Behavioral/Cognitive

Visual Prediction Error Spreads Across Object Features in Human Visual Cortex

Jiefeng Jiang (江界峰),[1] Christopher Summerfield,[3] and Tobias Egner[1,2]

[1]Center for Cognitive Neuroscience and [2]Department of Psychology and Neuroscience, Duke University, Durham, North Carolina 27708, and [3]Department of Experimental Psychology, University of Oxford, OX1 3UD Oxford, United Kingdom

Visual cognition is thought to rely heavily on contextual expectations. Accordingly, previous studies have revealed distinct neural signatures for expected versus unexpected stimuli in visual cortex. However, it is presently unknown how the brain combines multiple concurrent stimulus expectations such as those we have for different features of a familiar object. To understand how an unexpected object feature affects the simultaneous processing of other expected feature(s), we combined human fMRI with a task that independently manipulated expectations for color and motion features of moving-dot stimuli. Behavioral data and neural signals from visual cortex were then interrogated to adjudicate between three possible ways in which prediction error (surprise) in the processing of one feature might affect the concurrent processing of another, expected feature: (1) feature processing may be independent; (2) surprise might "spread" from the unexpected to the expected feature, rendering the entire object unexpected; or (3) pairing a surprising feature with an expected feature might promote the inference that the two features are not in fact part of the same object. To formalize these rival hypotheses, we implemented them in a simple computational model of multifeature expectations. Across a range of analyses, behavior and visual neural signals consistently supported a model that assumes a mixing of prediction error signals across features: surprise in one object feature spreads to its other feature(s), thus rendering the entire object unexpected. These results reveal neurocomputational principles of multifeature expectations and indicate that objects are the unit of selection for predictive vision.

Key words: expectation; feature-based attention; object vision; prediction error

Significance Statement

We address a key question in predictive visual cognition: how does the brain combine multiple concurrent expectations for different features of a single object such as its color and motion trajectory? By combining a behavioral protocol that independently varies expectation of (and attention to) multiple object features with computational modeling and fMRI, we demonstrate that behavior and fMRI activity patterns in visual cortex are best accounted for by a model in which prediction error in one object feature spreads to other object features. These results demonstrate how predictive vision forms object-level expectations out of multiple independent features.

Introduction

To recognize its surroundings, the visual brain has to infer accurately the causes of retinal stimulation. This process is greatly complicated by the inherent ambiguity of the visual signal: depending on viewpoint, occlusion, and lighting conditions, a single object can cast a vast number of different light patterns onto the retina, whereas myriad different stimuli can produce identical patterns of stimulation. To mitigate this problem, visual cognition is thought to rely heavily on contextually informed expectations to disambiguate bottom-up stimulation (Bar, 2004; Kersten et al., 2004; Summerfield and de Lange, 2014). Accordingly, objects are recognized more quickly if they occur in a typical context (e.g., a toaster on a kitchen counter) than when they are encountered in unusual circumstances (e.g., said toaster placed on a car roof) (Palmer, 1975; Biederman et al., 1982). Similarly, conditionally less probable (i.e., unexpected) stimuli appear to require more extensive neural processing in sensory cortex than more probable (expected) ones (Summerfield et al., 2008; den Ouden et al., 2009; Alink et al., 2010; Egner et al., 2010; Meyer and Olson, 2011).

Although the central role of expectations in perceptual inference is now widely acknowledged and some of its basic implications have been successfully modeled (Spratling, 2008; Jiang et al.,

Received May 12, 2016; revised Oct. 25, 2016; accepted Oct. 29, 2016.

Author contributions: J.J., C.S., and T.E. designed research; J.J. performed research; J.J. analyzed data; J.J., C.S., and T.E. wrote the paper.

This work was supported by the National Institute of Mental Health–National Institutes of Health (Grant R01MH097965 to T.E.). We thank Nadia Brashier for help with data acquisition.

The authors declare no competing financial interests.

Correspondence should be addressed to Jiefeng Jiang, Center for Cognitive Neuroscience, Duke University, P.O. Box 90999, Durham, NC 27708. E-mail: Jiefeng.jiang@duke.edu.

DOI:10.1523/JNEUROSCI.1546-16.2016

 0270-6474/16/3612746-18$15.00/0

2012; Wacongne et al., 2012), one particularly notable shortcoming is that we do not know how the visual brain manages multiple, simultaneous expectations for different features of an object such as its color, shape, and size. Prior studies have used only simple, one-dimensional scenarios in which predictions and surprise signals were limited to a single feature of a given object or object category (e.g., the forthcoming stimulus likely being a face, or a right-tilted Gabor patch; Egner et al., 2010; Kok et al., 2012a). In the real world, however, object expectations are rarely limited to a single feature. For instance, a soccer player must form expectations about both the motion of surrounding players and the color of their jerseys, to distinguish trajectories of teammates from those of opponents. Therefore, we typically acquire, and make use of, concurrent expectations about multiple features of an object. Importantly, this can give rise to circumstances in which one feature conforms to expectations, but another feature does not. A key unresolved question is thus how the brain resolves conflict between inconsistent feature expectations to produce unified object-level perception.

In the present study, we investigated how the processing of one stimulus feature (e.g., player motion) is affected by the violation of expectations concerning another feature (e.g., jersey color) of the same stimulus. To understand this core aspect of visual object cognition, we used behavioral and fMRI data to adjudicate between three rival hypotheses: First, the two feature expectations might operate independently of each other such that an expectation violation of one feature would not affect the processing of the other feature ("independence model"). Second, perceptual expectations may operate at an object level such that one surprising feature might render the entire object (including the expected feature) surprising ("reconciliation model"). A parallel to this scenario exists in the attention literature, in which attending to one feature (or part) of an object can lead to the attentional selection of the entire object (Egly et al., 1994; O'Craven et al., 1999). Third, the cooccurrence of an expected and an unexpected object feature might motivate the perceptual hypothesis that the two features are not in fact part of the same object ("segregation model"). This hypothesis echoes findings in figure–ground segmentation, in which subjects tend to interpret a single unusual shape as reflecting a collection of mutually occluding, common shapes (for review, see Wagemans et al., 2012). Finally, we investigated whether, and in what manner, a surprising feature affecting the processing of an expected feature could plausibly interact with feature-based attention (i.e., the feature's relevance to the current task; Summerfield and Egner, 2009). Our models therefore also incorporated effects of feature-based attention.

Materials and Methods

Design and rationale

Our goal was to determine how the visual brain processes expectations for multiple features of a single object as a function of whether a given feature is attended. We operationalized this problem with a perceptual categorization task involving a stimulus (a coherent motion field of dots) composed of two independently varying features: color and motion direction (see Figs. 1*A*,2*A*). Both of these features are known to drive neural responses in early visual cortex (EVC; Movshon and Newsome, 1996; Engel et al., 1997; Johnson et al., 2001; Kamitani and Tong, 2006), but are thereafter processed by specialized areas of the ventral (color: V4; Gegenfurtner, 2003) and dorsal (motion: area MT+; Born and Bradley, 2005) visual streams.

This provides an ideal scenario for testing how an expectation (or violations thereof) for one stimulus feature affects the processing of another feature of the same object both in feature-selective regions (i.e., V4 and MT+) and in regions sensitive to both of these features (i.e., EVC). To this end, we independently manipulated whether a given feature conformed to or violated perceptual expectations. These manipulations produced four experimental conditions: color-unexpected/motion-unexpected (CU/MU), color-expected/motion-expected (CE/ME), color-unexpected/motion-expected (CU/ME), and color-expected/motion-unexpected (CE/MU). Therefore, the expectation status across the two features is consistent in the CU/MU and CE/ME conditions, but inconsistent in the CU/ME and CE/MU conditions. To assess how multifeature expectations interact with attention and to dissociate expectation effects from attentional effects, we furthermore independently varied the task relevance of the two feature dimensions (attend to color vs attend to motion).

Using this experimental design, we compared three types of predictive coding models concerning how expectation and surprise interact between object features to produce unified object perception. This interaction relies on cross-feature exchange of prediction error (PE), which drives the updating of neural representation to match sensory input. Specifically, a parameter, β, is used to determine the proportion of PE that propagates from one feature stream to the other (see Materials and Methods: Computational simulation). When expectations are consistent across features (i.e., CE/ME and CU/MU conditions), the PEs are identical for both features (either both are low or both are high) such that any PE mixing across feature streams is balanced: the same amount of color PE would propagate to the motion stream as the other way around. Therefore, PE mixing does not alter feature processing in these conditions. Crucially, however, when expectations are inconsistent between features (i.e., CU/ME and CE/MU conditions), PE mixing affects the feature stream cross talk in different ways depending on the sign of β. (The absolute value of β does not qualitatively change the pattern of the interaction; see Fig. 7).

Setting β to 0 simulates the "independence model" in which no PE mixing occurs (see Fig. 3*A*), so PE in one feature exerts no influence on the processing of the other feature (e.g., violation of the expectation of a player's jersey color does not affect the processing of his or her motion). In contrast, setting β to a positive value simulates the "reconciliation model" (see Fig. 3*B*), which reduces the discrepancy of PE between the expected and the unexpected features by dampening PE in the unexpected feature and augmenting PE in the expected feature. Here, expectations for multiple features of a single object are effectively blended into an object-level expectation. For example, violation of the expectation of a player's jersey color—even in the presence of an expected motion direction—would make the perception of the player per se unexpected. The reconciliation model makes the following specific predictions. First, the positive β ensures that PE from one feature affects information processing in both features in the same direction (i.e., surprise in one stream enhances surprise in the other stream), which results in a reduced discrepancy between PEs across the two features. Second, this decreases the expectation effect (i.e., the discrepancy between unexpected and expected conditions; see Fig. 3*B*) in expectation-inconsistent conditions, thus making CU/ME and CE/MU less distinct from each other compared with expectation-consistent conditions (see Fig. 3*E*). And third, this type of PE mixing makes the unexpected feature less unexpected and the expected feature less expected (see Fig. 3*B*). Therefore, the PE mixing would interfere with within-feature information processing, making the neural representations of features in expectation-inconsistent trials weaker than in expectation-consistent trials.

Conversely, setting β to a negative value simulates the "segregation model" in which the unexpected feature sends PE to the expected feature stream to drive its processing in the opposite direction while enhancing its own PE to boost within-feature processing (see Fig. 3*C*). In other words, the segregation model resolves clashing expectations between features by discarding the premise that the features belong to the same object and producing segregated and enhanced perceptions for each feature instead. Observing an expected motion trajectory paired with an unexpected jersey color would result in an updated belief that the jersey color and object motion are caused by two different players. Compared with the reconciliation model, the reversed sign of β in the segregation

Table 1. Summary of key predictions from three different models of multifeature expectations in visual object cognition

	How PE in feature A affects feature B	Distinction between EI and EC	Representation strength
Independence model	No effect	EI = EC	EI = EC
Reconciliation model	Same direction as feature A	EI < EC	EI < EC
Segregation model	Opposite direction to feature A	EI > EC	EI > EC

EI, Expectation inconsistent conditions; EC, expectation consistent conditions.

model thus leads to the exact opposite predictions. All model predictions are summarized in Table 1.

We adjudicated between the three rival models using behavioral and neuroimaging data from the following two experiments. Note that all the model predictions concern differences in neural representations or PE between conditions. The key goal of our fMRI analyses was to quantify these distinctions. To this end, we adopted multivoxel pattern analysis (MVPA) as our hypothesis testing tool because MVPA measures how separable the neural activity patterns of different conditions are and the resulting classification accuracy is a natural quantification of condition separability. The rationale for focusing on MVPA (rather than GLM) results was also driven by additional considerations stemming from the predictive coding framework that underlies our models (see below). This framework assumes that computational units involved in producing expectations and PE are located in close spatial proximity (Bastos et al., 2012). Given random sampling of such units across fMRI voxels, previous studies have found spatially intermingled voxels with signals that were either primarily driven by expectation or PE signals (de Gardelle et al., 2013). This implies that mean regional BOLD signals derived from conventional univariate analysis with spatial smoothing blend together expectation and surprise signals (Egner et al., 2010) and therefore have limited sensitivity for distinguishing different expectation conditions (see also Kok et al., 2012a). In contrast, MVPA treats each voxel independently and is capable of exploiting heterogeneous response profile in adjacent voxels to distinguish activity patterns of different experimental conditions. For example, given two intermingled groups of voxels, one showing A > B activity and the other showing B > A activity, averaging across (e.g., smoothing) these voxels may cancel out any difference between these conditions, but MVPA can assign positive and negative weights to these two groups to "align" their opposite patterns of activity to distinguish between the A and B conditions.

Experiment 1 (behavior)

Subjects. Seventeen volunteers (11 females, 19–54 years old, mean age = 27 years, one left-handed) gave informed consent in accordance with institutional guidelines and completed this experiment. All subjects had normal or corrected-to-normal vision. This study was approved by the Duke University Health System Institutional Review Board.

Stimuli. The presentation of stimuli and response recording were controlled using Psychtoolbox version 3 (Brainard, 1997). The auditory stimuli were composed of four tones. Each tone consisted of four notes (200 ms each) that were ordered to produce either a rise or fall in pitch. Therefore, the rising and falling tones did not differ in the notes used, only in the way the notes were ordered. In addition, the tones were played in two distinct timbres, resulting in a two (rising/falling pitch) × two (timbres) factorial design. These auditory stimuli were delivered via noise-canceling headphones.

The visual stimuli consisted of clouds of colored (either red or green) moving (either up or down, 100% coherence) dots presented at the center of the screen against a gray background (duration = 1 s). The luminance of the dots and the background were identical. The moving dots display spanned ~6° of visual angle both vertically and horizontally and consisted of 200 dots of ~0.12° radius. The motion speed of each dot was drawn randomly from a uniform distribution from 13°/s to 15°/s. The visual stimuli were presented on a 17 inch LCD display at 60 Hz. The responses were recorded using a standard keyboard.

