, we categorized the response errors for each participant according to different conditions and modeled their distributions as a three-component mixture model[42]. This model comprised a von Mises distribution (∅) corresponding to trials in which the grating orientation was encoded and a uniform distribution (p_g) accounting for the probability of random guessing without encoding[43]:

$$p(\theta) = (1 - p_g)\emptyset_{\mu,k} + p_g(\frac{1}{2\pi})$$

where θ is the adjusted orientation value, $\emptyset$ denotes the Von Mises distribution with mean μ and shape parameter k, and p_g represents a uniform distribution. Specifically, the von Mises probability density function for the angle x is given by:

$$\emptyset(x \mid \mu, k) = \frac{e^{kcos(x-\mu)}}{2\pi I_0(k)}$$

where $I_0(k)$ is the modified Bessel function of the first kind of order 0, with this scaling constant chosen so that the distribution suns to unity:

$$\int_{-\pi}^{\pi} e^{k \cos x}\, dx = 2\pi I_0(k)$$

Here, we obtained maximum likelihood estimates for 3 parameters: (1) the systematic shift of von Mises distribution (*mu*), which reflects distribution shift away from the target grating orientation; (2) the dispersion of the von Mises distribution ($s.d.=\sqrt{1/k}$), which reflects response precision or resolution of representation; and (3) the height of the uniform distribution (*g*), which reflects the probability of guessing.

Orientation discrimination experiment

Experimental design. The protocol of orientation discrimination experiment was very similar with that of the orientation adjustment experiment, except for two aspects (Fig. 7a). First, there were 3 possible (20°, 45°, and 70°) orientations for the first grating: 20°/70° (Δ0° deviated from the expected orientation) and 45° (Δ25° deviated from the expected orientation). Second, in both baseline and main experiments, the second grating were 1°, 3°, 5°, 7°, and 9° deviated from the first grating, either clockwise or counterclockwise. Participants were asked to make a 2AFC judgment of the orientation of the second grating relative to the first, either clockwise or anticlockwise. Each participant completed 10 blocks of 120 trials in the baseline experiment, and 20 blocks of 160 trials in the main experiment.

Data fitting and analysis. We first constructed a psychometric function for each

condition shown in Fig. 7. We plotted the percentage of trials in which participants indicated the orientation of second grating that was anticlockwise or clockwise to the first for 20° (Baseline 20° and Expect 20°) and 70° (Baseline 70° and Expect 70°) conditions, respectively, as a function of the real orientation difference between two gratings. For each participant and each condition, the psychometric values at ten orientation differences were fitted to a cumulative Gaussian using Bayesian inference, implemented in the *Psignifit toolbox* for Matlab (Version 4)[128], and we interpolated the data to find the slope (orientation uncertainty) and PSE (point of subjective equality, which is the shift here) as an index for Tuning sharpening and Tuning shift models, respectively.

Artificial neural networks for the center-surround inhibition in expectation

We trained a deep predictive coding neural network (DPCNN), modified from Predify[44,45] to perform both the OD and SFD tasks. Relative to the reference, on the OD task, DPCNN was trained to classify whether the target was tilted clockwise or counterclockwise; whereas on the SFD task, it was trained to classify whether the target had lower or higher spatial frequency. For both tasks, the DPCNN consisted of 6 feedforward encoding layers (*e1-e6*), 5 generative feedback decoding layers (*d1-d5*), and 3 fully connected (fc) layers (Fig. 8a). The reconstruction error (*E1-E5*) is computed and used for the proposed predictive coding updates[31], denoting by *P.C.* loops. Note that the updating is only applied to *e1-e5* and for the last layer *e6*, there is

no feedback. Before the layer *e0*, we obtained the pixel difference between the target and reference images, which was then superimposed on the channels of the reference image. This superimposed image was set as the input of the network and remained constant over timesteps. Besides, we used a feedforward encoding layer (i.e., *e0*) to match the number of the channels between superimposed feature maps and the pre-trained DPCNN. Additionally, to further determine the contribution of predictive coding framework to center-surround inhibition in expectation, we also trained the same network but removing feedback predictive coding iterations (a standard feedforward CNN, i.e., a modified network of AlexNet[46]). Note that all these architects were built to mimic our hypothesis of the visual pathway involved in expectation[5-8,61] and could learn a lower-dimensional latent representation of a high-dimensional input space[129]. During the training, the last layer was trained to capture the difference between the target and reference and finally obtain the classification by softmax, to model decision making in our 2AFC paradigm (Fig. 1c), in which participants were asked to make a 2AFC judgment of the orientation (either clockwise or anticlockwise) or the spatial frequency (either lower or higher) of the second grating (target) relative to the first (reference) in OD and SFD tasks, respectively. Moreover, for the OD task, the orientation difference between the target and reference in the network was set to 5°; for the SFD task, the spatial wavelength difference between them was set to 0.5.

