
Research Articles: Behavioral/Cognitive

Visual prediction error spreads across object features in human visual cortex

Jiefeng Jiang (江界峰)¹, Christopher Summerfield² and Tobias Egner^{1,3}

¹Center for Cognitive Neuroscience, Duke University, Durham, North Carolina, 27708, USA

²Department of Experimental Psychology, University of Oxford, OX1 3UD, United Kingdom

³Department of Psychology and Neuroscience, Duke University, Durham, North Carolina, 27708, USA

DOI: 10.1523/JNEUROSCI.1546-16.2016

Received: 12 May 2016

Revised: 25 October 2016

Accepted: 29 October 2016

Published: 3 November 2016

Author contributions: J.J., C.S. and T.E. designed the study and wrote the paper. J.J. performed the data collection and analysis.

Conflict of Interest: The authors declare no competing financial interest.

This work was funded by NIMH award R01MH097965 (T.E.). We thank Nadia Brashier for help with data acquisition.

Corresponding author: Jiefeng Jiang, Jiefeng.jiang@duke.edu, P.O.Box 90999, Duke University, Durham, North Carolina, 27708, USA

Cite as: J. Neurosci 2016; 10.1523/JNEUROSCI.1546-16.2016

Alerts: Sign up at www.jneurosci.org/cgi/alerts to receive customized email alerts when the fully formatted version of this article is published.

Visual prediction error spreads across object features in human visual cortex

Abbreviated title: Visual prediction error spreads across object features

Jiefeng Jiang (江界峰)¹, Christopher Summerfield², and Tobias Egner^{1,3}

¹Center for Cognitive Neuroscience, Duke University, Durham, North Carolina, 27708, USA

²Department of Experimental Psychology, University of Oxford, OX1 3UD, United Kingdom

³Department of Psychology and Neuroscience, Duke University, Durham, North Carolina, 27708, USA

Corresponding author:

Jiefeng Jiang

Jiefeng.jiang@duke.edu

P.O.Box 90999

Duke University

Durham, North Carolina, 27708, USA

Number of pages: 58

Number of figures: 7

Number of tables: 3

Number of words for Abstract: 249

Number of words for Introduction: 645

Number of words for Discussion: 1456

Author contributions

J.J., C.S. and T.E. designed the study and wrote the paper. J.J. performed the data collection and analysis.

Conflict of interest

None declared.

Acknowledgments

This work was funded by NIMH award R01MH097965 (T.E.). We thank Nadia Brashier for help with data acquisition.

38 **Abstract**

39 Visual cognition is thought to rely heavily on contextual expectations. Accordingly,
40 previous studies have revealed distinct neural signatures for expected vs. unexpected
41 stimuli in visual cortex. However, it is presently unknown how the brain combines
42 multiple concurrent stimulus expectations, like those we have for different features of a
43 familiar object. To understand how an unexpected object feature affects the
44 simultaneous processing of other expected feature(s), we combined human functional
45 magnetic resonance imaging (fMRI) with a task that independently manipulated
46 expectations for color and motion features of moving-dot stimuli. Behavioral data and
47 neural signals from visual cortex were then interrogated to adjudicate between three
48 possible ways in which prediction error (surprise) in the processing of one feature might
49 affect the concurrent processing of another, expected feature: (1) feature processing
50 may be independent; (2) surprise might “spread” from the unexpected to the expected
51 feature, rendering the entire object unexpected; (3) pairing a surprising feature with an
52 expected feature might promote the inference that the two features are not in fact part of
53 the same object. To formalize these rival hypotheses, we implemented them in a simple
54 computational model of multi-feature expectations. Across a range of analyses,
55 behavior and visual neural signals consistently supported a model that assumes a
56 mixing of prediction error signals across features: surprise in one object feature spreads
57 to its other feature(s), thus rendering the entire object unexpected. These results reveal
58 neuro-computational principles of multi-feature expectations and indicate that objects
59 are the unit of selection for predictive vision.

60 **Significance statement**

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

70 Introduction

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

85 While the central role of expectations in perceptual inference is now widely
86 acknowledged, and some of its basic implications have been successfully modeled (e.g.,
87 (Spratling, 2008; Jiang et al., 2012; Wacongne et al., 2012)), one particularly notable
88 shortcoming is that we do not know how the visual brain manages multiple,
89 simultaneous expectations for different features of an object, such as its color, shape
90 and size. Prior studies have employed only simple, one-dimensional scenarios, where
91 predictions and surprise signals were limited to a single feature of a given object or
92 object category (e.g., the forthcoming stimulus likely being a face, or a right-tilted Gabor

93 patch) (Egner et al., 2010b; Kok et al., 2012a). In the real world, however, object
 94 expectations are rarely limited to a single feature. For instance, a soccer player must
 95 form expectations about both the motion of surrounding players and the color of their
 96 jerseys, in order to distinguish trajectories of teammates from those of opponents. Thus,
 97 we typically acquire, and make use of, concurrent expectations about multiple features
 98 of an object. Importantly, this can give rise to circumstances where one feature
 99 conforms to expectations but another feature does not. A key unresolved question, thus,
 100 is how the brain resolves conflict between inconsistent feature expectations to produce
 101 unified object-level perception.

102 In the present study, we therefore investigated how the processing of one
 103 stimulus feature (say, player motion) is affected by the violation of expectations
 104 concerning another feature (jersey color) of the same stimulus. To understand this core
 105 aspect of visual object cognition, we used behavioral and fMRI data to adjudicate
 106 between three rival hypotheses: First, the two feature expectations might operate
 107 independently of each other, such that an expectation violation of one feature would not
 108 affect the processing of the other feature (“independence model”). Second, perceptual
 109 expectations may operate at an object-level, such that one surprising feature might
 110 render the entire object (including the expected feature) surprising (“reconciliation
 111 model”). A parallel to this scenario exists in the attention literature, where attending to
 112 one feature (or part) of an object can lead to the attentional selection of the entire object
 113 (Egley et al., 1994; O’Craven et al., 1999). Third, the co-occurrence of an expected and
 114 an unexpected object feature might motivate the perceptual hypothesis that the two
 115 features are not in fact part of the same object (“segregation model”). This hypothesis

116 echoes findings in figure-ground segmentation, where subjects tend to interpret a single
 117 unusual shape as reflecting a collection of mutually occluding, common shapes
 118 (reviewed in (Wagemans et al., 2012)). Finally, whether, and in what manner, a
 119 surprising feature may affect the processing of an expected feature could plausibly
 120 interact with feature-based attention, that is, the feature's relevance to the current task
 121 (Summerfield and Egnér, 2009); therefore, our models also incorporated effects of
 122 feature-based attention.

123

124 **Materials and Methods**

125 **Design and Rationale**

126 Our goal was to determine how the visual brain processes expectations for multiple
 127 features of a single object, as a function of whether a given feature is attended. We
 128 operationalized this problem with a perceptual categorization task involving a stimulus
 129 (a coherent motion field of dots) composed of two independently varying features: color
 130 and motion direction (Fig. 1A, 2A). Both of these features are known to drive neural
 131 responses in early visual cortex (EVC) (Movshon and Newsome, 1996; Engel et al.,
 132 1997; Johnson et al., 2001; Kamitani and Tong, 2006), but are thereafter processed by
 133 specialized areas of the ventral (color: V4) (Gegenfurtner, 2003) and dorsal (motion:
 134 area MT+) (Born and Bradley, 2005) visual streams.

135 This provides an ideal scenario for testing how an expectation (or violations
 136 thereof) for one stimulus feature affects the processing of another feature of the same
 137 object, both in feature-selective regions (i.e., V4 and MT+) as well as in regions

138 sensitive to both of these features (i.e., EVC). To this end, we independently
 139 manipulated whether a given feature conformed to, or violated, perceptual expectations.
 140 These manipulations produced four experimental conditions: color-unexpected/motion-
 141 unexpected (CU/MU), color-expected/motion-expected (CE/ME), color-
 142 unexpected/motion-expected (CU/ME) and color-expected/motion-unexpected (CE/MU).
 143 Thus, the expectation status across the two features is consistent in CU/MU and CE/ME
 144 conditions, but inconsistent in CU/ME and CE/MU conditions. To assess how multi-
 145 feature expectations interact with attention and to dissociate expectation effects from
 146 attentional effects, we furthermore independently varied the task-relevance of the two
 147 feature dimensions (attend to color vs. attend to motion).