Procedure. Each trial started with the presentation of the auditory cue tone, which was followed by the moving dots display (see Fig. 1*A*). Therefore, the cue and stimulus processing did not overlap in sensory modality. The cue's timbre and pitch were predictive of the forthcoming dots' color and motion direction at 75% validity, respectively. To avoid potentially confusing violations in contingency, up/down motion was always predicted by rising/falling tones, respectively. For each trial, the participants were asked to identify the color or motion direction of the dots with button presses. The target feature (color or motion) was cued via written instruction (see below). The manipulation of target feature served the function of directing feature-based attention to either color or motion. Trials were separated by an intertrial interval (ITI) of 1.5 s.

Participants first went through a training and practice phase to learn the auditory cue-dots associations and task requirements: they first performed a training session of 20 trials (five trials for each tone) of 100% validity to promote learning. Participants were then asked to explicitly indicate the predicted color and motion direction of the dots for each cue tone. These training and test sessions repeated until the participants reached 100% correct rate in the test session. Then, the concurrent expectations (i.e., "the rising/falling of the pitch predicts the motion direction, and the timbre predicts the color") were further explained explicitly to the participant by the experimenter to reinforce the learned associations. Next, two practice sessions (one for each attention condition) of 20 trials each with the predictive validity of 75% were administrated to ensure that the participants comprehended the task instructions before performing the main task.

The main task consisted of six runs (three for each attention condition in an ABABAB order, with the attention condition in the first run counterbalanced across subjects) of 64 trials each. At the beginning of each run, an instructional cue was shown to specify the target feature (color or motion) that the subjects were to discriminate via a button press on each trial. The response mapping was displayed at the bottom of the screen throughout each run and counterbalanced across subjects. The numbers of presentations for each tone × color–motion combination were equated within each run and each condition of the factorial design to avoid bias in the analyses.

Analysis. The accuracy for each condition in the two (feature attention) × two (color expectation) × two (motion expectation) factorial design was calculated and entered into a repeated-measures three-way ANOVA. The same analysis was performed on response time (RT) means after excluding RTs from error trials or outlier trials (i.e., trials with RTs outside of the range of grand mean ± 2.5 SD).

Experiment 2 (fMRI)

Subjects. Twenty-five right-handed volunteers gave informed consent in accordance with institutional guidelines and completed this experiment. All subjects had normal or corrected-to-normal vision. Two subjects were excluded from further analysis due to excessive head movement during scanning (movement >6 mm or 6° within any run). The final sample consisted of 23 subjects (14 females, 22–35 years old, mean age = 27 years). This study was approved by the Duke University Health System Institutional Review Board.

Stimuli. The presentation of stimuli and response recording were accomplished using Psychtoolbox version 3. The auditory stimuli were identical to Experiment 1 and were delivered via MR-compatible, noise-canceling headphones. The visual stimuli were the same as Experiment 1 except with additional colors of blue and yellow (with equal luminance to the background) and additional motion directions of left and right sampled from the same uniform distribution of speed as in Experiment 1. The visual stimuli were presented on a back projection screen viewed via a mirror attached to the scanner head coil. The responses were recorded using two MR-compatible button boxes (one for each hand).

Procedure. The training, test, and practice sessions were identical to Experiment 1. The main task consisted of eight runs (in the order of ABABBABA, with the first run counterbalanced across subjects) of 64 trials each, with exponentially jittered ITIs (from 4 to 6 s with a step size of 500 ms). Different from Experiment 1, the goal of this task was to identify occasional changes in color/motion via button press. The target feature (e.g., color) was cued at the beginning of each run. The subjects were also explicitly informed that no change would occur in the nontarget feature to encourage the subjects to direct attention solely to the target feature. Therefore, similar to Experiment 1, this experimental design

resulted in a two (feature attention) × two (color expectation) × two (motion expectation) factorial design.

To manipulate feature-based attention and to keep subjects on task, eight trials (12.5%) per run were randomly selected as "change trials" (or target trials), in which the target feature (color/motion) changed to yellow or blue/left or right (at 50% probability) after 500 ms (see Fig. 2*A*), which had to be reported by the subjects based on a response mapping displayed at the bottom of the screen throughout each run. However, fMRI analysis only included the frequent nontarget trials to avoid confounds from motor responses or target-related processing (Summerfield et al., 2008). The auditory cues had no predictive value regarding the postchange color/motion in change trials. Nevertheless, in no-change trials, the expectation effects were still mediated by the auditory cues that preceded each dot cloud. The numbers of presentations for each tone × color–motion combination were equated within the no-change trials for each run and each condition of the factorial design to avoid bias in the analyses.

Behavioral data analysis. The accuracy in change trials and false alarm rate in no-change trials were calculated for each subject to give a descriptive assessment of task performance.

Image acquisition and preprocessing. Images were acquired parallel to the AC–PC line on a 3 T GE scanner. Structural images were scanned using a T1-weighted SPGR axial scan sequence (146 slices, slice thickness = 1 mm, TR = 8.124 ms, FoV = 256 mm * 256 mm, in-plane resolution = 1 mm * 1 mm). Functional images were scanned using a T2*-weighted single-shot gradient EPI sequence of 42 contiguous axial slices (slice thickness = 3 mm, TR = 2 s, TE = 28 ms, flip angle = 90 °, FoV = 192 mm * 192 mm, in-plane resolution = 3 mm * 3 mm). Functional data were acquired in eight runs of 206 images each. Preprocessing was done using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). After discarding the first five scans of each run, the remaining images underwent spatial realignment, slice-time correction, and spatial normalization, resulting in normalized functional images in their native resolution. As is customary in MVPA, no spatial smoothing was applied to the normalized fMRI images.

MVPA procedures. For each subject and each experimental condition in the factorial design (attention × color expectation × motion expectation), we generated an activation map that encodes the *t*-value of this condition at every gray matter (GM) voxel. Specifically, the normalized images were regressed against a general linear model (GLM) to estimate activation levels for each experimental condition. The GLM consisted of nine event-based regressors (convolved with SPM8's canonical hemodynamic response function) representing the onsets of no-change trials in each of the eight conditions of the factorial design, the onsets of change trials, and nuisance regressors representing head motion parameters, as well as the grand mean of the run (to remove the run-specific baseline signal and activity elicited by the response mapping instructions that were presented throughout each run). Note that the specific stimuli (e.g., red color, downward motion) were counterbalanced and collapsed within each cell of the design because we were interested in classifying neural patterns that distinguished the processing of different feature dimensions (i.e., color vs motion) rather than different intradimensional exemplars (e.g., red vs green). This approach applied to both expectation- and attention-based classifiers. In other words, within each cell of the factorial design, the presented color and motion stimuli belonged to the same attention and expectation conditions to enable the tests of generic (i.e., not specific to particular colors and motions) attention and expectation effects. This GLM also controlled for the unequal trial counts between expected and unexpected conditions because all trials within a particular condition were grouped into one regressor such that expected and unexpected conditions were represented by an equal number of regressors (or data points) for the MVPAs. As a result, for each subject and each experimental condition in the factorial design (attention × color expectation × motion expectation), this step generated an activation map that encodes the *t*-value of this condition at every GM voxel defined in the segmented SPM T1 template (dilated by one voxel). For each subject, activation estimates were further normalized within voxels and across the eight conditions to remove individual difference in baseline activation level and absolute amplitude of activations.

The MVPA was performed in a searchlight-based (Kriegeskorte et al., 2006), intersubject manner using a leave-one-out (LOO) cross-validation approach: the classifiers were trained on the data from 22 subjects and tested on the data from the remaining subject. The training and testing iterated until each subject served once as test subject. This LOO cross-validation procedure was applied to all classifiers. According to the predictive coding framework (see below), the effects of attention and expectation in one region (or level) mainly originate from the next lower or higher level in the processing hierarchy. Given the relatively small size of the searchlights (2 voxel radius up to 33 voxels in volume) in the MVPA, we did not expect one searchlight to cover more than one region modeled in the computational framework (e.g., EVC, MT+, and v4). Therefore, we used linear support vector machines, which assume no intervoxel interaction of fMRI activity within searchlights (Pereira et al., 2009), to quantify the differentiation of neural activity patterns between experimental conditions. The size of the searchlight, along with the box constraint of the linear support vector machine (1, also the default value in Matlab), are the same as in an earlier study investigating expectation and attention effects for single stimulus features (Jiang et al., 2013) and produced comparable results. Note that we did not remove the searchlight mean activity level before MVPA, so the MVPA did not make any assumptions about whether the signals of two experimental conditions diverge along a single dimension (i.e., a univariate difference in the average amplitude of the BOLD signal across a region) or multiple dimensions (i.e., a difference in the relative multivoxel pattern of activity evoked between conditions).

We took this cross-subject approach based on three considerations. First, this approach places the strong constraint on our findings that the mixture of computations driving the BOLD signal (while unknown) must be consistent (generalizable) across subjects at the voxel level after anatomical normalization, which is also the assumption of the widely used univariate fMRI analysis. This constraint is crucial in the present work because it focuses on the early visual cortex, one of the regions with the smallest degree of anatomical and functional individual differences in the cerebrum. Compared with within-subject MVPA, the assumptions underlying group results for cross-subject MVPA are in fact more similar to the standard mass-univariate analysis group results in that cross-subject MVPA requires the effects of interest to be in the same direction across subjects. Previous cross-subject MVPA studies have demonstrated this consistency by successfully decoding complex cognitive states such as task state (Mourão-Miranda et al., 2005; Poldrack et al., 2009), lying or telling the truth (Davatzikos et al., 2005), the ambiguity of a presented sentence (Mitchell et al., 2004), receiving monetary or social reward (Clithero et al., 2011), presence/absence of conflict in cognitive control (Jiang et al., 2015), experiencing pain (Gordon et al., 2014), fear conditioning (Onat and Buchel, 2015), and observing people touching different objects (Kaplan and Meyer, 2012). In visual cortex, a number of studies have demonstrated that, after standard anatomical alignment, high cross-subject MVPA accuracy can be achieved in the decoding of visual content (Haxby et al., 2011; Shinkareva et al. (2008); Shinkareva et al. (2011)). Of direct relevance to the current study, it has also been shown previously that this cross-subject generalizability held for the effects of different attention and expectation conditions on visual cortex signal (Jiang et al., 2013). Second, the current design, due to the importance of concurrently manipulating expectations in two features, necessitated the creation of some rare event conditions, namely the low probability events of CU/MU trials (16 trials/subject). This low trial count creates suboptimal conditions for running within-subject MVPA, a statistical power problem that can be countered by using the cross-subject MVPA approach that includes trials from all subjects to increase the trial count (to 23 subjects × 16 trials/subject) for the CU/MU condition in the MVPAs. As shown in Figures 4*F* and 6*C*, analyses involving CU/MU trials did in fact reveal significantly above-chance classification accuracies, suggesting that the chosen cross-subject MVPA approach was not hampered by low trial counts (high variance) in this condition. Third, the cross-subject approach allowed us to control for a potential confound introduced by specific response mappings because the mappings were counterbalanced across subjects.

To test the effects of the manipulation of feature-based attention and expectation, we built classifiers discriminating fMRI activity patterns of no-change trials between color and motion target runs (see Fig. 2*B*), CE and CU conditions (see Fig. 2*C*), and ME and MU conditions (see Fig. 2*D*), respectively. Furthermore, in conjunction with behavioral analyses in Experiment 1 (see Fig. 1*C*), we constructed expectation classifiers (i.e., expected vs unexpected) for the attended feature (see Fig. 2*E*) and unattended feature (see Fig. 2*F*), respectively, to further examine the interaction between attention and expectation. Moreover, to test how expectation (or violation thereof) of one feature affects the expectation of the other feature, we followed the model predictions in Figure 3, *A–F*, and Table 1 and compared the performance of two fMRI activity pattern classifiers: one that discriminated between CU/MU and CE/ME trials and another one that discriminated between CU/ME and CE/MU trials (see Fig. 4 *D*, *G*). Finally, to test whether/how the effect of attention varies as a function of concurrent expectation of color and motion, we constructed fMRI activity pattern classifiers between the attend-color and attend-motion conditions separately for each of the four color expectation × motion expectation conditions and tested whether classifier performance varies as a function of expectation conditions (see Fig. 6*A–C*).

As a result, for each classifier, a group-level classification accuracy map was computed in which each GM voxel represented the classification accuracy from the LOO cross-validation of the searchlight centered at that voxel. For each searchlight, the statistical significance of its performance was gauged using a binomial test. The difference of classification performance between two maps was compared using a Bayesian approach. This approach inferred the probability that two classification accuracies observed from the same searchlight over two different accuracy maps belonged to the same underlying classification accuracy (for details, see Jiang et al., 2013; Jiang et al., 2015).

Statistical analysis and control for false positives. For all aforementioned statistical analyses, false positives due to multiple comparisons were controlled for at $p < 0.05$ (for classification analyses, the *p*-values were obtained using binomial tests for each searchlight or ROI) for combined searchlight classification accuracy and cluster extent thresholds using the AFNI ClusterSim algorithm (http://afni.nimh.nih.gov/pub/dist/doc/program_help/3dClustSim.html). Ten thousand Monte Carlo simulations determined that an uncorrected voxelwise *p*-value threshold of <0.01 (for *p*-value transformed from binomial distribution, the largest *p*-value that was <0.01) in combination with a searchlight cluster size 21–32 searchlights (depending on the specific analysis) ensured a false-positive rate of <0.05.