For each task, both the DPCNN and CNN were independently and randomly

trained 12 times. For each time, the trained orientation was chosen randomly from 0° to 180°; the 9 test gratings were 0°, ± 10°, ± 20°, ± 30°, and ± 40° deviated (clockwise and counterclockwise) from the trained orientation. All grating stimuli (phase: random) were centered on 227 × 227-pixels images with gray background. To improve the robustness of our model, we trained the network on all combinations of several parameters: contrast (0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8), SD of the Gaussian additive noise (5, 25, and 45), and spatial wavelength (5, 10, 15, 20, 25, 30, 40, 50, 60, and 80 pixels) for the OD task; contrast ranging from 0.1 through 0.8 with a step size of 0.05 and SD of the Gaussian additive noise ranging from 3 through 60 with a step size of 3 for the SFD task. For each training, there were thus a total of 840 images; 600 images were the training set and the other 240 images were the test set. For both OD and SFD tasks, during the training set, there were 480 images for the expected orientation (Δ0°) and 30 images for each of unexpected orientations (Δ10°, Δ20°, Δ30°, and Δ40°); during the test set, there were 48 images for each of distances (Δ0°- Δ40°). For each distance, the training effect was defined as the accuracy difference ($ACC_{difference}$) between the pre- ($ACC_{baseline}$) and post- ($ACC_{trained}$) training.

References

1. Bar, M. Visual objects in context. *Nat. Rev. Neurosci*. **5**, 617–629 (2004).
2. Bar, M. The proactive brain: memory for predictions. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. **364**, 1235–1243 (2009).
3. Teufel, C. & Fletcher, P. C. Forms of prediction in the nervous system. *Nat. Rev. Neurosci*. **21**, 231–242 (2020).
4. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Ann. Rev. Neurosci*. **40**, 373–394 (2017).
5. De Lange, F. P., Heilbron, M. & Kok, P. How do expectations shape perception? *Trends Cogn. Sci*. **22**, 764–779 (2018).
6. Press, C., Kok, P. & Yon, D. The perceptual prediction paradox. *Trends Cogn. Sci*. **24**, 13–24 (2020).
7. Kok, P., Jehee, J. F. & De Lange, F. P. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* **75**, 265–270 (2012).
8. Summerfield, C. & De Lange, F. P. Expectation in perceptual decision making: neural and computational mechanisms. *Nat. Rev. Neurosci*. **15**, 745–756 (2014).
9. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis*. **20**, 1434–1448 (2003).
10. Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J. & Ungerleider, L. G. Encoding of stimulus probability in macaque inferior temporal cortex. *Curr. Biol*. **26**, 2280–2290 (2016).

11. Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M. & Keller, G. B. Experience-dependent spatial expectations in mouse visual cortex. *Nat. Neurosci.* **19**, 1658–1664 (2016).

12. Kaposvari, P., Kumar, S. & Vogels, R. Statistical learning signals in macaque inferior temporal cortex. *Cereb. Cortex* **28**, 250–266 (2018).

13. Meyer, T. & Olson, C. R. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl Acad. Sci. USA* **108**, 19401–19406 (2011).

14. Schwiedrzik, C. M. & Freiwald, W. A. High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron* **96**, 89–97 (2017).

15. Aitken, F., Turner, G. & Kok, P. Prior expectations of motion direction modulate early sensory processing. *J. Neurosci.* **40**, 6389–6397 (2020).

16. Kok, P., Mostert, P. & De Lange, F. P. Prior expectations induce prestimulus sensory templates. *Proc. Natl Acad. Sci. USA* **114**, 10473–10478 (2017).

17. Todorovic, A., van Ede, F., Maris, E. & de Lange, F. P. Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *J. Neurosci.* **31**, 9118–9123 (2011).

18. Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., ... & Griffiths, T. D. Neural signatures of perceptual inference. *eLife* **5**, e11476 (2016).

19. Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L. & Dehaene, S. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl Acad. Sci. USA* **108**, 20754–20759 (2011).

20. Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W. & Muckli, L. Stimulus predictability reduces responses in primary visual cortex. *J. Neurosci*. **30**, 2960–2966 (2010).

21. Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M. & Egner, T. Neural repetition suppression reflects fulfilled perceptual expectations. *Nat. Neurosci*. **11**, 1004–1006 (2008).

22. Yon, D., Gilbert, S. J., de Lange, F. P. & Press, C. Action sharpens sensory representations of expected outcomes. *Nat. Commun*. **9**, 4288 (2018).

23. Richter, D., Heilbron, M. & de Lange, F. P. Dampened sensory representations for expected input across the ventral visual stream. *Oxford Open Neurosci*. **1**, kvac013 (2022).

24. Kumar, S., Kaposvari, P. & Vogels, R. Encoding of predictable and unpredictable stimuli by inferior temporal cortical neurons. *J. Cogn. Neurosci.* **29**, 1445–54 (2017).

25. Blakemore, S., Wolpert, D. M. & Frith, C. D. Central cancellation of self-produced tickle sensation. *Nat. Neurosci.* **1**, 635–40 (1998).

26. Blank, H. & Davis, M. H. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biol*. **14**, 1–32 (2016).

27. Han, B., Mostert, P. & De Lange, F. P. Predictable tones elicit stimulus-specific suppression of evoked activity in auditory cortex. *NeuroImage* **200**, 242–9 (2019).

28. Richter, D., Ekman, M. & de Lange, F. P. Suppressed sensory response to

predictable object stimuli throughout the ventral visual stream. *J. Neurosci.* **38**, 7452–7461 (2018).

29. Feldman, H. & Friston, K. J. Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* **4**, 215 (2010).
30. Friston, K. J. A theory of cortical responses. Philos. *Trans. R. Soc. Lond. B Biol. Sci.* **360**, 815–836 (2005).
31. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
32. Summerfield, C. & Koechlin, E. A neural representation of prior information during perceptual inference. *Neuron* **59**, 336–347 (2008).
33. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
34. Watson, A. B. & Pelli, D. G. Quest: A Bayesian adaptive psychometric method. *Percept. Psychophys.* **33**, 113–120 (1983).
35. Shen, S., Sun, Y., Lu, J., Li, C., Chen, Q., Mo, C., Fang, F. & Zhang, X. Profiles of visual perceptual learning in feature space. *iScience* **27**, 109128 (2024).
36. Wang, S., Huang, L., Chen, Q., Wang, J., Xu, S. & Zhang, X. Awareness-dependent normalization framework of visual bottom-up attention. *J. Neurosci.* **41**, 9593–9607 (2021).
37. Akaike, H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–265 (1973).

38. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).

39. Burnham, K. P. & Anderson, D. R. A practical information-theoretic approach. *Model selection and multimodel inference* 2, (2002).

40. Wagenmakers, E. J. A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* **14**, 779–804 (2007).

41. Pitt, M. A., Myung, I. J. & Zhang, S. Toward a method of selecting among computational models of cognition. *Psychol. Rev.* **109**, 472 (2002).

42. Suchow, J. W., Brady, T. F., Fougnie, D. & Alvarez, G. A. Modeling visual working memory with the MemToolbox. *J. Vision* **13**, 9–9 (2013).

43. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).

44. Choksi, B., Mozafari, M., Biggs O'May, C., Ador, B., Alamia, A. & VanRullen, R. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Adv. Neural Inf. Process. Syst.* **34**, 14069–14083 (2021).

45. Pang, Z., O'May, C. B., Choksi, B. & VanRullen, R. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks* **144**, 164–175 (2021).

46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).

47. Cheadle, S., Egner, T., Wyart, V., Wu, C. & Summerfield, C. Feature expectation heightens visual sensitivity during fine orientation discrimination. *J. Vision* **15**,

14–14 (2015).

48. Esterman, M. & Yantis, S. Perceptual expectation evokes category-selective cortical activity. *Cereb. Cortex* **20**, 1245–1253 (2010).

49. Mareschal, I., Calder, A. J. & Clifford, C. W. Humans have an expectation that gaze is directed toward them. *Curr. Biol.* **23**, 717–721 (2013).

50. McAuley, J. D. & Kidd, G. R. Effect of deviations from temporal expectations on tempo discrimination of isochronous tone sequences. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 1786 (1998).

51. Stein, T. & Peelen, M. V. Content-specific expectations enhance stimulus detectability by increasing perceptual sensitivity. *J. Exp. Psychol.* Gen. **144**, 1089 (2015).

52. Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).

53. Fang, M. W. & Liu, T. The profile of attentional modulation to visual features. *J. Vis.* **19**, 13–13 (2019).

54. Liu, T., Fang, M. W. & Saba-Sadiya, S. Adaptive visual selection in feature space. *Psychon. Bull. Rev.* **30**, 994–1003 (2023).

55. Tombu, M. & Tsotsos, J. K. Attending to orientation results in an inhibitory surround in orientation space. *Percept. Psychophys.* **70**, 30–35 (2008).

56. Yoo, S. A., Tsotsos, J. K. & Fallah, M. The attentional suppressive surround: eccentricity, location-based and feature-based effects and interactions. *Front. Neurosci.* **12**, 710 (2018).

57. Tsotsos, J. K., Culhane, S. & Cutzu, F. From theoretical foundations to a hierarchical circuit for selective attention. *Visual Attention and Cortical Circuits* 285-306 (2001).

58. Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. & Nuflo, F. Modeling visual attention via selective tuning. *Artif. Intell.* **78**, 507–545 (1995).

59. Tsotsos, J. K., Rodriguez-Sanchez, A. J., Rothenstein, A. L. & Simine, E. The different stages of visual recognition need different attentional binding strategies. *Brain Res.* **1225**, 119–132 (2008).

60. Rungratsameetaweemana, N. & Serences, J. T. Dissociating the impact of attention and expectation on early sensory processing. *Curr. Opin. Psychol.* **29**, 181–186 (2019).

61. Summerfield, C. & Egner, T. Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* **13**, 403–409 (2009).

62. Gordon, N., Tsuchiya, N., Koenig-Robert, R. & Hohwy, J. Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLoS Biol.* **17**, e3000233 (2019).

63. Tal-Perry, N. & Yuval-Greenberg, S. The spatiotemporal link of temporal expectations: contextual temporal expectation is independent of spatial attention. *J. Neurosci.* **42**, 2516–2523 (2022).

64. Wilsch, A., Mercier, M. R., Obleser, J., Schroeder, C. E. & Haegens, S. Spatial attention and temporal expectation exert differential effects on visual and auditory discrimination. *J. Cogn. Neurosci.* **32**, 1562–1576 (2020).

65. Zuanazzi, A. & Noppeney, U. Distinct neural mechanisms of spatial attention and expectation guide perceptual inference in a multisensory world. *J. Neurosci.* **39**, 2301–2312 (2019).

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

67. Hopf, J. M., Boehler, C. N., Luck, S. J., Tsotsos, J. K., Heinze, H. J. & Schoenfeld, M. A. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proc. Natl Acad. Sci. USA* **103**, 1053–1058 (2006).

68. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).

69. Schall, J. D., Sato, T. R., Thompson, K. G., Vaughn, A. A. & Juan, C. H. Effects of search efficiency on surround suppression during visual selection in frontal eye field. *J. Neurophysiol.* **91**, 2765–2769 (2004).

70. Mounts, J. R. Evidence for suppressive mechanisms in attentional selection: Feature singletons produce inhibitory surrounds. *Percept. Psychophys.* **62**, 969–983 (2000).

71. Müller, N. G. & Kleinschmidt, A. The attentional ‘spotlight’s’ penumbra: Center–surround modulation in striate cortex. *Neuroreport* **15**, 977–980 (2004).

72. Müller, N. G., Mollenhauer, M., Rosler, A. & Kleinschmidt, A. The attentional field has a Mexican hat distribution. *Vision Res.* **45**, 1129–1137 (2005).

73. Boehler, C. N., Tsotsos, J. K., Schoenfeld, M. A., Heinze, H. J. & Hopf, J. M.

The center-surround profile of the focus of attention arises from recurrent processing in visual cortex. *Cereb. Cortex* **19**, 982–991 (2009).

74. Boehler, C. N., Tsotsos, J. K., Schoenfeld, M. A., Heinze, H. J. & Hopf, J. M. Neural mechanisms of surround attenuation and distractor competition in visual search. *J. Neurosci.* **31**, 5213–5224 (2011).

75. Fang, M. W., Becker, M. W. & Liu, T. Attention to colors induces surround suppression at category boundaries. *Sci. Rep.* **9**, 1443 (2019).