148 Using this experimental design, we compared three types of predictive coding
 149 models concerning how expectation and surprise interact between object features to
 150 produce unified object perception. This interaction relies on cross-feature exchange of
 151 prediction error (PE), which drives the updating of neural representation to match
 152 sensory input. Specifically, a parameter β is used to determine the proportion of PE that
 153 propagates from one feature stream to the other (see Materials and Methods:
 154 Computational simulation). When expectations are *consistent* across features (i.e.,
 155 CE/ME and CU/MU conditions), the PEs are identical for both features (either both are
 156 low or both are high), such that any PE mixing across feature streams is balanced: the
 157 same amount of color PE would propagate to the motion stream as the other way
 158 around. Therefore, PE mixing does not alter feature processing in these conditions.
 159 Crucially, however, when expectations are *inconsistent* between features (i.e., CU/ME
 160 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”, where no PE mixing occurs (Fig. 3A), and therefore 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). By contrast, setting β to a positive value simulates the “reconciliation model” (Fig. 3B), 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: (1) 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. (2) This decreases the expectation effect (i.e., the discrepancy between unexpected and expected conditions, Fig. 3B) in expectation-inconsistent conditions, thus making CU/ME and CE/MU less distinct from each other, as compared to expectation-consistent conditions (Fig. 3E). (3) Consequently, this type of PE mixing makes the unexpected feature less unexpected and the expected feature less expected (Fig. 3B). Therefore, the PE mixing would interfere with within-feature information processing,

183 making the neural representations of features in expectation-inconsistent trials weaker
 184 than in expectation-consistent trials.

185 Conversely, setting β to a negative value simulates the “segregation model”,
 186 where the unexpected feature sends PE to the expected feature stream to drive its
 187 processing in the *opposite* direction, while enhancing its own PE to boost within-feature
 188 processing (Fig. 3C). In other words, the segregation model resolves clashing
 189 expectations between features by discarding the premise that the features belong to the
 190 same object, and producing segregated and enhanced perceptions for each feature
 191 instead. To wit, observing an expected motion trajectory paired with an unexpected
 192 jersey color would result in an updated belief that the jersey color and object motion are
 193 caused by two different players. Compared to the reconciliation model, the reversed
 194 sign of β in the segregation model thus leads to the exact opposite predictions. All
 195 model predictions are summarized in Table 1.

196 We adjudicated between the three rival models using behavioral and
 197 neuroimaging data from the following two experiments. Note that all the model
 198 predictions concern differences in neural representations or prediction error between
 199 conditions. The key goal of our fMRI analyses was to quantify these distinctions. To this
 200 end, we adopted multi-voxel pattern analysis (MVPA) as our hypothesis testing tool,
 201 because MVPA measures how separable the neural activity patterns of different
 202 conditions are, and the resulting classification accuracy is a natural quantification of
 203 condition separability. The rationale for focusing on MVPA (rather than GLM) results
 204 was also driven by additional considerations stemming from the predictive coding
 205 framework that underlies our models (see below). This framework assumes that

206 computational units involved in producing expectations and prediction error are located
 207 in close spatial proximity (e.g., Bastos et al., 2012). Given random sampling of such
 208 units across fMRI voxels, previous studies have found spatially intermingled voxels with
 209 signals that were either primarily driven by expectation or prediction error signals (de
 210 Gardelle et al., 2013). This implies that mean regional BOLD signals derived from
 211 conventional univariate analysis with spatial smoothing blend together expectation and
 212 surprise signals (Egner et al., 2010a) and therefore have limited sensitivity for
 213 distinguishing different expectation conditions (see also Kok et al., 2012a). By contrast,
 214 MVPA treats each voxel independently and is capable of exploiting heterogeneous
 215 response profile in adjacent voxels in order to distinguish activity patterns of different
 216 experimental conditions. For example, given two intermingled groups of voxels, one
 217 showing A > B activity and the other showing B > A activity, averaging across (e.g.,
 218 smoothing) these voxels may cancel out any difference between these conditions, but
 219 MVPA can assign positive and negative weights to these two groups to ‘align’ their
 220 opposite patterns of activity, in order to distinguish between A and B conditions.

221 **Experiment 1 (Behavior)**

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

227 *Stimuli.* The presentation of stimuli and response recording were controlled using
 228 Psychtoolbox version 3 (Brainard, 1997). The auditory stimuli were comprised of four
 229 tones. Each tone consisted of four notes (200 ms each) which were ordered to produce
 230 either a rise or fall in pitch. Thus, the rising and falling tones did not differ in the notes
 231 used, but only in the way the notes were ordered. Additionally, the tones were played in
 232 two distinct timbres, resulting in a 2 (rising/falling pitch) \times 2 (timbres) factorial design.
 233 These auditory stimuli were delivered via noise-cancelling headphones.

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

242 *Procedure.* Each trial started with the presentation of the auditory cue tone, which was
 243 followed by the moving dots display (Fig. 1A). Thus, the cue and stimulus processing
 244 did not overlap in sensory modality. The cue's timbre and pitch were predictive of the
 245 forthcoming dots' color and motion direction at 75% validity, respectively. To avoid
 246 potentially confusing violations in contingency, up/down motion was always predicted by
 247 rising/falling tones, respectively. For each trial, the participants were asked to identify
 248 the color or motion direction of the dots with button-presses. The target feature (color or
 249 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 inter-trial interval (ITI) of 1.5s.

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 (5 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 prior to 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 counter-balanced 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. The response mapping was counter-balanced 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.

272 *Analysis.* The accuracy for each condition in the 2 (feature attention) × 2 (color-
 273 expectation) × 2 (motion-expectation) factorial design was calculated and entered into a
 274 repeated-measures 3-way ANOVA. The same analysis was performed on response
 275 time (RT) means, after excluding RTs from error trials or outlier trials (i.e., trials with
 276 RTs outside the range of grand mean \pm 2.5 SD).

277 **Experiment 2 (fMRI)**

278 *Subjects.* Twenty-five right-handed volunteers gave informed consent in accordance
 279 with institutional guidelines and completed this experiment. All subjects had normal or
 280 corrected-to-normal vision. Two subjects were excluded from further analysis due to
 281 excessive head movement during scanning (movement > 6mm or 6° within any run).
 282 The final sample consisted of 23 subjects (14 females, 22 – 35 years old, mean age =
 283 27 years). This study was approved by the Duke University Health System Institutional
 284 Review Board.

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

294 *Procedure.* The training, test, and practice sessions were identical to experiment 1. The
 295 main task consisted of eight runs (in the order of ABABBABA, with the first run counter-
 296 balanced across subjects) of 64 trials each, with exponentially jittered ITIs (from 4s to 6s
 297 with a step size of 500ms). Different from experiment 1, the goal of this task was to
 298 identify occasional changes in color/motion via button-press. The target feature (e.g.,
 299 color) was cued at the beginning of each run. The subjects were also explicitly informed
 300 that no change would occur in the non-target feature in order to encourage the subjects
 301 to direct attention solely to the target feature. Therefore, similar to experiment 1, this
 302 experimental design resulted in a 2 (feature attention) \times 2 (color-expectation) \times 2
 303 (motion-expectation) factorial design.

304 In order to manipulate feature-based attention and keep subjects on task, eight
 305 trials (12.5%) per run were randomly selected as “change trials” (or target trials), in
 306 which the target feature (color/motion) changed to yellow or blue/left or right (at 50%
 307 probability) after 500ms (Fig. 2A), which had to be reported by the subjects based on a
 308 response mapping displayed at the bottom of the screen throughout each run. However,
 309 fMRI analysis only included the frequent non-target trials to avoid confounds from motor
 310 responses or target-related processing (Summerfield et al., 2008). The auditory cues
 311 had no predictive value regarding the post-change color/motion in change trials.
 312 Nevertheless, in no-change trials, the expectation effects were still mediated by the
 313 auditory cues that preceded each dot cloud. The numbers of presentations for each
 314 tone \times color-motion combination were equated within the no-change trials for each run
 315 and each condition of the factorial design to avoid bias in the analyses.