Computational simulation

Computational modeling. To enable quantitative and formal predictions about responses under predictive coding framework, this study introduces a particular predictive coding scheme that was used to simulate perceptual inference under the three hypotheses above. This allowed us to simulate particular response profiles that we then tested for using behavioral reports and multivariate analysis of physiological responses. To this end, the aforementioned three rival models were implemented using a biologically feasible predictive coding model (Friston, 2005; Fig. 3*G*), which posits a continual interplay across the visual cortical hierarchy between the top-down passing of predictions concerning forthcoming inputs and the bottom-up passing of PE (Mumford, 1992; Rao and Ballard, 1999; Friston, 2005, 2010). Predictive coding models have been demonstrated to account for many empirical findings in the visual cognition literature (for review, see Summerfield and de Lange, 2014). To simulate the processing of the two features of color and motion, the model consists of two "visual streams" specialized in processing either feature (see Fig. 3*G*). The model streams comprise four levels: an input stage (level 0), followed by an EVC stage (level 1) that is sensitive to both color and motion direction, followed by higher-level, feature-selective visual cortex (level 2) that are sensitive to either color (i.e., V4) or motion direction (i.e., MT+), and finally, putative higher-level regions (level 3) that provide expectation inputs to the simulated lower level regions. Consistent with the tenets of predictive coding (Friston, 2005), each level consists of two types of computational units (except for the top level): "representation units" that encode predictions of bottom-up inputs and "error units" that receive top-down input from representation units at the next-higher level, calculate PE (i.e., the discrepancy between predicted and actual input), and pass that error back to the representation units at the next-higher level. The co-occurrence of predictive and surprise signals in visual cortex has been confirmed in previous studies (Egner et al., 2010; Keller et al., 2012; de Gardelle et al., 2013).

In this study, perception is considered as an inference process that integrates prior expectations with actual visual input and is thus implemented using a delta rule, which approximates the performance of the optimal (Bayesian) inference algorithm for our task with reduced running time (Nassar et al., 2010; Nassar et al., 2012). Within each level of the model, the error units' computation of PE guides the adjustment of prediction in representation units. This process is iterated until a stable state (i.e., a stable interpretation of the current visual input) is reached. In this model, representation and error units at level i of stream s ($s = 0$ and 1 for color and motion stream, respectively) are denoted by r_i^s and e_i^s, respectively. For simplicity, at each level of each stream, only one representation unit and one error unit were simulated.

To incorporate effects of attention into the model, we furthermore allowed feature relevance to impose a multiplicative gain on visual processing (Martinez-Trujillo and Treue, 2004) by an attentional factor a^s. In the framework of predictive coding, attention is modeled as the precision or confidence of the prediction errors (Feldman and Friston, 2010; Auksztulewicz and Friston, 2015; Kanai et al., 2015), where more attention equates to enhanced PE input forwarded to the next level. This assumption can successfully account for findings from behavioral cued attention studies (Feldman and Friston, 2010). Attentional sharpening of PE signals has also been documented at the level of fMRI signal in ventral visual cortex (Jiang et al., 2013). Attention/confidence-modulated PE can also be interpreted as a mathematical formulation of surprise that consists of two levels of uncertainty, namely the (violation) of prediction, and the confidence of this prediction (Yu and Dayan, 2005). This factor also simulates attentional modulation on representation units (Rao, 2005; Spratling, 2008).

We did not include an additive attentional gain (Thiele et al., 2009) because it would be canceled out when producing predictions for the empirical analyses, all of which compared the simulated activity between two conditions. Furthermore, we did not model an attention-induced shift of contrast-response function (Reynolds et al., 2000) because: (1) the stimuli used in the experiments had 100% coherence in both color and motion direction and thus had high contrast; (2) we only analyzed no-change trials, so there was no contrast due to change of features; and (3) our manipulation of attention did not direct the participants to any particular color or motion direction and provided no information for tuning the contrast-response function for a specific color or motion direction. To sum up, at any moment t, $e_i^s(t)$ was defined as follows:

$$e_i^s(t) = a^s \times (r_i^s(t) - \theta_i^s(t) r_{t+1}^s(t)) \quad (1)$$

Where a^s was higher in attended than unattended streams. For example, in a color detection change run, $a^0 > a^1$. We modeled attentional gain in both attended and unattended features because it has been reported that attention can also spread from attended features to other features of the same object (O'Craven et al., 1999). a^s set to 1 and 0.75 for attended and unattended conditions, respectively.

θ_i^s modulates the strength of expectation imposed by the next higher level and varied after Hebbian learning between e_i^s and r_{t+1}^s (Friston, 2005):

$$\frac{d\theta_i^s(t)}{dt} = e_i^s(t) r_{t+1}^s(t) \quad (2)$$

Similarly, the modulation of a^s on r_{t+1}^s was further implemented by applying a^s to the input; for example, $r_{0+s}^s = a^s u$ where u was the visual input, which remained constant during simulation. The noninput representation units were updated in the following manner:

$$\frac{dr_i^s(t)}{dt} = e_{t-1}^s(t) - e_i^s(t) \quad (3)$$

Therefore, updating of r_i^s was also modulated by a^s through the prediction errors. $e_3^s(t)$ was a constant of 0 due to the fact that level 3 had no error unit. In sum, attention and expectation were modeled separately using a^s and θ_i^s, respectively.

Crucially, the aforementioned crosstalk between the two stimulus features was modeled in EVC, which is sensitive to both motion and color. To introduce the effect of object-level perception on the processing of individual features, the above predictive coding model was extended to accommodate the belief that individual features were generated from the same object. Specifically, at each time point t, the updating of $r_1^s(t)$ is further modulated by this belief using the aforementioned parameter β and a mechanism that allowed for the "mixing" of the inputs from level 0 to level 1 across streams (see Fig. 3*G*, blue links) to mediate the updating of r_1^s in the following manner:

$$\frac{dr_1^s(t)}{dt} = e_0^s(t) - \beta \times e_0^s(t) + a^s \times \beta \times e_0^{1-s}(t) - e_1^s(t) \quad (4)$$

Where a^s was applied to the PE from the other stream to reflect the attentional modulation on the PE at the recipient stream. Therefore, $e_0^s(t) - a^s \times \beta \times e_0^s(t) + \beta \times e_0^{1-s}(t)$ represented a mixed PE from level 0. Specifically, when β is 0 (representing a neutral belief regarding whether the color and motion are from the same objects or not), (4) is identical to (3) to simulate the independence model. When $\beta > 0$ (representing the belief that the color and motion are from the same object), the updating of color and motion expectations are "synchronized" using β to facilitate an object-level expectation, as hypothesized in the reconciliation model. Last, when $\beta < 0$ (representing the belief that the color and motion come from different objects), the mixed PE differentiates and enhances the updating of expectations for individual features and therefore simulates the competition model. This model has two free parameters, namely the attentional modulator for the unattended stream (a) and β, which models modulation that is spread over attended and unattended streams.

Because the training and practice phases ensured that the subjects had learned the experimental manipulation of color and motion expectation, we did not model the learning effects of u and r_3^s during the simulation of the two experiments. To simulate the two tasks in this study, r_i^s ranged from -1 (completely tuned to represent the unexpected stimulus) to 1 (completely tuned to represent the expected stimulus), with 0 reflecting neutral selectivity. The absolute value of representation unit activity also represents the encoding strength of the observed features (e.g., an activity level of -0.8 represents a stronger neural representations of the observed unexpected feature than an activity level of -0.5). Accordingly, e_i^s ranged from -2 to 2. $u = 0$ when the cue was presented and 1 and -1 when the visual stimulus was expected or unexpected, respectively. $e_i^s(t) = 0.5$ (reflecting the 75% validity) during the presentation of the auditory cue to induce a top-down expectation of forthcoming visual stimuli. During the presentation of the visual stimuli, $e_i^s(t)$ changed based on (3) to reconcile the PE. The aforementioned parameter settings were applied to all three models. The only parameter that varied across models was β, which was set to 0, 0.3, and -0.3 for the independence model (no mixing of PE), the reconciliation model (mixing of PE), and the segregation model (enhancing PE within each feature), respectively. This ensured that the bias due to different model implementation details in model comparison was minimized. Therefore, the different model predictions can only be attributed to β or how the two features exchange prediction errors. The simulation results are robust to perturbation of model parameters (see Fig. 7) such that the magnitudes of a and β do not qualitatively change the pattern of simulation results. Consistent with the cross-subject MVPA approach that produced group-level results, we did not fit parameters to individual subjects. Instead, each model was run one time using the aforementioned parameters to simulate group-level results. The Matlab implementation of this framework and raw simulation results are available on request.

Simulation procedure. This $2 \times 2 \times 2$ factorial design was simulated using each of the three models. Because no randomness was introduced in the models, only one trial was simulated for each condition. Within each trial, the auditory cue was simulated for 200 time steps and the moving dots were simulated for 600 time steps to ensure that a steady state was reached (e.g., e_s^i converges to a minimum PE) to reflect that the subjects had learned the manipulations of expectation before the simulated tasks. The activity of r_i^s was estimated as its mean activity level over the last 10 time steps of the simulation to simulate the strength of representation.

Results

Experiment 1

We began by conducting a behavioral experiment that allowed us to establish how multifeature expectation interacts with attention and to adjudicate between rival model predictions of behavioral performance patterns. For the latter purpose, we simulated the task and used the model's neural activation estimates from the visual area sensitive to the attended visual feature (i.e., level 2 of the attended stream, see Materials and Methods, "Computational simulation," for details) as an index of RT. Consistent with empirical data, we treat greater simulated neural activity in category-selective visual cortex as reflective of stronger sensory evidence and thus faster RT (Ratcliff and Rouder, 1998).

Model predictions

All three models predicted that confirmed expectation in the relevant (attended) feature would facilitate performance (Fig. 1*B*). Crucially, the models' predictions diverged on the effect of expectation of the unattended feature on behavior. Specifically, the independence model predicted no effect, the reconciliation model predicted a positive effect (i.e., activity: expected > unexpected, and RT: expected < unexpected), and the segregation model predicted a negative effect due to their different assumptions of how PE in one feature affects the other feature (Table 1).

Behavioral data

To arbitrate among the models, we compared their predictions with the RT patterns of human participants judging expected versus unexpected attended features (collapsed across target feature). Using a 2-way ANOVA (feature: attended/unattended × expectation: expected/unexpected), we observed significant main effects of both attention ($F_{(1,16)} = 38.68$, $p < 0.001$; attended: 479 ± 24 ms, unattended: 514 ± 24 ms) and expectation ($F_{(1,16)} = 7.85$, $P = 0.01$; expected: 491 ± 24 ms, unexpected: 501 ± 24 ms). *Post hoc* analyses revealed a significant gain of expectation (i.e., responses on expected trials were faster than on unexpected trials) on the attended feature (34 ± 5 ms, $t_{(16)} = 6.52$, $p < 0.001$, one-sample t test; Fig. 1*C*). This finding was consistent with all three models' predictions (Fig. 1*B*).

Crucially, we also observed a significant expectation gain effect in the unattended feature (10 ± 3 ms, $t_{(16)} = 2.93$, $p = 0.01$; Fig. 1*C*). This finding exclusively supports the reconciliation model (cf. Fig. 1*B*), which assumes that surprise in one feature "spreads" to the other feature. This behavioral effect also rules out the possibility that only the attended feature expectations drove subjects' performance (which would predict no expectation gain in the unattended feature). Performing the corresponding ANOVA on the accuracy of motion/color categorization (Fig. 1*D*) replicated the main effect of attention ($F_{(1,16)} = 8.14$, $p = 0.01$), which was driven by more accurate responses when the color was attended (0.936 ± 0.008) than unattended (0.893 ± 0.021). The effect of expectation on the unattended feature was not observed in accuracy (-0.005 ± 0.007, n.s)., implying that the improved RT in expected conditions was not due to a speed–accuracy trade-off.