76. Bartsch, M. V., Loewe, K., Merkel, C., Heinze, H. J., Schoenfeld, M. A., Tsotsos, J. K. & Hopf, J. M. Attention to color sharpens neural population tuning via feedback processing in the human visual cortex hierarchy. *J. Neurosci.* **37**, 10346–10357 (2017).

77. Störmer, V. S. & Alvarez, G. A. Feature-based attention elicits surround suppression in feature space. *Curr. Biol.* **24**, 1985–1988 (2014).

78. Loach, D., Frischen, A., Bruce, N. & Tsotsos, J. K. An attentional mechanism for selecting appropriate actions afforded by graspable objects. *Psychol. Sci.* **19**, 1253–1257 (2008).

79. Kiyonaga, A. & Egner, T. Center-surround inhibition in working memory. *Curr. Biol.* **26**, 64–68 (2016).

80. Shi, R., Gao, H. & Zhang, Q. The extent of center-surround inhibition for colored items in working memory. *Mem. Cogn.* **49**, 733–746 (2021).

81. Shi, R., Qi, M. & Gao, H. The ERP correlates of color-based center-surround inhibition in working memory. *Int. J. Psychophysiol.* **181**, 160–169 (2022).

82. Eickhoff, S. B., Yeo, B. T. & Genon, S. Imaging-based parcellations of the human brain. *Nat. Rev. Neurosci*. **19**, 672–686 (2018).

83. Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... & Van Essen, D. C. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

84. Mountcastle, V. B. The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997).

85. Tanaka, K. Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* **13**, 90–99 (2003).

86. Van Essen, D. C. & Glasser, M. F. Parcellating cerebral cortex: how invasive animal studies inform noninvasive mapmaking in humans. *Neuron* **99**, 640–663 (2018).

87. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual field maps in human cortex. *Neuron* **56**, 366–383 (2007).

88. Connor, C. E., Preddie, D. C., Gallant, J. L. & Van Essen, D. C. Spatial attention effects in macaque area V4. *J. Neurosci*. **17**, 3201–3214 (1997).

89. Fox, K. J., Birman, D. & Gardner, J. L. Gain, not concomitant changes in spatial receptive field properties, improves task performance in a neural network attention model. *eLife* **12**, e78392 (2023).

90. Klein, B. P., Harvey, B. M. & Dumoulin, S. O. Attraction of position preference by spatial attention throughout human visual cortex. *Neuron* **84**, 227–237 (2014).

91. Sheremata, S. L. & Silver, M. A. Hemisphere-dependent attentional modulation of human parietal visual field representations. *J. Neurosci.* **35**, 508–517 (2015).

92. Tolias, A. S., Moore, T., Smirnakis, S. M., Tehovnik, E. J., Siapas, A. G. & Schiller, P. H. Eye movements modulate visual receptive fields of V4 neurons. *Neuron* **29**, 757–767 (2001).

93. Vo, V. A., Sprague, T. C. & Serences, J. T. Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. *J. Neurosci.* 37, 3386–3401 (2017).

94. Womelsdorf, T., Anton-Erxleben, K., Pieper, F., & Treue, S. Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat. Neurosci.* **9**, 1156–1160 (2006).

95. Cukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* **16**, 763–770 (2013).

96. David, S. V., Hayden, B. Y., Mazer, J. A. & Gallant, J. L. Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* **59**, 509–521 (2008).

97. Motter, B. C. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J. Neurosci.* **14**, 2190–2199 (1994).

98. van Es, D. M., Theeuwes, J. & Knapen, T. Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. *eLife* **7**, e36928 (2018).

99. Carrasco, M., Ling, S. & Read, S. Attention alters appearance. *Nat. Neurosci.* **7**, 308–313 (2004).

100. Compte, A. & Wang, X. J. Tuning curve shift by attention modulation in cortical neurons: a computational study of its mechanisms. *Cereb. Cortex* **16**, 761–778 (2006).

101. Lee, D. K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nat. neurosci.* **2**, 375–381 (1999).

102. Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719 (1993).

103. Rao, R. P. & Ballard, D. H. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput.* **9**, 721–763 (1997).

104. Hanson, H. M. Effects of discrimination training on stimulus generalization. *J. Exp. Psychol.* **58**, 321 (1959).

105. Schumacher, J. W., McCann, M. K., Maximov, K. J. & Fitzpatrick, D. Selective enhancement of neural coding in V1 underlies fine-discrimination learning in tree shrew. *Curr. Biol.* **32**, 3245–3260 (2022).

106. Spence, K. W. The differential response in animals to stimuli varying within a single dimension. *Psychol. Rev.* **44**, 430 (1937).