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

319 *Image acquisition and preprocessing.* Images were acquired parallel to the AC-PC line
 320 on a 3T GE scanner (Milwaukee, WI). Structural images were scanned using a T1-
 321 weighted SPGR axial scan sequence (146 slices, slice thickness = 1mm, TR = 8.124ms,
 322 FoV = 256mm * 256mm, in-plane resolution = 1mm * 1mm). Functional images were
 323 scanned using a T2*-weighted single-shot gradient EPI sequence of 42 contiguous axial
 324 slices (slice thickness = 3mm, TR = 2s, TE = 28ms, flip angle = 90 °, FoV = 192mm *
 325 192mm, in-plane resolution = 3mm * 3mm). Functional data were acquired in 8 runs of
 326 206 images each. Preprocessing was done using SPM8
 327 (<http://www.fil.ion.ucl.ac.uk/spm/>). After discarding the first five scans of each run, the
 328 remaining images underwent spatial realignment, slice-time correction, spatial
 329 normalization and resulted in normalized functional images in their native resolution. As
 330 is customary in MVPA, no spatial smoothing was applied to the normalized fMRI images.

331 *MVPA procedures.* For each subject and each experimental condition in the factorial
 332 design (attention × color- × motion expectation), we generated an activation map that
 333 encodes the t-value of this condition at every grey matter (GM) voxel. Specifically, the
 334 normalized images were regressed against a general linear model (GLM) to estimate
 335 activation levels for each experimental condition. The GLM consisted of nine event-
 336 based regressors (convolved with SPM 8's canonical hemodynamic response function)
 337 representing the onsets of no-change trials in each of the eight conditions of the
 338 factorial design, the onsets of change trials, and nuisance regressors representing head

339 motion parameters, as well as the grand mean of the run (to remove the run-specific
 340 baseline signal and activity elicited by the response mapping instructions that were
 341 presented throughout each run). Note that the specific stimuli (for example, red color,
 342 downward motion) were counter-balanced and collapsed within each cell of the design,
 343 as we were interested in classifying neural patterns that distinguished the processing of
 344 different feature dimensions (that is, color vs. motion) rather than different intra-
 345 dimensional exemplars (for example, red vs. green). This approach applied to both
 346 expectation and attention based classifiers. In other words, within each cell of the
 347 factorial design, the presented color and motion stimuli belonged to the same attention
 348 and expectation conditions in order to enable the tests of generic (i.e., not specific to
 349 particular colors and motions) attention and expectation effects. This GLM also
 350 controlled for the unequal trial counts between expected and unexpected conditions, as
 351 all trials within a particular condition were grouped into one regressor, such that
 352 expected and unexpected conditions were represented by an equal number of
 353 regressors (or data points) for the MVPAs. As a result, for each subject and each
 354 experimental condition in the factorial design (attention \times color expectation \times motion
 355 expectation), this step generated an activation map that encodes the t-value of this
 356 condition at every grey matter (GM) voxel defined in the segmented SPM T1 template
 357 (dilated by one voxel). For each subject, activation estimates were further normalized
 358 within voxels and across the eight conditions to remove individual difference in baseline
 359 activation level and absolute amplitude of activations.

360 The MVPA was performed in a searchlight-based (Kriegeskorte et al., 2006) ,
 361 inter-subject manner using a leave-one-out (LOO) cross-validation approach: the

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

382 We took this cross-subject approach based on three considerations. First, this
 383 approach places the strong constraint on our findings that the mixture of computations
 384 driving the BOLD signal (whilst unknown) must be consistent (generalizable) across

385 subjects at the voxel level after anatomical normalization, which is also the assumption
 386 of the widely used univariate fMRI analysis. This constraint is crucial in the present work,
 387 as it focuses on the early visual cortex, one of the regions with the smallest degree of
 388 anatomical and functional individual differences in the cerebrum. Compared to within-
 389 subject MVPA, the assumptions underlying group results for cross-subject MVPA are in
 390 fact more similar to the standard mass-univariate analysis group results, in that cross-
 391 subject MVPA requires the effects of interest to be in the same direction across subjects.
 392 Previous cross-subject MVPA studies have demonstrated this consistency by
 393 successfully decoding complex cognitive states such as task state (Mourao-Miranda et
 394 al., 2005; Poldrack et al., 2009), lying or telling the truth (Davatzikos et al., 2005), the
 395 ambiguity of a presented sentence (Mitchell et al., 2004), receiving monetary or social
 396 reward (Clithero et al., 2011), presence/absence of conflict in cognitive control (Jiang et
 397 al., 2015), experiencing pain (Gordon et al., 2014), fear conditioning (Onat and Buchel,
 398 2015), and observing people touching different objects (Kaplan and Meyer, 2012). In
 399 visual cortex, a number of studies have demonstrated that, after standard anatomical
 400 alignment, high cross-subject MVPA accuracy can be achieved in the decoding of visual
 401 content (Haxby et al., 2011; Shinkareva et al., (2008); Shinkareva et al., (2011)). Of
 402 direct relevance to the current study, it has also previously been shown that this cross-
 403 subject generalizability held for the effects of different attention and expectation
 404 conditions on visual cortex signal (Jiang et al., 2013). Second, the current design, due
 405 to the importance of concurrently manipulating expectations in two features,
 406 necessitated the creation of some rare event conditions, namely the low probability
 407 events of CU/MU trials (16 trials/subject). This low trial count creates sub-optimal

408 conditions for running within-subject MVPA, a statistical power problem that can be
 409 countered by employing the cross-subject MVPA approach that includes trials from all
 410 subjects to increase the trial count (to 23 subjects \times 16 trials/subject) for the CU/MU
 411 condition in the MVPAs. As shown in Fig. 4F and Fig. 6C, analyses involving CU/MU
 412 trials did in fact reveal significantly above-chance classification accuracies, suggesting
 413 the chosen cross-subject MVPA approach was not hampered by low trial counts (high
 414 variance) in this condition. Third, the cross-subject approach allowed us to control for a
 415 potential confound introduced by specific response mappings, because the mappings
 416 were counter-balanced across subjects.

417 To test the effects of the manipulation of feature-based attention and expectation,
 418 we built classifiers discriminating fMRI activity patterns of no-change trials between
 419 color- and motion-target runs (Fig. 2B), color-expected and color-unexpected conditions
 420 (Fig. 2C), and motion-expected and motion-unexpected conditions (Fig. 2D),
 421 respectively. Furthermore, in conjunction with behavioral analyses in experiment 1 (Fig.
 422 1C), we constructed expectation classifiers (i.e., expected vs. unexpected) for the
 423 attended (Fig. 2E) and unattended feature (Fig. 2F), respectively, in order to further
 424 examine the interaction between attention and expectation. Moreover, to test how
 425 expectation (or violation thereof) of one feature affects the expectation of the other
 426 feature, we followed the model predictions in Fig. 3A-F and Table 1, and compared the
 427 performance of two fMRI activity pattern classifiers: one that discriminated between
 428 CU/MU and CE/ME trials and another one that discriminated between CU/ME and
 429 CE/MU trials (Fig. 4D-G). Finally, to test whether/how the effect of attention varies as a
 430 function of concurrent expectation of color and motion, we constructed fMRI activity

431 pattern classifiers between attend-color and attend-motion conditions separately for
 432 each of the four color- × motion-expectation conditions and tested whether classifier
 433 performance varies as a function of expectation conditions (Fig. 6A-C).

434 As a result, for each classifier, a group-level classification accuracy map was
 435 computed, where each GM voxel represented the classification accuracy from the LOO
 436 cross-validation of the searchlight centered at that voxel. For each searchlight, the
 437 statistical significance of its performance was gauged using a binomial test. The
 438 difference of classification performance between two maps was compared using a
 439 Bayesian approach. This approach inferred the probability that two classification
 440 accuracies observed from the same searchlight over two different accuracy maps
 441 belonged to the same underlying classification accuracy based on the distributions of
 442 accuracy in these two accuracy maps (for details, see (Jiang et al., 2013; Jiang et al.,
 443 2015)).

444 *Statistical analysis and control for false positives.* For all aforementioned statistical
 445 analyses, false positives due to multiple comparisons were controlled for at $P < 0.05$ (for
 446 classification analyses, the P values were obtained using binomial tests for each
 447 searchlight or ROI) for combined searchlight classification accuracy and cluster extent
 448 thresholds, using the AFNI ClusterSim algorithm
 449 (http://afni.nimh.nih.gov/pub/dist/doc/program_help/3dClustSim.html). Ten thousand
 450 Monte Carlo simulations determined that an uncorrected voxelwise P value threshold of
 451 < 0.01 (for P value transformed from binomial distribution, the largest P value that was
 452 less than 0.01) in combination with a searchlight cluster size 21 to 32 searchlights
 453 (depending on the specific analysis) ensured a false positive rate of < 0.05 .