These results clearly demonstrate that the experimental manipulations successfully induced concurrent color and motion

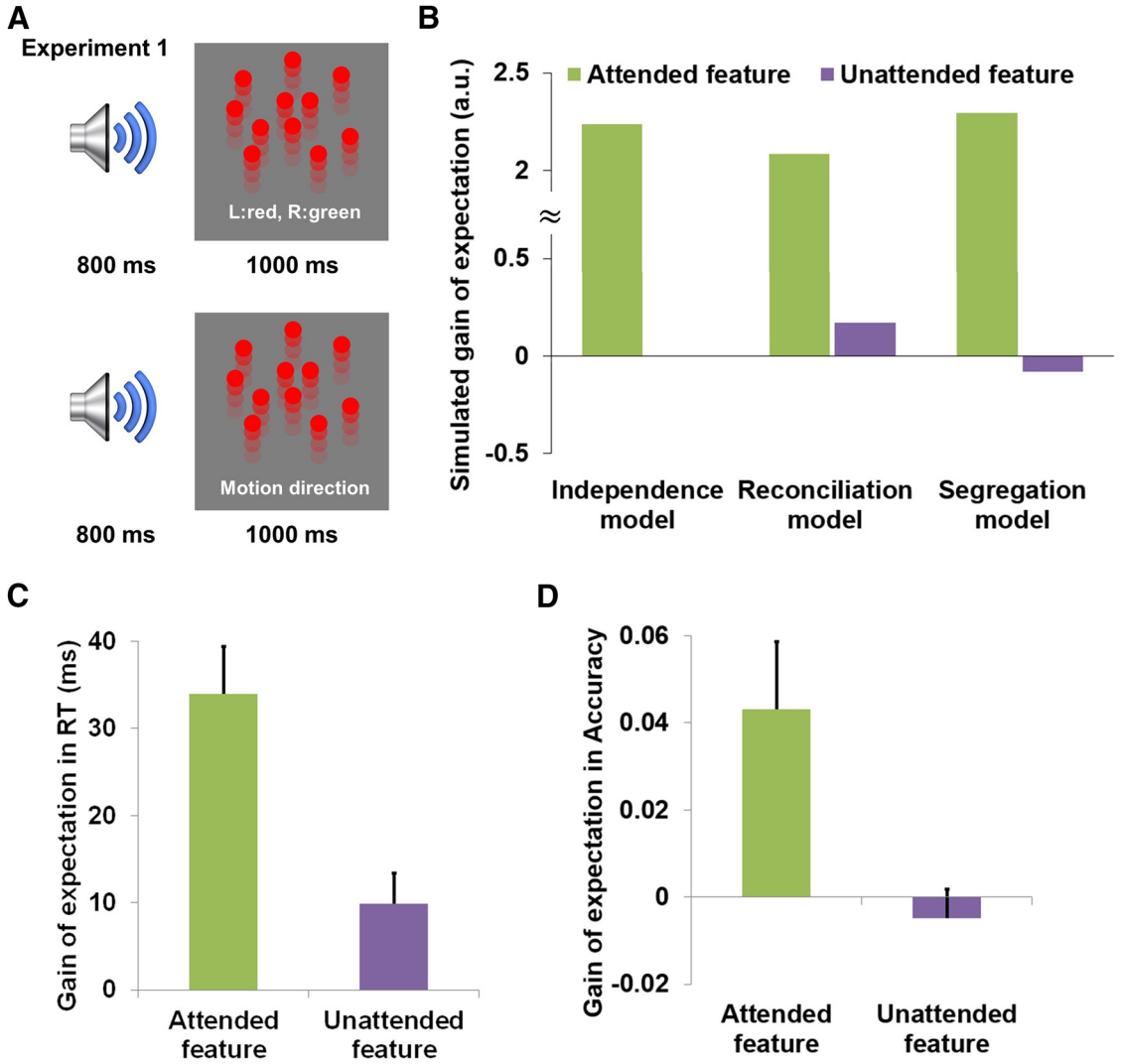

Figure 1. Experiment 1 task, model predictions, and behavioral results. ***A***, Two example trials in Experiment 1. Note that the number and size of the dots differ from the actual experimental displays for illustrative purposes. The top/bottom example trial requires a participant to respond to the color/motion direction of the dots. ***B***, Model predictions of the effects of expectation on the attended and the unattended features. ***C***, ***D***, Group mean and MSE of the gain of expectation (i.e., improved performance for expected > unexpected features if value on *y*-axis is positive) in RT (***C***) and accuracy (***D***) in Experiment 1 plotted as a function of whether the feature in question was attended (i.e., was the current target feature).

expectations in the participants. Moreover, the behavioral data were best accounted for by the reconciliation model with cross-feature blending of PEs.

Experiment 2

We next sought to investigate how multiple feature expectations and attention interact to shape neural stimulus representations in the visual system, allowing us to further adjudicate between predictions of the three rival models. Subjects first learned the aforementioned concurrent expectation cues in a training session and then performed a visual change detection task during simultaneous fMRI scanning (see Materials and Methods, "Experiment 2"). As expected, subjects correctly indicated the changed color or motion direction on target trials with high accuracy (mean accuracy = 0.947 ± 0.012) and committed few false alarms (mean false alarm rate = 0.006 ± 0.002) in nontarget trials. In addition, participants were more accurate in motion change runs (mean accuracy = 0.975 ± 0.024) than color change runs (mean accuracy = 0.919 ± 0.010, $t_{(16)} = 3.08$, $p = 0.006$), possibly due to a more intuitive response mapping in the former (e.g., left key = dots moving left) than the latter (e.g., left key = yellow). These findings document that the participants followed instructions and were focused on the task, thus providing a solid basis for interpreting the fMRI data from nontarget trials.

Imaging data and model comparison

The predictive coding framework claims that there are neurons encoding prediction and prediction errors and that these neurons will respond in opposing ways to our factors of interest. Therefore, a model-based univariate approach has two caveats: there is always the potential that they will cancel one another out in univariate signals and the interpretation of univariate results will depend on assumptions about the relative numbers of prediction versus error units. Alternatively, a more conservative way to test the rival hypotheses is to look at multivariate activity pattern divergence/convergence between experimental conditions, which is directly inspired by the models and does not suffer from the two caveats. Therefore, imaging data were analyzed using whole-brain searchlight-based (Kriegeskorte et al., 2006), cross-subject MVPA to classify activation patterns between different

experimental conditions (see Materials and Methods, "MVPA procedure"). The classification accuracy quantifies the distinction between the activation patterns, or neural representations of the two conditions being classified, whereby higher classification accuracy indicates more distinct neural representations.

The multivariate fMRI analyses resembled a three-way ANOVA on the attention × color expectation × motion expectation factorial design. All classifiers were trained and tested on independent portions of the data using a leave-one-out approach over participants. We began with a positive control that involved testing the main effects for each of the three factors. For example, for the color factor, we trained classifiers on CE versus CU stimuli and used the resulting classifiers to predict which trials involved expected or unexpected stimuli in a left out participant. Similarly, for the motion factor, we trained and tested on ME versus MU stimuli, and for the attention factor, we trained and tested on "attend color" versus "attend motion" trials. These results are reported in the section entitled "Representation of feature-expectations in visual cortex." Subsequently, to adjudicate between the rival hypotheses regarding whether/how color- and motion-expectation interact, we performed the crucial test on the interaction between color- and motion-expectation (in the section titled "Contagion of surprise signals across stimulus features in EVC"). Then, to examine the role of feature-based attention in modulating fMRI activity patterns, we tested the interactions involving the attention factor (i.e., color expectation × attention, motion expectation × attention, and color expectation × motion expectation × attention) in the section titled "Attentional gain on feature representations in EVC depends on consistency of feature expectations." Finally, we conducted some control analyses to control for multiple comparisons and activity pattern consistency across subjects in order to further validate the results under our MVPA approach.

Representation of feature-expectations in visual cortex. By testing the main effects of each of the three factors using classifiers discriminating the two levels of the respective factor (e.g., testing the main effect of attention using classifiers discriminating color target vs motion target trials), we confirmed our a priori model assumption that information concerning whether stimulus features were expected is represented for both motion and color in EVC and selectively for motion and color in dorsal (area MT+) and ventral (V4) visual cortex, respectively (Grill-Spector and Malach, 2004; Fig. 2*B–D*). To follow up on the analyses of expectation effects on attended and unattended features (collapsed across feature dimensions) in Experiment 1, we further tested whether fMRI activation patterns allow reliable decoding of the expected and unexpected conditions with respect to the attended feature (e.g., classifiers discriminating CU/MU and CU/ME vs CE/MU and CE/ME trials in color target runs) and found significantly above-chance classifier performance in the EVC and nearby extrastriate visual cortex (binomial tests, $p < 0.05$, corrected; Fig. 2*E*). A repetition of this analysis using the unattended feature (e.g., classifiers discriminating CU/MU and CU/ME vs CE/MU and CE/ME trials in motion target runs) yielded similar findings (binomial tests, $p < 0.05$, corrected; Fig. 2*F*). In sum, these data replicate previous findings to validate our basic model structure and lay the groundwork for our main analyses of interest, namely, how the concurrent expectations in color and motion streams interact to shape neural stimulus representations.

Contagion of surprise signals across stimulus features in EVC. As outlined above (see Materials and Methods. "Design and rationale"), the three models make different predictions about the relative distance (distinction) between simulated neural activity in different experimental conditions (Table 1, also shown schematically in Fig. 3*D–F*). For this analysis, we divided our trials into four key conditions: (1) CU/MU, (2) CE/MU, (3) CU/ME, and (4) CE/ME according to whether the color, the motion, both, or neither was expected based on the conditional cue (Fig. 2). Specifically, the reconciliation model predicts that CE/ME and CU/MU conditions, in which both features are either expected or unexpected, will be more distinct (i.e., that neural classifiers will be more successful in distinguishing them) than the converse CU/ME and CE/MU conditions. In contrast, the segregation model predicts the converse, namely that neural patterns associated with CU/ME and CE/MU conditions will become more dissimilar, so classifiers will distinguish these conditions better than CE/ME versus CU/MU conditions. Finally, the independence model predicts that there will be no difference in classification accuracy between the CE/ME versus CU/MU and CE/MU versus CU/ME conditions. We calculated the distance in simulated neural signals (i.e., magnitude of *r* unit activity) in the EVC (due to its sensitivity to both color and motion information) that were output by each model in the CE/ME, CE/MU, CU/ME, and CU/MU conditions, collapsing across the attention factor. As can be seen in Figure 4*A*, the results were similar to the qualitative predictions outlined in Figure 3, *D–F*.

To adjudicate between these model predictions, we tested the interaction between color and motion expectation. Specifically, for each searchlight, we calculated the classification accuracy, which quantifies the distinction between two conditions on the basis of the pattern of neural activity they evoke. To test the hypotheses associated with each of the three models, we ran whole-brain searches focused on the relative ability of the classifier to distinguish between two pairs of conditions: CE/ME versus CU/MU ("expectation-consistent classifiers") and CE/MU versus CU/ME ("expectation-inconsistent classifiers"). For both types of classifiers, the expectancies were different between the two classes for both color and motion features. Therefore, the comparison between expectation-consistent and expectation-inconsistent classifiers was not biased by design. Within each searchlight, each color × motion expectation condition included two data points: one for each color-/motion-attended activation pattern.

This analysis revealed significant differences in classification accuracy in bilateral EVC ($p < 0.05$, corrected; Fig. 4*D*,*E*). Specifically, expectation-consistent classifiers (CU/MU vs CE/ME, mean accuracy = 0.718, $p < 0.001$, binomial test, $n = 92$, or 23 subjects × 2 classes × 2 attention conditions; Fig. 4*F*) outperformed expectation-inconsistent classifiers (CU/ME vs CE/MU, mean accuracy = 0.492, n.s., binomial test, $n = 92$; Fig. 4*F*) in a large region of EVC (peaking at 9, −88, −2, Brodmann area 17). To further demonstrate that this effect cannot be solely explained by attention, we repeated this analysis separately on color and motion change runs. In the same EVC region (Fig. 4*F*), CU/MU versus CE/ME classifiers performed significantly above chance level (color target trials: mean accuracy = 0.659, $p < 0.05$; motion target trials: mean accuracy = 0.654, $p < 0.05$, binomial tests, $n = 46$), whereas the CU/ME versus CE/MU classifiers had accuracy at chance level for both target conditions (color target trials: mean accuracy = 0.484, n.s.; motion target trials: mean accuracy = 0.506, n.s., binomial tests, $n = 46$). These results indicate that the representations of feature expectations in EVC were more distinct when the expectations were consistent than when they were inconsistent between streams. These results are consistent with predictions from the reconciliation model (i.e., a larger

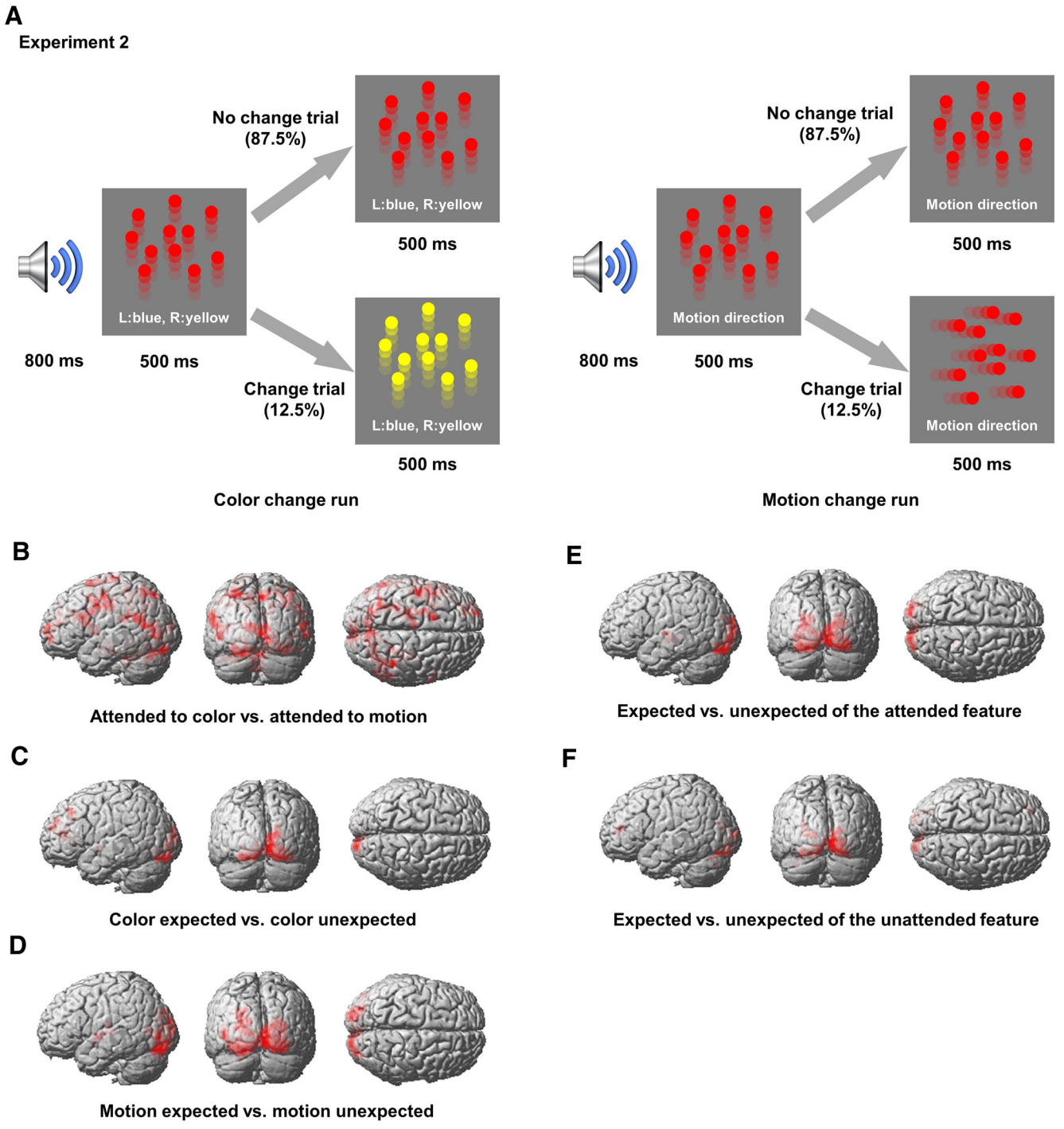

Figure 2. Experiment 2 task and fMRI validation results. ***A***, Example trials from color change detection (left) and motion change detection runs (right) in Experiment 2. Identical to Experiment 1, a trial started with a predictive auditory cue followed by moving dots. In 87.5% of all trials, neither the color nor the motion direction of the moving dots changed. In the other 12.5% of trials, the color (in color change runs) or motion (in motion change runs) changed after 500 ms. Subjects were required to identify the postchange feature. ***B***–***F***, Lateral, posterior, and dorsal views of brain areas showing significant (in red; binomial tests, $p < 0.05$, corrected) performance for attentional classifiers (***B***), color expectation classifiers (***C***), motion expectation classifiers (***D***), expectation of attended feature classifiers (***E***), and expectation of unattended feature classifiers (***F***).

distance between consistent than inconsistent conditions; Fig. 4*B*), but not with those of the independence and segregation models (Fig. 4*A*,*C*). No brain regions were found where neural activation patterns were more distinct when expectations were inconsistent than consistent.