107. Lucci, G., Berchicci, M., Perri, R. L., Spinelli, D. & Di Russo, F. Effect of target probability on pre-stimulus brain activity. *Neuroscience* **322**, 121–128 (2016).

108. van Ede, F., Jensen, O. & Maris, E. Tactile expectation modulates pre-stimulus

β-band oscillations in human sensorimotor cortex. *Neuroimage* **51**, 867–876 (2010).

109. Foley, N. C., Kelly, S. P., Mhatre, H., Lopes, M. & Gottlieb, J. Parietal neurons encode expected gains in instrumental information. *Proc. Natl Acad. Sci. USA* **114**, E3315–E3323 (2017).

110. Kok, P., Van Lieshout, L. L. & De Lange, F. P. Local expectation violations result in global activity gain in primary visual cortex. *Sci. Rep.* **6**, 37706 (2016).

111. Voss, M., Ingram, J. N., Wolpert, D. M. & Haggard, P. Mere expectation to move causes attenuation of sensory signals. *PLoS One* **3**, e2866 (2008).

112. Adrian, E. D. & Matthews, R. The action of light on the eye: Part I. The discharge of impulses in the optic nerve and its relation to the electric changes in the retina. *J. Physiol.* **63**, 378 (1927).

113. Barlow, H. B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1**, 371–394 (1972).

114. David, S. V., Hayden, B. Y., Mazer, J. A. & Gallant, J. L. Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* **59**, 509–521 (2008).

115. Doetsch, G. S. Patterns in the brain: Neuronal population coding in the somatosensory system. *Physiol. Behav.* **69**, 187–201 (2000).

116. Marr, D. Vision. New York : W. H. Freeman Co (1982).

117. Bang, J. W. & Rahnev, D. Stimulus expectation alters decision criterion but not sensory signal in perceptual decision making. *Sci. Rep.* **7**, 17072 (2017).

118. Gold, J. I. & Stocker, A. A. Visual decision-making in an uncertain and dynamic world. *Annu. Rev. Vis. Sci.* **3**, 227–250 (2017).

119. Rungratsameetaweemana, N., Itthipuripat, S., Salazar, A. & Serences, J. T. Expectations do not alter early sensory processing during perceptual decision-making. *J. Neurosci.* **38**, 5632–5648 (2018).

120. Rahnev, D., Lau, H. & De Lange, F. P. Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. *J. Neurosci.* **31**, 10741–10748 (2011).

121. Manenti, G. L., Dizaji, A. S. & Schwiedrzik, C. M. Variability in training unlocks generalization in visual perceptual learning through invariant representations. *Curr. Biol.* **33**, 817–826 (2023).

122. Wenliang, L. K. & Seitz, A. R. Deep neural networks for modeling visual perceptual learning. *J. Neurosci.* **38**, 6028–6044 (2018).

123. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).

124. Zhou, L., Yang, A., Meng, M. & Zhou, K. Emerged human-like facial expression representation in a deep convolutional neural network. *Sci. Adv.* **8**, eabj4383 (2022).

125. Boutin, V., Franciosini, A., Chavane, F., Ruffier, F. & Perrinet, L. Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLoS Comput. Biol.* **17**, e1008629 (2021).

126. Brainard, D. H. & Vision, S. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436

(1997).

127. Raftery, A. E. Bayes factors and BIC: Comment on “A critique of the Bayesian information criterion for model selection”. *Sociol. Methods Res.* **27**, 411–427 (1999).

128. Schütt, H. H., Harmeling, S., Macke, J. H. & Wichmann, F. A. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vis. Res.* **122**, 105–123 (2016).

129. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv* 1312.6114 (2013).

Acknowledgements

We acknowledge the participants for their contribution to this study. This work was supported by National Natural Science Foundation of China (32271099), Research Center for Brain Cognition and Human Development of Guangdong Province, and Striving for the First-Class, Improving Weak Links and Highlighting Features (SIH) Key Discipline for Psychology in South China Normal University.

Author contributions

Conceptualization, X.Z.; Methodology, L.H., S.S., Y.S., R.Z., F.L., and X.Z.; Formal Analysis, L.H., S.S., and Y.S.; Investigation, L.H., S.S., Y.S., and S.O.; Writing – Original Draft, X.Z.; Writing – Review and Editing, L.H., S.S., Y.S., F.L., and X.Z.; Supervision, X.Z.

Competing interests

The authors declare no competing interests.