454 **Computational simulation**

455 *Computational modeling.* To enable quantitative and formal predictions about responses
 456 under predictive coding, this paper introduces a particular predictive coding scheme that
 457 was used to simulate perceptual inference under the three hypotheses above. This
 458 allowed us to simulate particular response profiles that we then tested for using
 459 behavioral reports and multivariate analysis of physiological responses. To this end, the
 460 aforementioned three rival models were implemented using a biologically feasible
 461 predictive coding model (Friston, 2005) (Fig. 3G), which posits a continual interplay
 462 across the visual cortical hierarchy between the top-down passing of predictions
 463 concerning forthcoming inputs and the bottom-up passing of PE (Mumford, 1992; Rao
 464 and Ballard, 1999; Friston, 2005, 2010). Predictive coding models have been
 465 demonstrated to account for many empirical findings in the visual cognition literature (for
 466 review, see Summerfield and de Lange (2014)). To simulate the processing of the two
 467 features of color and motion, the model consists of two “visual streams”, specialized in
 468 processing either feature (Fig. 3G). The model streams comprise four levels, namely, an
 469 input stage (level 0), followed by an early visual cortex (EVC) stage (level 1) that is
 470 sensitive to both color and motion direction, then followed by higher-level, feature-
 471 selective visual cortex (level 2) that are sensitive to either color (i.e., V4) or motion
 472 direction (i.e., MT+), and finally, putative higher-level regions (level 3) that provide
 473 expectation inputs to the simulated lower level regions. In line with the tenets of
 474 predictive coding (Friston, 2005), each level consists of two types of computational units
 475 (except for the top level): “representation units” that encode predictions of bottom-up
 476 inputs, and “error units” that receive top-down input from representation units at the

477 next-higher level, calculate prediction error (i.e., the discrepancy between predicted and
 478 actual input), and pass that error back to the representation units at the next-higher
 479 level. The co-occurrence of predictive and surprise signals in visual cortex has been
 480 confirmed in previous studies (Egner et al., 2010a; Keller et al., 2012; de Gardelle et al.,
 481 2013).

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

492 In order to incorporate effects of attention into the model, we furthermore allowed
 493 feature-relevance to impose a multiplicative gain on visual processing (Martinez-Trujillo
 494 and Treue, 2004) by an attentional factor a^s . In the framework of predictive coding,
 495 attention is modeled as the precision or confidence of the prediction errors (Feldman
 496 and Friston, 2010; Aukstulewicz and Friston, 2015; Kanai et al., 2015), whereby more
 497 attention equates to enhanced prediction error input forwarded to the next level. This
 498 assumption can successfully account for findings from behavioral cued attention studies
 499 (e.g., Feldman & Friston, 2010), and attentional sharpening of prediction error signals

500 has also been documented at the level of fMRI signal in ventral visual cortex (e.g., Jiang
 501 et al (2013)). Attention/confidence-modulated prediction error can also be interpreted as
 502 a mathematical formulation of surprise that consists of two levels of uncertainty, namely
 503 the (violation) of prediction (i.e., prediction error), and the confidence of this prediction
 504 (Yu and Dayan, 2005). This factor also simulates attentional modulation on
 505 representation units (Rao, 2005; Spratling, 2008).

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

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

517
 518 Where a^s was higher in attended than unattended streams. For example, in a
 519 color-detection change run, $a^0 > a^1$. We modeled attentional gain in both attended and
 520 unattended features because it has been reported that attention can also spread from

521 attended features to other features of the same object (O'Craven et al., 1999). a^s was
522 set to 1 and 0.75 for attended and unattended conditions, respectively.

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

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

525 Similarly, the modulation of a^s on r_{i+1}^s was further implemented by applying a^s to
526 the input, for example: $r_0^s = a^s u$, where u was the visual input, which remained constant
527 during simulation. The non-input representation units were updated in the following
528 manner:

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

529

530 Thus, updating of r_i^s was also modulated by a^s through the prediction errors.
531 $e_3^s(t)$ was a constant of 0 due to the fact that level 3 had no error unit. In sum, attention
532 and expectation were modeled separately using a^s and θ_i^s , respectively.

533 Crucially, the aforementioned crosstalk between the two stimulus features was modeled
534 in EVC, which is sensitive to both motion and color. To introduce the effect of object-
535 level perception on the processing of individual features, the above predictive coding
536 model was extended to accommodate the belief that individual features were generated
537 from the same object. Specifically, at each time point t , the updating of $r_1^s(t)$ is further
538 modulated by this belief using the aforementioned parameter β and a mechanism that

539 allowed for the “mixing” of the inputs from level 0 to level 1 across streams (Fig. 3G,
 540 blue links) to mediate the updating of r_1^s in the following manner:

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

541 Where a^s was applied to the prediction error from the other stream, in order to reflect
 542 the attentional modulation on the prediction error at the recipient stream. Thus, $e_0^s(t) -$
 543 $a^s \times \beta \times e_0^s(t) + \beta \times e_0^{1-s}(t)$ represented a mixed prediction error from level 0.
 544 Specifically, when β is 0 (representing a neutral belief regarding whether the color and
 545 motion are from the same objects or not), Eq (4) is identical to Eq (3) to simulate the
 546 independence model. When $\beta > 0$ (representing the belief that the color and motion are
 547 from the same object), the updating of color and motion expectations are “synchronized”
 548 using β to facilitate an object-level expectation, as hypothesized in the reconciliation
 549 model. Lastly, when $\beta < 0$ (representing the belief that the color and motion come from
 550 different objects), the mixed prediction error differentiates and enhances the updating of
 551 expectations for individual features, and therefore simulates the competition model.
 552 This model has two free parameters, namely the attentional modulator for the
 553 unattended stream (a) and β , which models modulation that is spread over attended
 554 and unattended streams.

555 Because the training and practice phases ensured that the subjects had learned
 556 the experimental manipulation of color- and motion-expectation, we did not model the
 557 learning effects of u and r_3^s during the simulation of the two experiments. To simulate
 558 the two tasks in this study, r_i^s ranged from -1 (completely tuned to represent the
 559 unexpected stimulus) to 1 (completely tuned to represent the expected stimulus), with 0

560 reflecting neutral selectivity. The absolute value of representation unit activity also
 561 represents the encoding strength of the observed features (e.g., an activity level of -0.8
 562 represents a stronger neural representations of the observed unexpected feature than
 563 an activity level of -0.5). Accordingly, e_i^s ranged from -2 to 2. $u = 0$ when the cue was
 564 presented, and 1 and -1 when the visual stimulus was expected or unexpected,
 565 respectively. $r_3^s = 0.5$ (reflecting the 75% validity) during the presentation of the auditory
 566 cue in order to induce a top-down expectation of forthcoming visual stimuli. During the
 567 presentation of the visual stimuli, r_3^s changed based on Eq (3) to reconcile the prediction
 568 error. The aforementioned parameter settings were applied to all three models. The only
 569 parameter that varied across models was β , which was set to 0, 0.3, and -0.3 for the
 570 independence model (no mixing of PE), the reconciliation model (mixing of PE), and the
 571 segregation model (enhancing PE within each feature), respectively. This ensured that
 572 the bias due to different model implementation details in model comparison was
 573 minimized. Thus, the different model predictions can only be attributed to β , or how the
 574 two features exchange prediction errors. The simulation results are robust to
 575 perturbation of model parameters (Fig. 7), such that the magnitudes of a and β do not
 576 qualitatively change the pattern of simulation results. In accordance with the cross-
 577 subject MVPA approach that produced group-level results, we did not fit parameters to
 578 individual subjects. Instead, each model was run one time using the aforementioned
 579 parameters to simulate group-level results. The Matlab implementation of this
 580 framework and raw simulation results are available on request.

581 *Simulation procedure.* This $2 \times 2 \times 2$ factorial design was simulated using each of the
 582 three models. Because no randomness was introduced in the models, only one trial was

583 simulated for each condition. Within each trial, the auditory cue was simulated for 200
 584 time steps and the moving dots were simulated for 600 time steps to ensure a steady
 585 state was reached (e.g., θ_i^s converges to a minimum prediction error), in order to reflect
 586 that the subjects had learned the manipulations of expectation prior to the simulated
 587 tasks. The activity of r_i^s was estimated as its mean activity level over the last 10 time
 588 steps of the simulation to simulate the strength of representation.