Alternatively, this result could also be driven by a single outlier condition (either CU/MU or CE/ME, given the consistent > inconsistent classification accuracy) that was more distinct from all other three conditions. This interpretation would also predict a modulation of one feature expectation on the other. For example, if the CU/MU condition were the outlier condition, then it would follow that the distinction between CU/MU and CU/ME conditions is greater than the distinction between CE/MU and CE/ME conditions. To test this prediction, we conducted additional whole-brain analyses that tested the modulation of color expectation on motion expectation (i.e., does the performance of the

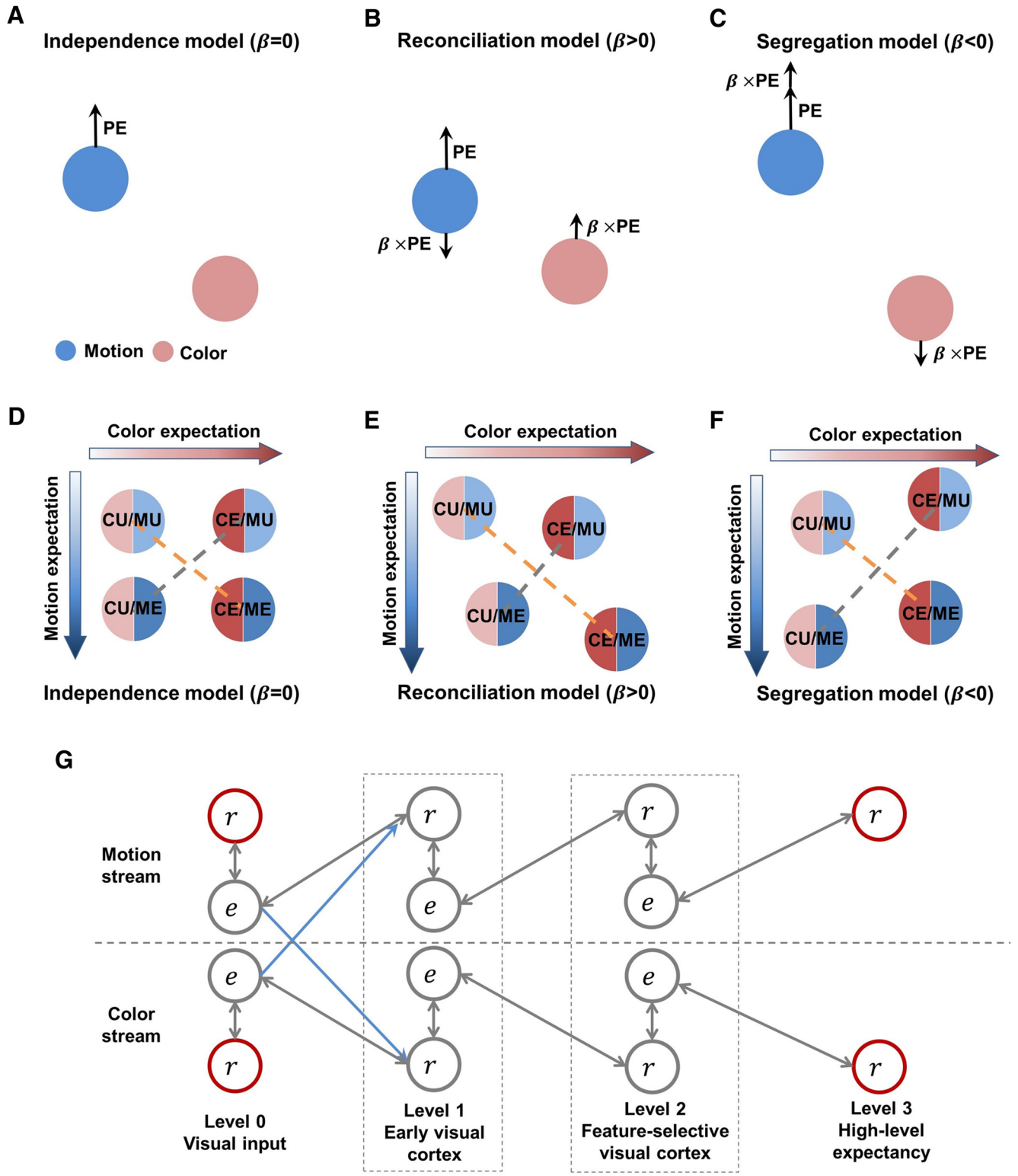

Figure 3. Model structure and rival hypotheses. ***A–C***, Schematic illustration of how PE in one feature affects the representation of the other feature in the expectation-inconsistent (here, CE/MU) conditions. The vertical axis represents PE level (i.e., the higher a disk, the greater the PE). Note how different signs of β lead to different mixed PEs (i.e., $\beta \times$ PE) that drive the representations of both features (disks) in different directions and then produce different levels of PE discrepancy between features (i.e., the distance between disks along the vertical direction). ***D–F***, Schematic illustration of different model predictions of color $\times$ motion expectation interactions. The lengths of the orange and gray dotted lines reflect the CU/MU-CE/ME distance and the CU/ME-CE/MU distance. ***G***, Structure of the predictive coding implementation of the conceptual models (same structure for all three models). This implementation consists of two visual processing streams (top: motion stream; bottom, color stream, separated by the dashed line) of four levels each. The levels used for model comparisons are surrounded by dotted boxes. Each level contains one representation (*r*) unit that encodes the prediction of the incoming input and up to one PE (*e*) unit that computes the PE of the prediction. The edges indicate information flow. At each moment, the *e* units send PEs to higher levels, which consequently adjust their prediction to account for the PE and then guide the adjustment of prediction at lower levels. The red nodes can receive input from outside of the model (e.g., visual input in level 0 and predictive information from the auditory cue in level 3). The interaction between the two features was implemented by the cross-stream edges from level 0 to level 1 (blue arrows). The three computational models only differ in their patterns of this interaction.

CE/MU vs CE/ME classifier differ from the CU/MU vs CU/ME classifier?) and vice versa. We did not find any brain regions showing such modulation (for the results in the aforementioned EVC region, see Fig. 4*G*), thus corroborating our interpretation, consistent with the reconciliation model.

Finally, this set of results could in principle also be explained by a generic encoding of PE, a feature-general surprise signal, for both color and motion direction (e.g., color and motion PE are encoded along the same dimension). Following this logic, CU/ME and CE/MU trials are inherently similar to each other because both are generically unexpected. However, note that this explanation is simply a restatement of the reconciliation model (i.e., surprise in one feature renders other features unexpected).

In summary, consistent with the behavioral results in Experiment 1, we found that multivariate information in EVC was best explained by the reconciliation model in which a positive PE mixing parameter results in surprise signals being spread from one visual object feature to another.

Attentional gain on feature representations in EVC depends on consistency of feature expectations. The effects of expectation on visual cognition are thought to interact with attention (Summerfield and Egner, 2009; Summerfield and de Lange, 2014). In this set of analyses, we therefore further tested whether the above findings can be solely attributed to attention and assessed how well the rival multifeature expectation models would be able to account for possible modulatory effects of (in)consistent feature expectations on the effects of feature-based attention. Specifically, for each color × motion expectation condition (CU/MU, etc.), we extracted the attentional effect of each of the two features defined as the (unsigned) difference of simulated activity between color-attended and motion-attended trials. This attentional effect on model activity allowed us to estimate, in a monotonic fashion, the predicted neural dissimilarity between the two attentional conditions (attended vs unattended) while keeping the expectation settings identical. For predictions about feature-selective visual areas (i.e., model levels 2: simulated V4 and MT+), the attentional effect was computed separately for color and motion. For the model simulation of EVC, sensitive to both color and motion, the two features' attentional effects were summed. Note that the size of the attentional effect is positively correlated with the magnitude of simulated neural activity (i.e., encoding strength) because attention was modeled as a multiplicative gain modulator on simulated neural activity.

All three models generated qualitatively similar predictions for color- and motion-selective regions (Fig. 5*A*,*B*) in which the attentional gain effect was larger when the preferred feature (e.g., color in the color stream) was expected than when it was unexpected. These effects resemble the two-way interaction between attention and color/motion expectation. In contrast, the predictions of possible color × motion expectation interaction effects on attentional gain were distinct between the three models at the level of EVC (Fig. 5*C*, Table 1), depicting different patterns of a three-way interaction among attention, color expectation, and motion expectation (with an emphasis on how the attentional effect is modulated by different combinations of color and motion expectancy). Specifically, the independence model predicted that the two feature expectations would independently modulate the multivariate effect of attention due to no difference in neural representation strength among expectation conditions. The reconciliation model predicted that the attentional effects would be larger in expectation-consistent conditions (CU/MU and CE/MU) than in expectation-inconsistent conditions

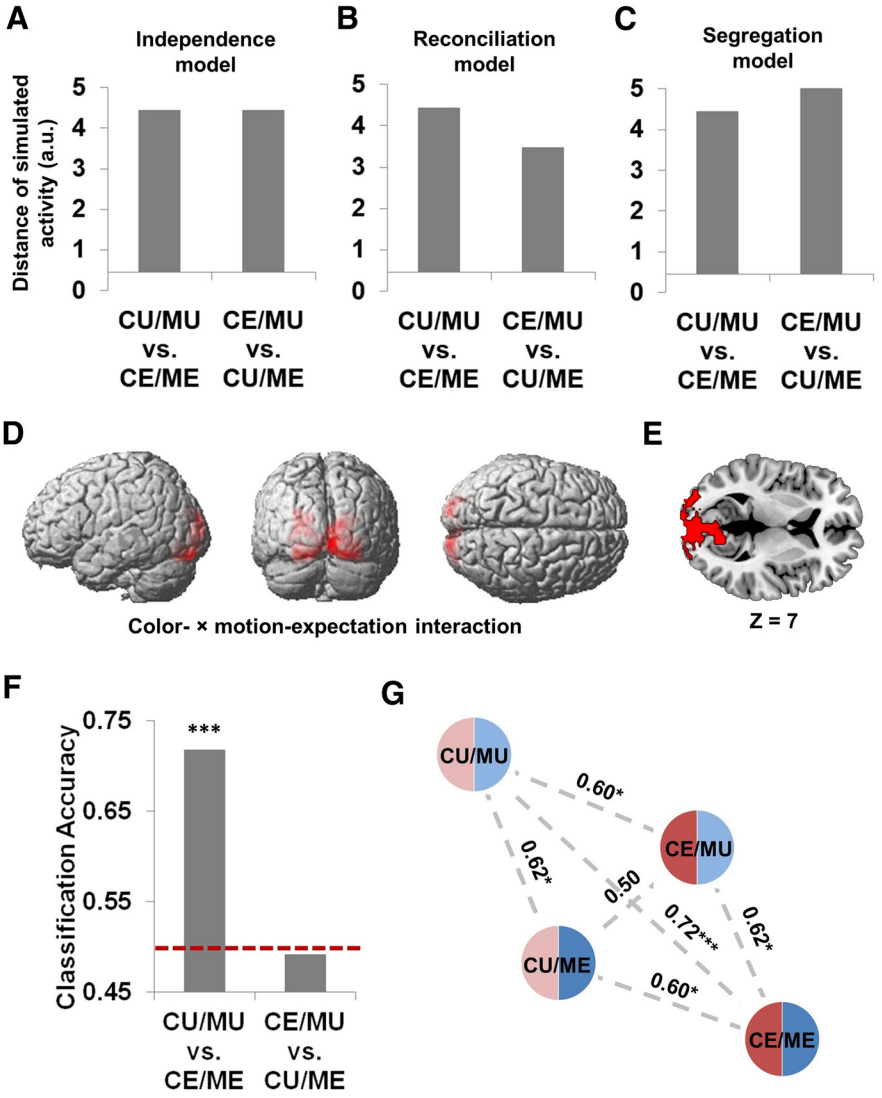

Figure 4. Joint effects of color and motion expectation in simulation and fMRI data. ***A–C***, Simulation results of the distances of *r* unit activity within expectation consistent and expectation inconsistent conditions at model level 1 (EVC, see Fig. 1*A*) using the independence model, reconciliation model, and segregation model, respectively. ***D***, Left to right, Lateral, posterior, and dorsal views of the EVC cluster (in red) showing a significant (Text S8, $p < 0.05$, corrected) interaction between color and motion expectation. ***E***, Axial slice showing the same cluster (in red) as in ***D***. ***F***, Mean classification accuracy in the EVC cluster in ***D*** and ***E*** plotted as a function of classifiers. The red dotted line marks the chance level (50%). ***G***, Classification accuracy for each pair of the color × motion expectation conditions. Using the cluster in ***D***, the length of a dotted line represents the cluster mean accuracy of discriminating activation patterns of the two conditions connected by that line. The numbers above the lines are the cluster-mean accuracy of the classifiers represented by those lines. *$p < 0.05$; ***$p < 0.001$ using binomial tests ($n = 92$).