589 **Results**

590 **Experiment 1**

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

599 **Model predictions**

600 All three models predicted that confirmed expectation in the relevant (attended) feature
 601 would facilitate performance (Fig. 1B). Crucially, the models' predictions diverged on the
 602 effect of expectation of the *unattended* feature on behavior. Specifically, the
 603 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 to the RT patterns of human participants judging expected versus unexpected attended features (collapsed across target feature). Employing 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 (that is, 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. 1C). This finding was consistent with all three models' predictions (Fig. 1B).

Crucially, we also observed a significant expectation gain effect in the unattended feature (10 ± 3 ms, $t_{16} = 2.93$, $P = 0.01$, Fig. 1C). This finding exclusively supports the reconciliation model (cf. Fig. 1B), 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. 1D) 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 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 non-target 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 non-target 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. Thus, 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 vs. 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 multi-voxel pattern analysis (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 3-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 3 factors. For example, for the color factor, we trained classifiers on color-expected vs. color-unexpected 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 motion-expected vs. motion-unexpected stimuli; and for the attention factor, we trained and tested on “attend color” vs. “attend motion” trials. These results are reported in the section entitled “Representation of feature-expectations in visual cortex”.

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. 2B-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. 2E). 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. 2F). 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. 3D-F). For this analysis, we divided our trials into 4 key conditions: (i) CU/MU, (ii) CE/MU, (iii) CU/ME and (iv) CE/ME, according to whether the color, the motion, both, or neither were 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. By contrast, the segregation model predicts the converse, namely that neural patterns associated with CU/ME and CE/MU conditions will become more dissimilar, and thus classifiers will distinguish these conditions better than CE/ME vs. CU/MU conditions. Finally, the independence model predicts that there will be no difference in classification accuracy between the CE/ME vs. CU/MU and CE/MU vs. CU/ME conditions. These analyses were all conducted after collapsing over the attention factor. 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,

715 collapsing across the attention factor. As can be seen in Fig. 4A, the results were
 716 similar to the qualitative predictions outlined in Fig. 3D-F.

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

729 This analysis revealed significant differences in classification accuracy in bilateral
 730 EVC (Fig. 4D, E, $P < 0.05$, corrected). Specifically, expectation-consistent classifiers
 731 (CU/MU vs. CE/ME, mean accuracy = 0.718, $P < 0.001$, binomial test, $n = 92$, or 23
 732 subjects × 2 classes × 2 attention conditions; Fig. 4F) outperformed expectation-
 733 inconsistent classifiers (CU/ME vs. CE/MU, mean accuracy = 0.492, n.s., binomial test,
 734 $n = 92$, Fig. 4F) in a large region of EVC (peaking at 9, -88, -2, Brodmann area 17). To
 735 further demonstrate that this effect cannot be solely explained by attention, we repeated
 736 this analysis separately on color- and motion-change runs. In the same EVC region (Fig.
 737 4F), CU/MU vs. 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 vs. 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 in accord with predictions from the reconciliation model (that is, a larger distance between consistent than inconsistent conditions, Fig. 4B), but not with those of the independence and segregation models (Fig. 4A, 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, 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 (that is, does the performance of the 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 (see Fig. 4G for the results in the aforementioned EVC region), thus corroborating our interpretation in line with the reconciliation model.

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

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

772 ***Attentional gain on feature representations in EVC depends on consistency of***
 773 ***feature expectations***

774 The effects of expectation on visual cognition are thought to interact with attention
 775 (Summerfield and Egner, 2009; Summerfield and de Lange, 2014). In this set of
 776 analyses, we therefore further tested whether the above findings can be solely
 777 attributed to attention, and assessed how well the rival multi-feature expectation models
 778 would be able to account for possible modulatory effects of (in)consistent feature
 779 expectations on the effects of feature-based attention. Specifically, for each color- ×
 780 motion-expectation condition (CU/MU, etc.), we extracted the attentional effect of each
 781 of the two features, defined as the (unsigned) difference of simulated activity between
 782 color-attended and motion-attended trials. This attentional effect on model activity

783 allowed us to estimate, in a monotonic fashion, the predicted neural dissimilarity
 784 between the two attentional conditions (attended vs. unattended) while keeping the
 785 expectation settings identical. For predictions about feature-selective visual areas (i.e.,
 786 model levels 2: simulated V4 and MT+), the attentional effect was computed separately
 787 for color and motion. For the model simulation of EVC, sensitive to both color and
 788 motion, the two features' attentional effects were summed. Note that the size of the
 789 attentional effect is positively correlated with the magnitude of simulated neural activity
 790 (i.e., encoding strength), because attention was modeled as a multiplicative gain
 791 modulator on simulated neural activity.

792 All three models generated qualitatively similar predictions for color- and motion-
 793 selective regions (Fig. 5A, B), whereby the attentional gain effect was larger when the
 794 preferred feature (e.g., color in the color stream) was expected than unexpected. These
 795 effects resemble the two-way interaction between attention and color-/motion-
 796 expectation. By contrast, the predictions of possible color- \times motion-expectation
 797 interaction effects on attentional gain were distinct between the three models at the
 798 level of EVC (Table 1, Fig. 5C), depicting different patterns of a 3-way interaction
 799 between attention, color-expectation and motion-expectation (with an emphasis on how
 800 the attentional effect is modulated by different combinations of color- and motion-
 801 expectancy). Specifically, the independence model predicted that the two feature
 802 expectations would independently modulate the multivariate effect of attention, due to
 803 no difference in neural representation strength among expectation conditions. The
 804 reconciliation model predicted that the attentional effects would be larger in expectation-
 805 consistent conditions (CU/MU and CE/MU) than expectation-inconsistent conditions

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

814 We next adjudicated between these model predictions using fMRI data. To this
 815 end, we constructed whole-brain searchlight-based, cross-subject attention classifiers
 816 (discriminating between attend-color- and attend-motion activation patterns) for each
 817 color- × motion-expectancy condition (e.g., CU/MU trials in color-target runs vs. CU/MU
 818 trials in motion-target runs). Note that because identical stimuli were used across the
 819 color- and motion-change detection runs, classification performance must reflect purely
 820 attentional effects. We then conducted a two-way ANOVA on the performance of these
 821 attention classifiers based on the 2 (color-expectation) × 2 (motion-expectation) design
 822 at each searchlight throughout the brain. We found a region in the anterior collateral
 823 sulcus (aCos) where color-attended and motion-attended trials evoked more dissimilar
 824 patterns of neural activity when color was expected than when it was unexpected (Fig.
 825 6A, $P < 0.05$, corrected). As expected, based on our study design, this region
 826 corresponds closely to color-sensitive cortex defined in previous studies (Cavina-Pratesi
 827 et al., 2010). We also detected a region in lateral occipital cortex where classifiers were
 828 better able to distinguish color-attended from motion-attended trials when motion

829 direction was expected than when it was unexpected (Fig. 6B, $P < 0.05$, corrected); this
 830 region corresponds closely to prior studies' localization of area MT+ (Rahnev et al.,
 831 2011). These findings were consistent with the activation predictions for feature-
 832 selective level 2 nodes of all three models (Fig. 5A, B).

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

851 **The patterns of simulation results only rely on the sign of β**

852 Finally, to show that the model predictions were not biased by the specific
 853 choices of model parameters, we ran the simulations with a wide range of attentional
 854 gain (α) and PE mixing (β) parameters and found that the qualitative pattern of
 855 simulation results (i.e., the sign of the effect of the unattended feature expectation in Fig.
 856 2B; whether expectation consistent classifiers outperform expectation inconsistent
 857 classifiers in Fig 4A-C; and the color- \times motion-expectation interaction pattern on
 858 attentional effects in Fig. 5C) only depended on the sign of β , which, by definition, was
 859 how the rival models are distinguished (Fig. 7).

860 **Univariate fMRI results**

861 To explore the relationship between the above multivariate results and mean signal
 862 neural strength in the corresponding visual regions, we conducted univariate analyses
 863 on the area-mean activity levels in these regions (Fig. 6A-C). First, because the rival
 864 hypotheses did not predict any difference between the color- and motion-stream level 2
 865 areas, we collapsed across the aCos (Fig. 6A) and MT+ (Fig. 6B) areas and performed
 866 a repeat measure 3-way ANOVA (attention \times preferred feature expectation \times non-
 867 preferred feature expectation, Fig. 6D). We found a significant main effect of attention
 868 ($F_{1,22} = 5.91$, $P < 0.05$), driven by higher activity level when the target feature was the
 869 preferred feature (0.15 ± 0.11) than the non-preferred feature (0.00 ± 0.10). This is
 870 consistent with the finding of increased neuronal firing rate driven by an attended
 871 stimulus (for review, see (Reynolds and Chelazzi, 2004)). We also observed a
 872 marginally significant attention-reversed expectation effect in the preferred feature ($F_{1,22}$
 873 $= 3.52$, $P = 0.07$) as reported in previous work (Kok et al., 2012b). We then conducted a
 874 repeat measure 3-way ANOVA (attention \times color expectation \times motion expectation, Fig.