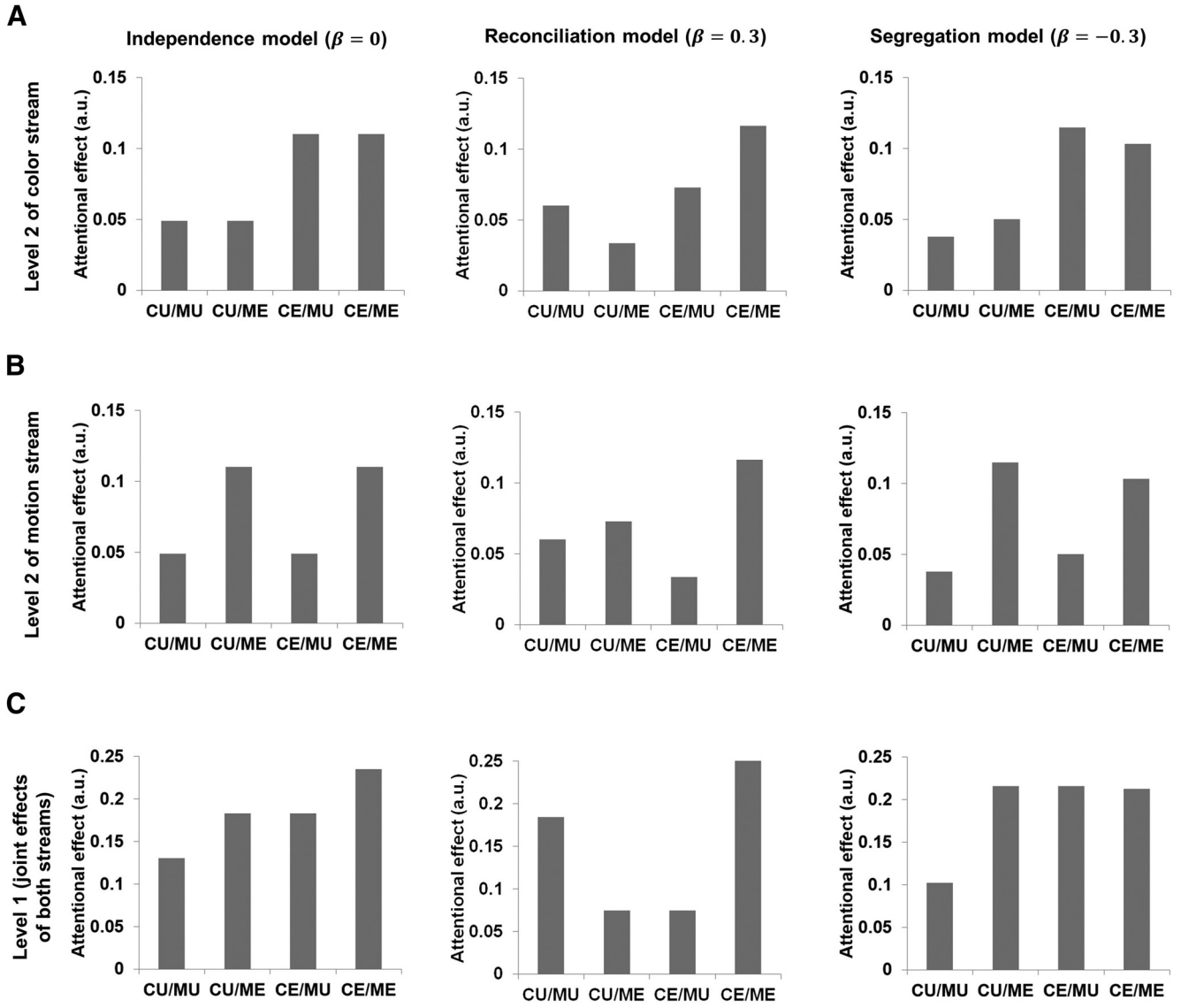

Figure 5. Model simulation results of the color × motion expectation modulation on attentional gain effects. ***A***, Simulation results using data from level 2 of the color stream (intended to simulate color selective V4). The bar graphs represent attentional gain effects or distance of *r* unit activity between color change and motion change conditions. The attentional effects are plotted as a function of color × motion expectation. Left to right, Simulation results using the independence model, reconciliation model, and segregation model, respectively. The simulation results using data from level 2 of the motion stream (intended to simulate color -motion MT+) and data from level 1 of both streams (intended to simulate EVC) are shown in the same format in ***B*** and ***C***, respectively.

(CU/ME and CE/MU) because of weakened neural feature representations caused by PE mixing in the latter conditions. In contrast, the segregation model predicted smaller attentional effects when expectations for the two features were consistent than when they were inconsistent as a result of enhanced processing within each feature in expectation-inconsistent conditions. We also included in this comparison an additional model that assumes that surprise attracts attention and hence overrides the manipulation of attention by task-relevance. Due to this override mechanism, this model would predict no significant attentional effects when either feature is unexpected.

We next adjudicated between these model predictions using fMRI data. To this end, we constructed whole-brain, searchlight-based, cross-subject attention classifiers (discriminating between attend color and attend motion activation patterns) for each color × motion expectancy condition (e.g., CU/MU trials in color target runs vs CU/MU trials in motion target runs). Note that because identical stimuli were used across the color and motion change detection runs, classification performance must reflect purely attentional effects. We then conducted a two-way ANOVA on the performance of these attention classifiers based on the two (color expectation) × two (motion expectation) design at each searchlight throughout the brain. We found a region in the anterior collateral sulcus (aCos) where color-attended and motion-attended trials evoked more dissimilar patterns of neural activity when color was expected than when it was unexpected ($p < 0.05$, corrected; Fig. 6*A*). As expected, based on our study design, this region corresponds closely to color-sensitive cortex defined in previous studies (Cavina-Pratesi et al., 2010). We also detected a region in lateral occipital cortex where classifiers were better able to distinguish color-attended from motion-attended trials when motion direction was expected than when it was unexpected ($p < 0.05$, corrected; Fig. 6*B*); this region corresponds closely to prior studies' localization of area MT+ (Rahnev et al.,

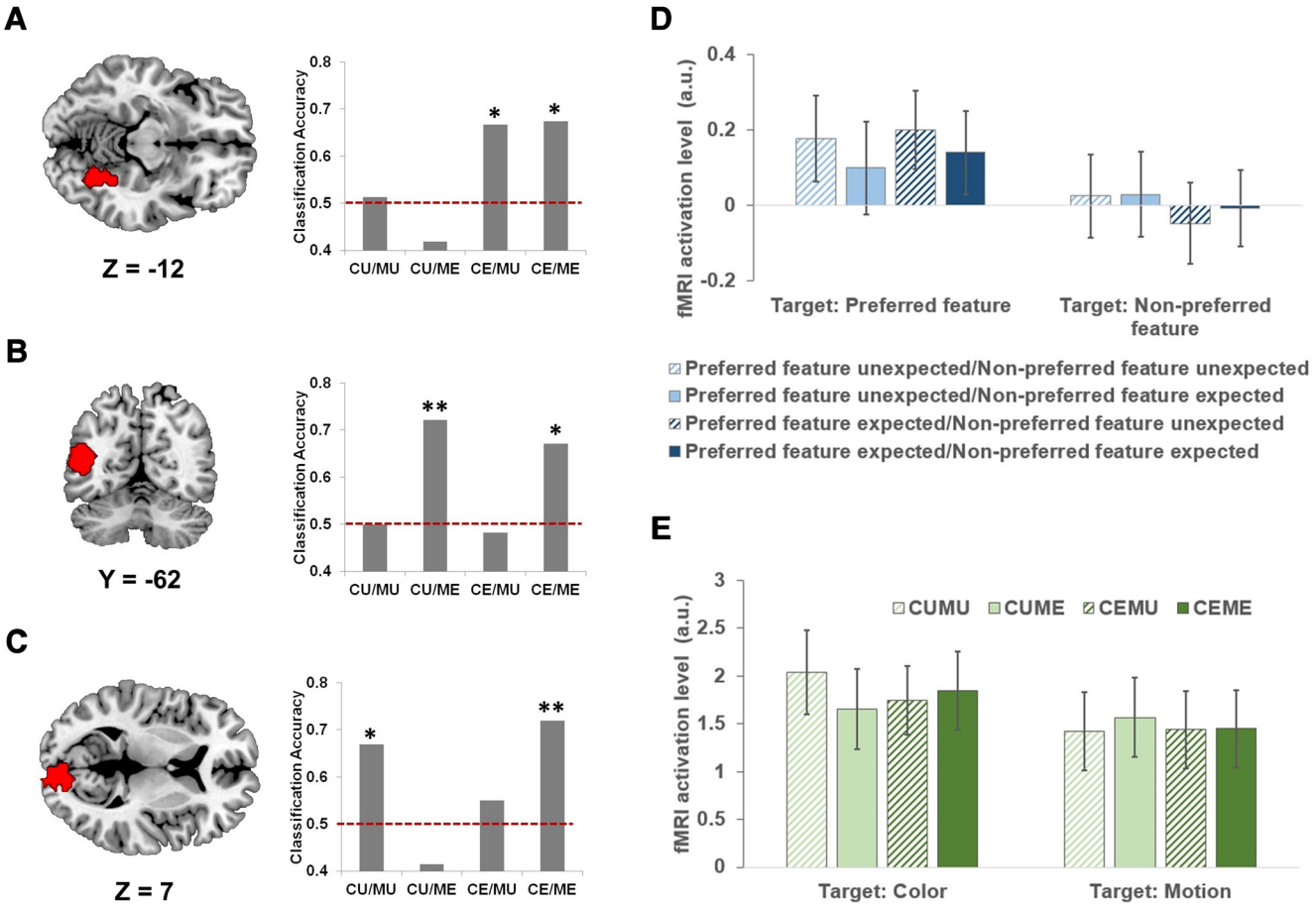

Figure 6. Effects of multifeature expectation on the accuracy of attention classifiers in visual cortex. ***A***, Left, Cluster of searchlights in the right aCos that displayed a significant (Text S7, $p < 0.05$, corrected) main effect of color expectation on attention classifiers. Right, Cluster mean attention classifier accuracy plotted as a function of color and motion expectation. ***B***, Left, Cluster of searchlights in the left lateral occipital cortex that displayed a significant (Text S7, $p < 0.05$, corrected) main effect of color expectation on attention classifiers. Right, Cluster mean attention classifier accuracy plotted as a function of color and motion expectation. ***C***, Left, Cluster of searchlights in early visual cortex that displayed a significant (Text S7, $p < 0.05$, corrected) interaction between color and motion expectation on attention classifiers' performance. Right, Cluster-mean attention classifier accuracy plotted as a function of color and motion expectation. The red dotted lines represent chance level classification (i.e., accuracy = 0.5). $^*p < 0.05$, $^{**}p < 0.005$, binomial tests ($n = 46$). ***D***, Mean fMRI activation level (±MSE) of the areas showing significant main effect of the preferred feature expectation on attentional effects (i.e., collapsed across the areas shown in ***A*** and ***B*** plotted as a function of target feature) and the expectation of preferred and nonpreferred features. ***E***, Mean fMRI activation level (±MSE) of the area shown in ***C*** plotted as a function of target feature, color, and motion expectation.

2011). These findings were consistent with the activation predictions for feature-selective level two nodes of all three models (Fig. 5*A*,*B*).

Crucially, however, we detected an interaction effect of color and motion expectation on attentional gain in EVC ($p < 0.05$, corrected; Fig. 6*C*) and this interaction selectively resembled the predictions of the reconciliation model (Fig. 5*C*). Specifically, the activation patterns differed significantly as a function of the attended feature (i.e., color or motion) only in expectation-consistent conditions (i.e., CU/MU and CE/ME), which was consistent with the reconciliation model's prediction of enhanced processing of visual information in these conditions. Importantly, these results did not support the model that surprise attracts attention, again suggesting that the results cannot be accounted for by attentional mechanisms only. In summary, whole-brain searchlight MVPA of attentional gain effects in the context of multifeature expectation interactions showed that discriminant information in EVC conforms to predictions of the reconciliation model, in which attentional effects are larger when the two feature predictions are either both confirmed or both violated compared with when their expectation statuses are inconsistent with each other. Consistent with the prior analyses of behavioral data and neural stimulus expectations, these results again provide selective support for a model in which a positive PE mixing parameter attenuates visual representation strength, and thus the multiplicative attentional gain effect, in expectation-inconsistent conditions.

Patterns of simulation results only rely on the sign of β

To show that the model predictions were not biased by the specific choices of model parameters, we ran the simulations with a wide range of attentional gain (α) and PE mixing (β) parameters and found that the qualitative pattern of simulation results (i.e., the sign of the effect of the unattended feature expectation in Fig. 2*B*; whether expectation consistent classifiers outperform expectation inconsistent classifiers in Fig. 4, *A*–*C*; and the color × motion expectation interaction pattern on attentional effects in Fig. 5*C*) only depended on the sign of β, which by definition was how the rival models are distinguished (Fig. 7).

Univariate fMRI results

To explore the relationship between the above multivariate results and mean signal neural strength in the corresponding visual regions, we conducted univariate analyses on the area-mean ac-

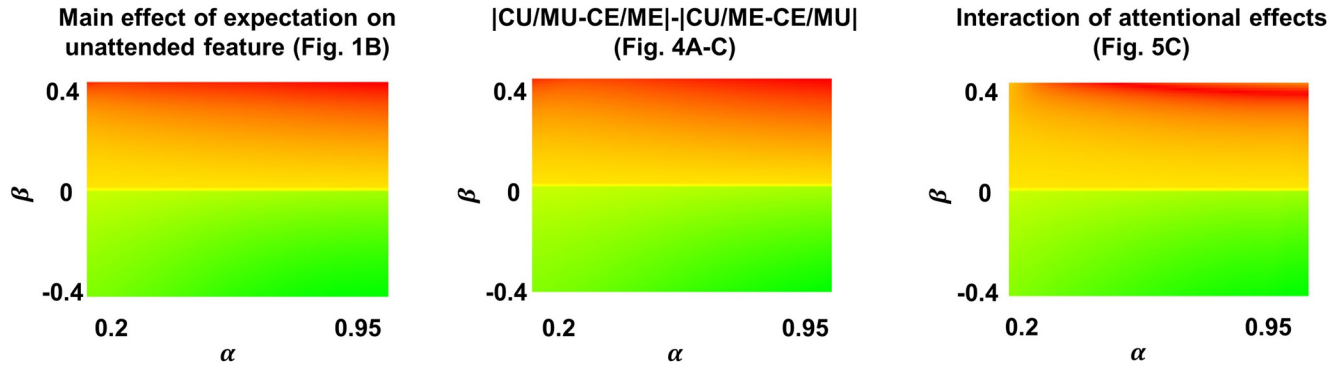

Figure 7. The simulation results are only sensitive to the sign of β. Each heat map visualizes an effect that has divergent model predictions. Left to right, Expectation effect on the unattended feature shown in Figure 1*B*, distance between expectation-consistent conditions minus the distance between expectation-inconsistent conditions shown in Figure 4, *A–C*, and the interaction of the two features' expectation on attentional effect shown in Figure 5*C*. We conducted the same analyses of model outputs as in the main text with a wide range of free parameter settings. Specifically, α (horizontal axis) ranges from 0.2 (400% of attentional gain) to 0.95 (~5% of attentional gain) and β (vertical axis) ranges from −0.4 to 0.4. Each cell on a heat map represents the result from a model in which the α and β parameters are determined by the horizontal and vertical coordinates, respectively. The color encodes the simulated effect. Positive, zero, and negative effects were color coded in red, yellow and green, respectively. The "redness" and "greenness" further indicate the magnitude of the effect. In all three heat maps, the size of the simulated effects displayed a similar dependence on the parameters. Crucially, the signs of all simulated effects are only sensitive to the sign of β. Therefore, our simulation results are not biased by the choices of specific model parameters.

tivity levels in these regions (Fig. 6*A–C*). First, because the rival hypotheses did not predict any difference between the color and motion stream level two areas, we collapsed across the aCos (Fig. 6*A*) and MT+ (Fig. 6*B*) areas and performed a repeated-measures three-way ANOVA (attention × preferred feature expectation × nonpreferred feature expectation; Fig. 6*D*). We found a significant main effect of attention ($F_{(1,22)} = 5.91$, $p < 0.05$) driven by higher activity level when the target feature was the preferred feature (0.15 ± 0.11) than the nonpreferred feature (0.00 ± 0.10). This is consistent with the finding of increased neuronal firing rate driven by an attended stimulus (for review, see Reynolds and Chelazzi, 2004). We also observed a marginally significant attention-reversed expectation effect in the preferred feature ($F_{(1,22)} = 3.52$, $p = 0.07$), as described previously (Kok et al., 2012b). We then conducted a repeated-measures three-way ANOVA (attention × color expectation × motion expectation; Fig. 6*E*) on the EVC area and found a marginally significant main effect of attention ($F_{(1,22)} = 3.91$, $p < 0.06$) and a significant three-way interaction ($F_{(1,22)} = 9.83$, $P = 0.005$) that mimics the pattern found in MVPA results (i.e., larger attentional effects in expectation-consistent than expectation-inconsistent conditions; Fig. 6*C*). Therefore, whereas the univariate analyses, as expected a priori, were less sensitive in distinguishing the experimental conditions, the mean regional BOLD responses were broadly consistent with the MVPA findings and reflected known effects of expectation and attention.

Validation of cross-subject MVPA

To test whether our MVPA approach was prone to false-positive findings, we compared the cluster size of the four reported ROIs (EVC reported in Fig. 4*E*, aCos in Fig. 6*A*, MT+ in Fig. 6*B*, and EVC in Fig. 6*C*) with a null distribution of cluster sizes using the same voxelwise height threshold of uncorrected $p < 0.01$. The null distribution was obtained by randomly shuffling fMRI activation levels in the visual brain (including occipital cortex and ventral and dorsal visual pathway regions of the superior and inferior parietal sulci, fusiform gyri, and middle and inferior temporal gyri, based on the AAL template), conducting the exact same cross-subject MVPA analyses (i.e., expectation consistent vs inconsistent, Fig. 4, and the two-way ANOVA on attentional classifiers, Fig. 6) and then evaluating the sizes of all clusters obtained using the threshold of $p < 0.01$. For each analysis, this procedure was repeated 50 times, resulting a total of ~11,000 clusters for forming the null distribution of cluster size. Consistent with the results of the standard correction for multiple comparisons, all 4 ROIs were significantly larger than clusters obtained from scrambled data (EVC in Fig. 4*E*: $p < 0.0001$, aCos: $p < 0.001$, MT+: $p < 0.0005$, EVC in Fig. 6*C*: $p < 0.0001$). Therefore, our analysis approach was not prone to false positives.

Table 2. Results of cross-subject fMRI activity pattern consistency

ROI name	Random scramble *z*-value range	*z*-value from real data	*p*-value
EVC (Fig. 4*E*)	[−0.0006, 0.0006]	0.1314	<0.0001
aCos (Fig. 6*A*)	[−0.0083, 0.0105]	0.0691	<0.0001
MT+ (Fig. 6*B*)	[−0.0072, 0.0080]	0.0314	<0.0001
EVC (Fig. 6*C*)	[−0.0052, 0.0060]	0.1234	<0.0001

Cross-subject MVPA requires that neural activity patterns are consistent across subjects. To gauge such consistency, we calculated the correlation of activity patterns between subjects. Specifically, this analysis was conducted separately for each of the four reported ROIs. To further test whether signal (as opposed to noise) exists at the level of single searchlights, for each searchlight in a given ROI, we calculated the difference of activation patterns between each pair of the eight conditions in the experimental design and computed the *z*-transformed correlation coefficients for each pair of subjects. The reason for using the difference of activation patterns between two conditions is to simulate the MVPAs. The *z*-values were then averaged across conditions, subjects, and searchlights. The resulting mean *z*-value, which represents pattern consistency across subjects, was then compared with the mean *z*-values calculated using randomly scrambled data in the same ROI (repetition = 10,000 times). The results are summarized in Table 2. These data show that the univariate activity, which was used in MVPA, indeed contained signal patterns that were consistent across subjects and can be decoded using cross-subject MVPA.

The assumption of pattern consistency across subjects also predicts that the voxelwise weights in the classifiers were preserved across subjects. To test this prediction, for each searchlight in each of the aforementioned four ROIs, we randomly split the

Table 3. Results of the preservation of voxelwise weights in cross-subject fMRI

ROI name	Contrast	Random scramble *z*-value range	*z*-value from real data	*p*-value
EVC (Fig. 4*E*)	Expectation consistent versus expectation inconsistent	[−0.2307, 0.3049]	0.6743	<0.001
aCos (Fig. 6*A*)	Attentional classifier: color expected/motion unexpected	[−0.4064, 0.4463]	0.8935	<0.001
aCos (Fig. 6*A*)	Attentional classifier: color expected/motion expected	[−0.3426, 0.4587]	0.8159	<0.001
MT+ (Fig. 6*B*)	Attentional classifier: color unexpected/motion expected	[−0.3727, 0.3999]	0.8720	<0.001
MT+ (Fig. 6*B*)	Attentional classifier: color expected/motion expected	[−0.3816, 0.3960]	0.7967	<0.001
EVC (Fig. 6*C*)	Attentional classifier: color unexpected/motion unexpected	[−0.2027, 0.2248]	0.4898	<0.001
EVC (Fig. 6*C*)	Attentional classifier: color expected/motion expected	[−0.2128, 0.2366]	0.7459	<0.001

subjects into two groups, calculated the voxelwise weights of classifiers for each group, and tested the correlation of weights between groups. This procedure was repeated 100 times for each searchlight and the mean *z*-transformed correlation coefficients were used as a quantification of the preservation of voxelwise weights in cross-subject MVPA. Due to the high computational cost, we only computed contrasts that we reported as statistically significant. The ROI mean *z*-value was compared with *z*-values computed using randomly scrambled data of the same ROI (repetition = 1000 times). The results are summarized in Table 3. As can be seen in Table 3, the obtained correlations in the empirical data were significantly greater than correlations generated from scrambled fMRI data (all $p < 0.001$). Therefore, these results clearly support the crucial assumption that the weights of classifiers were indeed preserved across subjects at the voxel level.

Even though the neural populations (e.g., cortical columns) calculating the prediction and prediction errors operate at a much finer spatial scale than the spatial resolution of fMRI, previous MVPA studies have shown that the voxel-level fMRI response is sensitive to changes in columnar level neural activity in the EVC and can thus be used to decode orientation in visual stimuli (Haynes and Rees, 2005; Kamitani and Tong, 2005). In the framework of predictive coding, the canonical microcircuits model (Bastos et al., 2012) ties the conceptual roles of computing prediction and prediction errors and the hierarchy of the predictive coding framework to the functions and connectivity of cortical columns. Following this logic, a match/mismatch between expectations and bottom-up input could lead to different columnar activity even for the same stimulus. Furthermore, given that columns are tuned to respond to different features (e.g., specific motion directions, specific colors), different columns may have different neural responses to the same stimulus. As a result, voxel-level fMRI activity may be modulated by the proportions of cortical columns it samples and by expectation. Our control analyses showed consistent fMRI activity patterns across subjects (Tables 2, 3), which leads us to speculate that the distributions of columnar responses may vary as a function of the spatial locations of the columns in the EVC at a spatial scale similar to the spatial resolution of fMRI.

Discussion

Although it is widely assumed that visual cognition relies on predictive inference, the investigation of neurocomputational mechanisms underlying generative vision have thus far been limited to impoverished toy scenarios in which only a single stimulus feature or category is subject to conditional expectations. Here, we built on this work to tackle the more complex but realistic scenario of the visual brain managing concurrent expectations for multiple object features and to shed light on the transformation from expectations concerning individual stimulus features to a unified, object-level expectation. To develop and test formal hypotheses, we harnessed computational modeling in combination with behavioral and neuroimaging data, which allowed us to adjudicate between rival possibilities concerning how different feature expectations (and attention) interact in driving perceptual decisions and neural representations (Table 1). Behavioral data (Fig. 1) and fMRI data (Figs. 4, 6) from two experiments unanimously supported predictions of a "reconciliation model" that assumes PE mixing, or a spreading of surprise, across different features of an object: when one feature expectation is violated, PE spreads to other features, rendering the object as a whole unexpected. This PE contagion provides a mechanism to promote object-level prediction and perceptual inference.

The dual-prediction modeling framework developed here is grounded in basic tenets of predictive coding (Friston, 2005) and attention (Reynolds and Chelazzi, 2004), as well as prior findings on the nature of color and motion processing in visual cortex (Gegenfurtner, 2003; Born and Bradley, 2005). The present fMRI data confirmed all of the key model assumptions, including the encoding of feature-selective color and motion expectations (Fig. 2*B–D*) in ventral and dorsal extrastriate visual cortex, respectively, paired with mixed selectivity for color and motion expectation (and their attentional modulation) in EVC. Moreover, all of the simulated neural activity patterns predicted by the reconciliation model (Table 1) were observed in fMRI activations patterns in the EVC (Figs. 2*B–F*, 4*D–G*, 6*C*). This is precisely consistent with our model implementation, in which the cross-feature blending of PE occurs at the simulated EVC level, an assumption that was based on prior demonstrations that neurons in primary visual cortex are sensitive to both color and motion information (Movshon and Newsome, 1996; Engel et al., 1997; Johnson et al., 2001; Kamitani and Tong, 2006). At the microscopic level, this PE mixing in EVC could stem from an intermingling of parvocellular color-sensitive (Perry et al., 1984) and magnocellular motion-sensitive (Wiesel and Hubel, 1966) inputs from the lateral geniculate nucleus of the thalamus, which has been documented in previous studies of V1 (for review, see Sincich and Horton, 2005). Although our model clearly represents a gross simplification of the rich interplay between early and later stages of the visual system, it successfully captured some basic neural population signatures of multifeature expectations while adhering to a biologically plausible architecture and processing principles.