6E) on the EVC area, and found a marginally significant main effect of attention ($F_{1,22} = 3.91$, $P < 0.06$), and a significant 3-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. 6C). Thus, while the univariate analyses – as expected *a priori* – were less sensitive in distinguishing the experimental conditions, the mean regional BOLD responses were in broadly in line 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 4 reported ROIs (early visual cortex (EVC) reported in Fig. 4E, anterior collateral sulcus (aCos) in Fig. 6A, MT+ in Fig. 6B, and EVC in Fig. 6C) to a null distribution of cluster sizes using the same voxel-wise 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 2-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 for 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.

897 4E: $P < 0.0001$, aCos: $P < 0.001$, MT+: $P < 0.0005$, EVC in Fig. 6C: $P < 0.0001$).

898 Therefore, our analysis approach was not prone to false positives.

899 Cross-subject MVPA requires that neural activity patterns are consistent across
 900 subjects. To gauge such consistency, we calculated the correlation of activity patterns
 901 between subjects. Specifically, this analysis was conducted separately for each of the 4
 902 reported ROIs. To further test if signal (as opposed to noise) exists at the level of single
 903 searchlights, for each searchlight in a given ROI, we calculated the difference of
 904 activation patterns between each pair of the 8 conditions in the experimental design,
 905 and computed the z-transformed correlation coefficients for each pair of subjects. The
 906 reason for using the difference of activation patterns between 2 conditions is to simulate
 907 the MVPAs. The z-values were then averaged across conditions, subjects and
 908 searchlights. The resulting mean z-value, which represents pattern consistency across
 909 subjects, was then compared to the mean z-values calculated using randomly
 910 scrambled data in the same ROI (repetition = 10,000 times). The results are
 911 summarized in Table 2. These data show that the univariate activity, which was used in
 912 MVPA, indeed contained signal patterns that were consistent across subjects and can
 913 be decoded using cross-subject MVPA.

914 The assumption of pattern consistency across subjects also predicts that the
 915 voxel-wise weights in the classifiers were preserved across subjects. To test this
 916 prediction, for each searchlight in each of the aforementioned 4 ROIs, we randomly split
 917 the subjects into 2 groups, calculated the voxel-wise weights of classifiers for each
 918 group, and tested the correlation of weights between groups. This procedure was
 919 repeated 100 times for each searchlight, and the mean z-transformed correlation

920 coefficients were used as a quantification of the preservation of voxel-wise weights in
 921 cross-subject MVPA. Due to the high computational cost, we here only computed
 922 contrasts that we reported as statistically significant in the manuscript. The ROI mean z-
 923 value was compared to z-values computed using randomly scrambled data of the same
 924 ROI (repetition = 1,000 times). The results are summarized in Table 3. As can be seen
 925 in Table 3, the obtained correlations in the empirical data were significantly greater than
 926 correlations generated from scrambled fMRI data (all P s < 0.001). Thus, these results
 927 clearly support that crucial assumption that the weights of classifiers were indeed
 928 preserved across subjects at the voxel level.

929 Even though the neural populations (e.g., cortical columns) calculating the
 930 prediction and prediction errors operate at a much finer spatial scale than the spatial
 931 resolution of fMRI, previous MVPA studies have shown that the voxel-level fMRI
 932 response is sensitive to changes in columnar level neural activity in the EVC and can
 933 hence be used to decode orientation in visual stimuli (Haynes and Rees, 2005; Kamitani
 934 and Tong, 2005). In the framework of predictive coding, the canonical microcircuits
 935 model (Bastos et al., 2012) ties the conceptual roles of computing prediction and
 936 prediction errors and the hierarchy of the predictive coding framework to the functions
 937 and connectivity of cortical columns. Following this logic, a match/mismatch between
 938 expectations and bottom-up input could lead to different columnar activity even for the
 939 same stimulus. Furthermore, given that columns are tuned to respond to different
 940 features (e.g., specific motion directions, specific colors), different columns may have
 941 different neural responses to the same stimulus. As a result, voxel-level fMRI activity
 942 may be modulated by the proportions of cortical columns it samples and by expectation.

943 Our control analyses showed consistent fMRI activity patterns across subjects (Table 2
944 and 3), which leads us to speculate that the distributions of columnar responses may
945 vary as a function of the spatial locations of the columns in the EVC at a spatial scale
946 similar to the spatial resolution of fMRI.

947 **Discussion**

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

965 The dual-prediction modeling framework developed here is grounded in basic
 966 tenets of the predictive coding (e.g., (Friston, 2005)) and attention literatures (e.g.,
 967 (Reynolds and Chelazzi, 2004)), as well as prior findings on the nature of color and
 968 motion processing in visual cortex (Gegenfurtner, 2003; Born and Bradley, 2005). The
 969 present fMRI data confirmed all of the key model assumptions, including the encoding
 970 of feature-selective color and motion expectations (Fig. 2B-D) in ventral and dorsal
 971 extrastriate visual cortex, respectively, paired with mixed selectivity for color- and
 972 motion-expectation (and their attentional modulation) in EVC. Moreover, all of the
 973 simulated neural activity patterns predicted by the reconciliation model (Table 1) were
 974 observed in fMRI activations patterns in the EVC (Fig. 2B-F, Fig. 4D-G, Fig. 6C). This is
 975 precisely consistent with our model implementation, where the cross-feature blending of
 976 PE occurs at the simulated EVC level, an assumption that was based on prior
 977 demonstrations that neurons in primary visual cortex are sensitive to both color and
 978 motion information (Movshon and Newsome, 1996; Engel et al., 1997; Johnson et al.,
 979 2001; Kamitani and Tong, 2006). At the microscopic level, this PE mixing in EVC could
 980 stem from an intermingling of parvocellular color-sensitive (Perry et al., 1984) and
 981 magnocellular motion-sensitive (Wiesel and Hubel, 1966) inputs from the lateral
 982 geniculate nucleus of the thalamus, which has been documented in previous studies of
 983 V1 (for review, see (Sincich and Horton, 2005)). While our model clearly represents a
 984 gross simplification of the rich interplay between early and later stages of the visual
 985 system, it successfully captured some basic neural population signatures of multi-
 986 feature expectations, while adhering to a biologically plausible architecture and
 987 processing principles.

988 Our main findings document that, rather than treating expectations concerning
 989 different object features as independent, or promoting the assumption that expected
 990 and unexpected features belong to different objects, the visual brain appears to
 991 exchange PE between visual features to form object-level expectations, such that
 992 surprise in one feature spreads to other features and ultimately renders the perception
 993 of all features of the object unexpected. The idea of object-level selection has a long
 994 history in the study of attention (Duncan, 1984), where a number of behavioral (e.g.,
 995 (Egley et al., 1994; He and Nakayama, 1995)) and neural studies (e.g., (Roelfsema et al.,
 996 1998; O'Craven et al., 1999)) have shown that attending to one location on, or feature of,
 997 an object confers an attentional advantage to other locations and features of that object.
 998 Importantly, the present data now show that objects, rather than single features or
 999 spatial locations, represent the default unit of selection not only for relevance-driven (i.e.,
 1000 attention) but also for probability-driven (i.e., expectation) endogenous determinants of
 1001 visual cognition. Furthermore, object-level selection implied by the reconciliation model
 1002 would also predict that the mixed PE should increase the similarity between the cue-
 1003 feature associations learned from different features. This similarity should in principle
 1004 also facilitate the learning of a unified cue-object association across trials. Future
 1005 studies are encouraged to test this prediction.

1006 Interestingly, our findings also document an interaction between expectation and
 1007 attention in the modulation of multi-feature processing. In particular, while attention
 1008 generally enhanced feature representations in higher visual regions (Fig. 6A, B) and in
 1009 expectation-consistent conditions in the EVC (Fig. 6C, CU/MU and CE/ME conditions),
 1010 this attentional modulation effect was absent in EVC for expectation-inconsistent

1011 conditions (Fig. 6C, CE/MU and CU/ME conditions). According to the reconciliation
 1012 model, this is because in expectation-inconsistent conditions PE mixing results in
 1013 attenuated neural feature representations (Table 1), which in turn dampens their
 1014 attentional modulation. On the other hand, the attention-modulated PE enters the PE
 1015 mixing process and spreads to unattended features associated with the same object. In
 1016 other words, PE mixing also transfers the attentional modulation to unattended features,
 1017 which is again consistent with the above-mentioned spreading of attention across object
 1018 features.