Our main findings document that, rather than treating expectations concerning different object features as independent or promoting the assumption that expected and unexpected features belong to different objects, the visual brain appears to exchange PE between visual features to form object-level expectations such that surprise in one feature spreads to other features and ultimately renders the perception of all features of the object unexpected. The idea of object-level selection has a long history in the study of attention (Duncan, 1984), for which a number of behavioral (Egly et al., 1994; He and Nakayama, 1995) and neural

(Roelfsema et al., 1998; O'Craven et al., 1999) studies have shown that attending to one location on, or feature of, an object confers an attentional advantage to other locations and features of that object. Importantly, the present data now show that objects, rather than single features or spatial locations, represent the default unit of selection, not only for relevance-driven (i.e., attention), but also for probability-driven (i.e., expectation) endogenous determinants of visual cognition. Furthermore, object-level selection implied by the reconciliation model would also predict that the mixed PE should increase the similarity between the cue–feature associations learned from different features. This similarity should in principle also facilitate the learning of a unified cue–object association across trials. Future studies are encouraged to test this prediction.

Interestingly, our findings also document an interaction between expectation and attention in the modulation of multifeature processing. In particular, although attention generally enhanced feature representations in higher visual regions (Fig. 6*A*,*B*) and in expectation-consistent conditions in the EVC (CU/MU and CE/ME conditions; Fig. 6*C*), this attentional modulation effect was absent in EVC for expectation-inconsistent conditions (CE/MU and CU/ME conditions; Fig. 6*C*). According to the reconciliation model, this is because, in expectation-inconsistent conditions, PE mixing results in attenuated neural feature representations (Table 1), which in turn dampens their attentional modulation. Conversely, the attention-modulated PE enters the PE mixing process and spreads to unattended features associated with the same object. In other words, PE mixing also transfers the attentional modulation to unattended features, which is again consistent with the above-mentioned spreading of attention across object features.

Although our study and model were designed to focus on how object-level expectation is implemented in visual cortical processing of individual features, an important question to ask is where the belief that these features belong to the same object might originate. Possible answers to this question may be found in the literature on feature binding (or "feature integration"), which has long been considered integral to object perception (Treisman and Gelade, 1980; Treisman, 1998) and proposed to be an obligatory operation in human cognition (Ashby et al., 1996; Hommel, 2004). Prior lesion and neuroimaging studies have observed involvement of parietal cortex (Treisman, 1998) and of both classic learning systems of the brain, the medial temporal lobe/ hippocampus (Mitchell et al., 2000; Jiang et al., 2015) and the striatum (Jiang et al., 2015), in the perceptual and mnemonic binding of different event features. These regions therefore constitute prime candidates for generating the integrated, object-level predictions that drive the effects we here documented in visual cortex; assessing the exact mechanisms by which these or other more anterior regions (e.g., hippocampus, see Hindy et al., 2016) impose top-down object-level expectations represents a key goal for future studies.

Given the close relationship between attention and expectation (Summerfield and Egner, 2009, 2016), we took several measures to ensure that the present results were not due to attentional mechanisms. First, in the experimental design, attention and expectation were dissociated. Second, we conducted a key analysis that compared expectation-consistent and expectation-inconsistent classifiers (Fig. 4*D*–*F*) by collapsing across color and motion target trials and performing this analysis on these two types of trials separately. All three analyses revealed the same results, thus strongly suggesting that attention to target features cannot account for the current results. Third, we tested whether a hypothesis that one unexpected feature attracts attention can explain some of the results. This hypothesis, along with the findings of a significant main effect of attention in the EVC, would predict significantly distinct fMRI activity patterns between CU/ME and CE/MU trials as a result of an attentional effect (i.e., attention was attracted to color and motion in CU/ME and CE/MU trials, respectively). However, the CU/ME versus CE/MU classifiers did not perform above chance level (Fig. 4*F*). Moreover, we conducted another analysis that directly contradicted this hypothesis by showing a significant attentional effect on EVC neural activity patterns in CU/MU trials (Fig. 6*C*), which would not be expected to show attentional effects under this hypothesis. Fourth, another alternative hypothesis could be that violation of prediction in any feature would result in reallocation of attention to both features. Assuming that the BOLD signal reflects a joint effect of feature-based attention from task instruction and the redistribution of attention due to high PE this hypothesis would predict reduced performance of attention classifiers in any condition with expectation violation given that the redistribution of attention would increase similarity in BOLD signal between color and motion target trials. In fact, these predictions were consistent with chance-level performance observed in CU/ME and CE/MU conditions. Similarly, chance-level classifier performance should also be expected in CU/MU conditions. However, this was not supported by the significant attentional effects in the CU/MU condition in EVC (Fig. 6*C*). In general, compared with various attentional mechanisms that may be able to explain only part of the reported results, the reconciliation model provides a parsimonious account for all empirical findings in this study.

In conclusion, we have shown how the visual brain implements concurrent predictive coding of multiple stimulus features. Our modeling and empirical data converge on the conclusion that feature expectations interact to drive object-level predictions: surprise from one unexpected feature spreads to other features to render the object unexpected. These findings constitute a major advance in our understanding of the neurocomputational substrates of active vision in the human brain.

References

Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L (2010) Stimulus predictability reduces responses in primary visual cortex. J Neurosci 30:2960–2966. CrossRef Medline

Ashby FG, Prinzmetal W, Ivry R, Maddox WT (1996) A formal theory of feature binding in object perception. Psychol Rev 103:165–192. CrossRef Medline

Auksztulewicz R, Friston K (2015) Attentional enhancement of auditory mismatch responses: a DCM/MEG study. Cereb Cortex 25:4273–4283. CrossRef Medline

Bar M (2004) Visual objects in context. Nat Rev Neurosci 5:617–629. CrossRef Medline

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical microcircuits for predictive coding. Neuron 76:695–711. CrossRef Medline

Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol 14:143–177. CrossRef Medline

Born RT, Bradley DC (2005) Structure and function of visual area MT. Annu Rev Neurosci 28:157–189. CrossRef Medline

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436. CrossRef Medline

Cavina-Pratesi C, Kentridge RW, Heywood CA, Milner AD (2010) Separate channels for processing form, texture, and color: evidence from FMRI adaptation and visual object agnosia. Cereb Cortex 20:2319–2332. CrossRef Medline

Clithero JA, Smith DV, Carter RM, Huettel SA (2011) Within- and cross-participant classifiers reveal different neural coding of information. Neuroimage 56:699–708. CrossRef Medline

Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, Gur RC, Langleben DD (2005) Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. Neuroimage 28:663–668. CrossRef Medline

de Gardelle V, Waszczuk M, Egner T, Summerfield C (2013) Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. Cereb Cortex 23:2235–2244. CrossRef Medline

den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in associative learning. Cereb Cortex 19: 1175–1185. CrossRef Medline

Duncan J (1984) Selective attention and the organization of visual information. J Exp Psychol Gen 113:501–517. CrossRef Medline

Egly R, Driver J, Rafal RD (1994) Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. J Exp Psychol Gen 123:161–177. CrossRef Medline

Egner T, Monti JM, Summerfield C (2010) Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 30:16601–16608. CrossRef Medline

Engel S, Zhang X, Wandell B (1997) Colour tuning in human visual cortex measured with functional magnetic resonance imaging. Nature 388:68–71. CrossRef Medline

Feldman H, Friston KJ (2010) Attention, uncertainty, and free-energy. Front Hum Neurosci 4:215. CrossRef Medline

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836. CrossRef Medline

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138. CrossRef Medline

Gegenfurtner KR (2003) Cortical mechanisms of colour vision. Nat Rev Neurosci 4:563–572. CrossRef Medline

Gordon AM, Rissman J, Kiani R, Wagner AD (2014) Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. Cereb Cortex 24:3350–3364. CrossRef Medline

Grill-Spector K, Malach R (2004) The human visual cortex. Annu Rev Neurosci 27:649–677. CrossRef Medline

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ (2011) A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72:404–416. CrossRef Medline

Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–691. CrossRef Medline

He ZJ, Nakayama K (1995) Visual attention to surfaces in three-dimensional space. Proc Natl Acad Sci U S A 92:11155–11159. CrossRef Medline

Hindy NC, Ng FY, Turk-Browne NB (2016) Linking pattern completion in the hippocampus to predictive coding in visual cortex. Nat Neurosci 19: 665–667. CrossRef Medline

Hommel B (2004) Event files: feature binding in and across perception and action. Trends Cogn Sci 8:494–500. CrossRef Medline

Jiang J, Schmajuk N, Egner T (2012) Explaining neural signals in human visual cortex with an associative learning model. Behav Neurosci 126: 575–581. CrossRef Medline

Jiang J, Summerfield C, Egner T (2013) Attention sharpens the distinction between expected and unexpected percepts in the visual brain. J Neurosci 33:18438–18447. CrossRef Medline

Jiang J, Brashier NM, Egner T (2015) Memory meets control in hippocampal and striatal binding of stimuli, responses, and attentional control states. J Neurosci 35:14885–14895. CrossRef Medline

Johnson EN, Hawken MJ, Shapley R (2001) The spatial transformation of color in the primary visual cortex of the macaque monkey. Nat Neurosci 4:409–416. CrossRef Medline

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685. CrossRef Medline

Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–1102. CrossRef Medline

Kanai R, Komura Y, Shipp S, Friston K (2015) Cerebral hierarchies: predictive processing, precision and the pulvinar. Philos Trans R Soc Lond B Biol Sci 370 pii: 20140169. CrossRef Medline

Kaplan JT, Meyer K (2012) Multivariate pattern analysis reveals common neural patterns across individuals during touch observation. Neuroimage 60:204–212. CrossRef Medline

Keller GB, Bonhoeffer T, Hübener M (2012) Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. Neuron 74:809–815. CrossRef Medline

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304. CrossRef Medline

Kok P, Jehee JF, de Lange FP (2012a) Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75:265–270. CrossRef Medline

Kok P, Rahnev D, Jehee JF, Lau HC, de Lange FP (2012b) Attention reverses the effect of prediction in silencing sensory signals. Cereb Cortex 22: 2197–2206. CrossRef Medline

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline

Martinez-Trujillo JC, Treue S (2004) Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr Biol 14:744–751. CrossRef Medline

Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci U S A 108:19401–19406. CrossRef Medline

Mitchell KJ, Johnson MK, Raye CL, D'Esposito M (2000) fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. Brain Res Cogn Brain Res 10:197–206. CrossRef Medline

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang XR, Just M, Newman S (2004) Learning to decode cognitive states from brain images. Machine Learning 57:145–175. CrossRef

Mourão-Miranda J, Bokde AL, Born C, Hampel H, Stetter M (2005) Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage 28:980–995. CrossRef Medline

Movshon JA, Newsome WT (1996) Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. J Neurosci 16:7733–7741. Medline

Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66:241–251. CrossRef Medline

Nassar MR, Wilson RC, Heasly B, Gold JI (2010) An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci 30:12366–12378. CrossRef Medline

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. Nat Neurosci 15:1040–1046. CrossRef Medline

O'Craven KM, Downing PE, Kanwisher N (1999) fMRI evidence for objects as the units of attentional selection. Nature 401:584–587. CrossRef Medline

Onat S, Büchel C (2015) The neuronal basis of fear generalization in humans. Nat Neurosci 18:1811–1818. CrossRef Medline

Palmer TE (1975) The effects of contextual scenes on the identification of objects. Mem Cognit 3:519–526. CrossRef Medline

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199–209. CrossRef Medline

Perry VH, Oehler R, Cowey A (1984) Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey. Neuroscience 12:1101–1123. CrossRef Medline

Poldrack RA, Halchenko YO, Hanson SJ (2009) Decoding the large-scale structure of brain function by classifying mental States across individuals. Psychol Sci 20:1364–1372. CrossRef Medline

Rahnev D, Lau H, de Lange FP (2011) Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. J Neurosci 31:10741–10748. CrossRef Medline

Rao RP (2005) Bayesian inference and attentional modulation in the visual cortex. Neuroreport 16:1843–1848. CrossRef Medline

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87. CrossRef Medline

Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. Psychol Sci 9:347–356. CrossRef

Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. Annu Rev Neurosci 27:611–647. CrossRef Medline

Reynolds JH, Pasternak T, Desimone R (2000) Attention increases sensitivity of V4 neurons. Neuron 26:703–714. CrossRef Medline

Roelfsema PR, Lamme VA, Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. Nature 395:376–381. CrossRef Medline

Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA (2008) Using FMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS One 3:e1394. CrossRef Medline

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA (2011) Commonality of neural representations of words and pictures. Neuroimage 54:2418–2425. CrossRef Medline

Sincich LC, Horton JC (2005) The circuitry of V1 and V2: integration of color, form, and motion. Annu Rev Neurosci 28:303–326. CrossRef Medline

Spratling MW (2008) Reconciling predictive coding and biased competition models of cortical function. Front Comput Neurosci 2:4. CrossRef Medline

Summerfield C, de Lange FP (2014) Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci 15: 745–756. CrossRef Medline

Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends Cogn Sci 13:403–409. CrossRef Medline

Summerfield C, Egner T (2016) Feature-based attention and feature-based expectation. Trends Cogn Sci 20:401–404. CrossRef Medline

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004–1006. CrossRef Medline

Thiele A, Pooresmaeili A, Delicato LS, Herrero JL, Roelfsema PR (2009) Additive effects of attention and stimulus contrast in primary visual cortex. Cereb Cortex 19:2970–2981. CrossRef Medline

Treisman A (1998) Feature binding, attention and object perception. Philos Trans R Soc Lond B Biol Sci 353:1295–1306. CrossRef Medline

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136. CrossRef Medline

Wacongne C, Changeux JP, Dehaene S (2012) A neuronal model of predictive coding accounting for the mismatch negativity. J Neurosci 32:3665–3678. CrossRef Medline

Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R (2012) A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. Psychol Bull 138: 1172–1217. CrossRef Medline

Wiesel TN, Hubel DH (1966) Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. J Neurophysiol 29:1115–1156. Medline

Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681–692. CrossRef Medline