1019 While our study and model were designed to focus on how object-level
 1020 expectation is implemented in visual cortical processing of individual features, an
 1021 important question to ask is where the belief that these features belong to the same
 1022 object might originate. Possible answers to this question may be found in the literature
 1023 on feature binding (or “feature integration”), which has long been considered integral to
 1024 object perception (Treisman and Gelade, 1980; Treisman, 1998) and proposed to be an
 1025 obligatory operation in human cognition (Ashby et al., 1996; Hommel, 2004). Prior
 1026 lesion and neuroimaging studies have observed involvement of parietal cortex
 1027 (Treisman, 1998) as well as of both classic learning systems of the brain, the medial
 1028 temporal lobe/ hippocampus (Mitchell et al., 2000; Jiang et al., 2015), and the striatum
 1029 (Jiang et al., 2015) in the perceptual and mnemonic binding of different event features.
 1030 These regions therefore constitute prime candidates for generating the integrated,
 1031 object-level predictions that drive the effects we here documented in visual cortex;
 1032 assessing the exact mechanisms by which these or other more anterior regions (e.g.,

1033 hippocampus, see (Hindy et al., 2016)) impose top-down object-level expectations
 1034 represents a key goal for future studies.

1035 Given the close relationship between attention and expectation (Summerfield and
 1036 Egner, 2009, 2016), we took several measures to ensure that the present results are
 1037 not due to attentional mechanisms. First, in the experimental design, attention and
 1038 expectation were dissociated. Second, we conducted the key analysis that compared
 1039 expectation-consistent and expectation-inconsistent classifiers (Fig. 4D-F) by collapsing
 1040 across color- and motion-target trials *and* performing this analysis on these two types of
 1041 trials separately. All three analyses revealed the same results, thus strongly suggesting
 1042 that attention to target features cannot account for the current results. Third, we tested
 1043 whether a hypothesis that one unexpected feature attracts attention can explain some
 1044 of the results. This hypothesis, along with the findings of a significant main effect of
 1045 attention in the EVC, would predict significantly distinct fMRI activity patterns between
 1046 CU/ME and CE/MU trials as a result of an attentional effect (i.e., attention was attracted
 1047 to color and motion in CU/ME and CE/MU trials, respectively). However, the CU/ME vs.
 1048 CE/MU classifiers did not perform above chance level (Fig. 4F). Moreover, we
 1049 conducted another analysis that directly contradicted this hypothesis by showing a
 1050 significant attentional effect on EVC neural activity patterns in CU/MU trials (Fig. 6C),
 1051 which would not be expected to show attentional effects under this hypothesis. Fourth,
 1052 another alternative hypothesis could be that violation of prediction in any feature would
 1053 result in re-allocation of attention to both features. Assuming that the BOLD signal
 1054 reflects a joint effect of feature-based attention from task instruction and the
 1055 redistribution of attention due to high prediction error, 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. 6C). In general, compared to 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. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30:2960-2966.
- Ashby FG, Prinzmetal W, Ivry R, Maddox WT (1996) A formal theory of feature binding in object perception. *Psychological review* 103:165-192.
- Auksztulewicz R, Friston K (2015) Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. *Cerebral cortex* 25:4273-4283.
- Bar M (2004) Visual objects in context. *Nature reviews Neuroscience* 5:617-629.
- Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical microcircuits for predictive coding. *Neuron* 76:695-711.
- Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: detecting and judging objects undergoing relational violations. *Cognitive psychology* 14:143-177.

- 1084 Born RT, Bradley DC (2005) Structure and function of visual area MT. *Annual review of neuroscience*
1085 28:157-189.
- 1086 Brainard DH (1997) The Psychophysics Toolbox. *Spatial vision* 10:433-436.
- 1087 Cavina-Pratesi C, Kentridge RW, Heywood CA, Milner AD (2010) Separate channels for processing form,
1088 texture, and color: evidence from fMRI adaptation and visual object agnosia. *Cerebral cortex*
1089 20:2319-2332.
- 1090 Clithero JA, Smith DV, Carter RM, Huettel SA (2011) Within- and cross-participant classifiers reveal
1091 different neural coding of information. *NeuroImage* 56:699-708.
- 1092 Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughhead JW, Gur RC, Langleben DD (2005)
1093 Classifying spatial patterns of brain activity with machine learning methods: application to lie
1094 detection. *NeuroImage* 28:663-668.
- 1095 de Gardelle V, Waszczuk M, Egner T, Summerfield C (2013) Concurrent repetition enhancement and
1096 suppression responses in extrastriate visual cortex. *Cerebral cortex* 23:2235-2244.
- 1097 den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in
1098 associative learning. *Cerebral cortex* 19:1175-1185.
- 1099 Duncan J (1984) Selective attention and the organization of visual information. *Journal of experimental*
1100 *psychology General* 113:501-517.
- 1101 Egly R, Driver J, Rafal RD (1994) Shifting visual attention between objects and locations: evidence from
1102 normal and parietal lesion subjects. *Journal of experimental psychology General* 123:161-177.
- 1103 Egner T, Monti JM, Summerfield C (2010a) Expectation and surprise determine neural population
1104 responses in the ventral visual stream. *The Journal of neuroscience : the official journal of the*
1105 *Society for Neuroscience* 30:16601-16608.
- 1106 Egner T, Ely S, Grinband J (2010b) Going, going, gone: characterizing the time-course of congruency
1107 sequence effects. *Frontiers in psychology* 1:154.
- 1108 Engel S, Zhang X, Wandell B (1997) Colour tuning in human visual cortex measured with functional
1109 magnetic resonance imaging. *Nature* 388:68-71.
- 1110 Feldman H, Friston KJ (2010) Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*
1111 4:215.
- 1112 Friston K (2005) A theory of cortical responses. *Philosophical transactions of the Royal Society of London*
1113 *Series B, Biological sciences* 360:815-836.
- 1114 Friston K (2010) The free-energy principle: a unified brain theory? *Nature reviews Neuroscience* 11:127-
1115 138.
- 1116 Gegenfurtner KR (2003) Cortical mechanisms of colour vision. *Nature reviews Neuroscience* 4:563-572.
- 1117 Gordon AM, Rissman J, Kiani R, Wagner AD (2014) Cortical reinstatement mediates the relationship
1118 between content-specific encoding activity and subsequent recollection decisions. *Cerebral*
1119 *cortex* 24:3350-3364.
- 1120 Grill-Spector K, Malach R (2004) The human visual cortex. *Annual review of neuroscience* 27:649-677.
- 1121 Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary
1122 visual cortex. *Nature neuroscience* 8:686-691.
- 1123 He ZJ, Nakayama K (1995) Visual attention to surfaces in three-dimensional space. *Proceedings of the*
1124 *National Academy of Sciences of the United States of America* 92:11155-11159.
- 1125 Hindy NC, Ng FY, Turk-Browne NB (2016) Linking pattern completion in the hippocampus to predictive
1126 coding in visual cortex. *Nature neuroscience* 19:665-667.
- 1127 Hommel B (2004) Event files: feature binding in and across perception and action. *Trends in cognitive*
1128 *sciences* 8:494-500.
- 1129 Jiang J, Schmajuk N, Egner T (2012) Explaining neural signals in human visual cortex with an associative
1130 learning model. *Behav Neurosci* 126:575-581.

- 1131 Jiang J, Summerfield C, Egner T (2013) Attention sharpens the distinction between expected and
 1132 unexpected percepts in the visual brain. *The Journal of neuroscience : the official journal of the*
 1133 *Society for Neuroscience* 33:18438-18447.
- 1134 Jiang J, Brashier NM, Egner T (2015) Memory Meets Control in Hippocampal and Striatal Binding of
 1135 Stimuli, Responses, and Attentional Control States. *The Journal of neuroscience : the official*
 1136 *journal of the Society for Neuroscience* 35:14885-14895.
- 1137 Johnson EN, Hawken MJ, Shapley R (2001) The spatial transformation of color in the primary visual
 1138 cortex of the macaque monkey. *Nature neuroscience* 4:409-416.
- 1139 Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nature*
 1140 *neuroscience* 8:679-685.
- 1141 Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human
 1142 visual cortex. *Current biology : CB* 16:1096-1102.
- 1143 Kanai R, Komura Y, Shipp S, Friston K (2015) Cerebral hierarchies: predictive processing, precision and
 1144 the pulvinar. *Philosophical transactions of the Royal Society of London Series B, Biological*
 1145 *sciences* 370.
- 1146 Kaplan JT, Meyer K (2012) Multivariate pattern analysis reveals common neural patterns across
 1147 individuals during touch observation. *NeuroImage* 60:204-212.
- 1148 Keller GB, Bonhoeffer T, Hubener M (2012) Sensorimotor mismatch signals in primary visual cortex of
 1149 the behaving mouse. *Neuron* 74:809-815.
- 1150 Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. *Annu Rev Psychol*
 1151 55:271-304.
- 1152 Kok P, Jehee JF, de Lange FP (2012a) Less is more: expectation sharpens representations in the primary
 1153 visual cortex. *Neuron* 75:265-270.
- 1154 Kok P, Rahnev D, Jehee JF, Lau HC, de Lange FP (2012b) Attention reverses the effect of prediction in
 1155 silencing sensory signals. *Cerebral cortex* 22:2197-2206.
- 1156 Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proceedings*
 1157 *of the National Academy of Sciences of the United States of America* 103:3863-3868.
- 1158 Martinez-Trujillo JC, Treue S (2004) Feature-based attention increases the selectivity of population
 1159 responses in primate visual cortex. *Current biology : CB* 14:744-751.
- 1160 Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex.
 1161 *Proceedings of the National Academy of Sciences of the United States of America* 108:19401-
 1162 19406.
- 1163 Mitchell KJ, Johnson MK, Raye CL, D'Esposito M (2000) fMRI evidence of age-related hippocampal
 1164 dysfunction in feature binding in working memory. *Brain Res Cogn Brain Res* 10:197-206.
- 1165 Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang XR, Just M, Newman S (2004) Learning to
 1166 decode cognitive states from brain images. *Machine Learning* 57:145-175.
- 1167 Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M (2005) Classifying brain states and
 1168 determining the discriminating activation patterns: Support Vector Machine on functional MRI
 1169 data. *NeuroImage* 28:980-995.
- 1170 Movshon JA, Newsome WT (1996) Visual response properties of striate cortical neurons projecting to
 1171 area MT in macaque monkeys. *The Journal of neuroscience : the official journal of the Society*
 1172 *for Neuroscience* 16:7733-7741.
- 1173 Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical
 1174 loops. *Biol Cybern* 66:241-251.
- 1175 Nassar MR, Wilson RC, Heasly B, Gold JI (2010) An approximately Bayesian delta-rule model explains the
 1176 dynamics of belief updating in a changing environment. *The Journal of neuroscience : the official*
 1177 *journal of the Society for Neuroscience* 30:12366-12378.

- 1178 Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI (2012) Rational regulation of learning
 1179 dynamics by pupil-linked arousal systems. *Nature neuroscience* 15:1040-1046.
- 1180 O'Craven KM, Downing PE, Kanwisher N (1999) fMRI evidence for objects as the units of attentional
 1181 selection. *Nature* 401:584-587.
- 1182 Onat S, Buchel C (2015) The neuronal basis of fear generalization in humans. *Nature neuroscience*
 1183 18:1811-1818.
- 1184 Palmer TE (1975) The effects of contextual scenes on the identification of objects. *Mem Cognit* 3:519-
 1185 526.
- 1186 Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview.
 1187 *NeuroImage* 45:S199-209.
- 1188 Perry VH, Oehler R, Cowey A (1984) Retinal ganglion cells that project to the dorsal lateral geniculate
 1189 nucleus in the macaque monkey. *Neuroscience* 12:1101-1123.
- 1190 Poldrack RA, Halchenko YO, Hanson SJ (2009) Decoding the large-scale structure of brain function by
 1191 classifying mental States across individuals. *Psychological science* 20:1364-1372.
- 1192 Rahnev D, Lau H, de Lange FP (2011) Prior expectation modulates the interaction between sensory and
 1193 prefrontal regions in the human brain. *The Journal of neuroscience : the official journal of the*
 1194 *Society for Neuroscience* 31:10741-10748.
- 1195 Rao RP (2005) Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16:1843-
 1196 1848.
- 1197 Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some
 1198 extra-classical receptive-field effects. *Nature neuroscience* 2:79-87.
- 1199 Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. *Psychological science*
 1200 9:347-356.
- 1201 Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. *Annual review of*
 1202 *neuroscience* 27:611-647.
- 1203 Reynolds JH, Pasternak T, Desimone R (2000) Attention increases sensitivity of V4 neurons. *Neuron*
 1204 26:703-714.
- 1205 Roelfsema PR, Lamme VA, Spekreijse H (1998) Object-based attention in the primary visual cortex of the
 1206 macaque monkey. *Nature* 395:376-381.
- 1207 Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA (2011) Commonality of neural
 1208 representations of words and pictures. *NeuroImage* 54:2418-2425.
- 1209 Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA (2008) Using FMRI brain activation
 1210 to identify cognitive states associated with perception of tools and dwellings. *PLoS one* 3:e1394.
- 1211 Sincich LC, Horton JC (2005) The circuitry of V1 and V2: integration of color, form, and motion. *Annual*
 1212 *review of neuroscience* 28:303-326.
- 1213 Spratling MW (2008) Reconciling predictive coding and biased competition models of cortical function.
 1214 *Frontiers in computational neuroscience* 2:4.
- 1215 Summerfield C, Egnér T (2009) Expectation (and attention) in visual cognition. *Trends in cognitive*
 1216 *sciences* 13:403-409.
- 1217 Summerfield C, de Lange FP (2014) Expectation in perceptual decision making: neural and
 1218 computational mechanisms. *Nature reviews Neuroscience* 15:745-756.
- 1219 Summerfield C, Egnér T (2016) Feature-Based Attention and Feature-Based Expectation. *Trends in*
 1220 *cognitive sciences*.
- 1221 Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egnér T (2008) Neural repetition suppression
 1222 reflects fulfilled perceptual expectations. *Nature neuroscience* 11:1004-1006.
- 1223 Thiele A, Pooremaeili A, Delicato LS, Herrero JL, Roelfsema PR (2009) Additive effects of attention and
 1224 stimulus contrast in primary visual cortex. *Cerebral cortex* 19:2970-2981.

- 1225 Treisman A (1998) Feature binding, attention and object perception. Philosophical transactions of the
1226 Royal Society of London Series B, Biological sciences 353:1295-1306.
- 1227 Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognitive psychology 12:97-
1228 136.
- 1229 Wacongne C, Changeux JP, Dehaene S (2012) A neuronal model of predictive coding accounting for the
1230 mismatch negativity. The Journal of neuroscience : the official journal of the Society for
1231 Neuroscience 32:3665-3678.
- 1232 Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R (2012) A century of
1233 Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization.
1234 Psychological bulletin 138:1172-1217.
- 1235 Wiesel TN, Hubel DH (1966) Spatial and chromatic interactions in the lateral geniculate body of the
1236 rhesus monkey. J Neurophysiol 29:1115-1156.
- 1237 Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681-692.

1238

1239

1240 **Tables**

	How PE in feature A affects feature B	Distinction between EI, compared to 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

1241

1242 Table 1. Summary of key predictions from three different models of multi-feature expectations in visual
 1243 object cognition. EI: expectation inconsistent conditions. EC: expectation consistent conditions.

1244

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

1245 Table 2. Results of cross-subject fMRI activity pattern consistency.

1246

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

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

Figure legends

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 (C) RT and (D) accuracy in experiment 1, plotted as a function of whether the feature in question was attended (i.e., was the current target feature).

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 500ms. Subjects were required to identify the post-change feature. (B-F) Lateral, posterior and dorsal views of brain areas showing significant (in red; binomial tests, $P < 0.05$, corrected) performance for (B) attentional classifiers, (C) color-expectation classifiers, (D) motion-expectation classifiers, (E) expectation of attended feature classifiers, and (F) expectation of unattended feature classifiers.

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 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 grey dotted lines reflect the CU/MU-CE/ME distance and the CU/ME-CE/MU distance. (G) The structure of the predictive coding implementation of the conceptual models (same structure for all three models). This implementation consists of two visual processing streams (upper: motion stream, lower: 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.

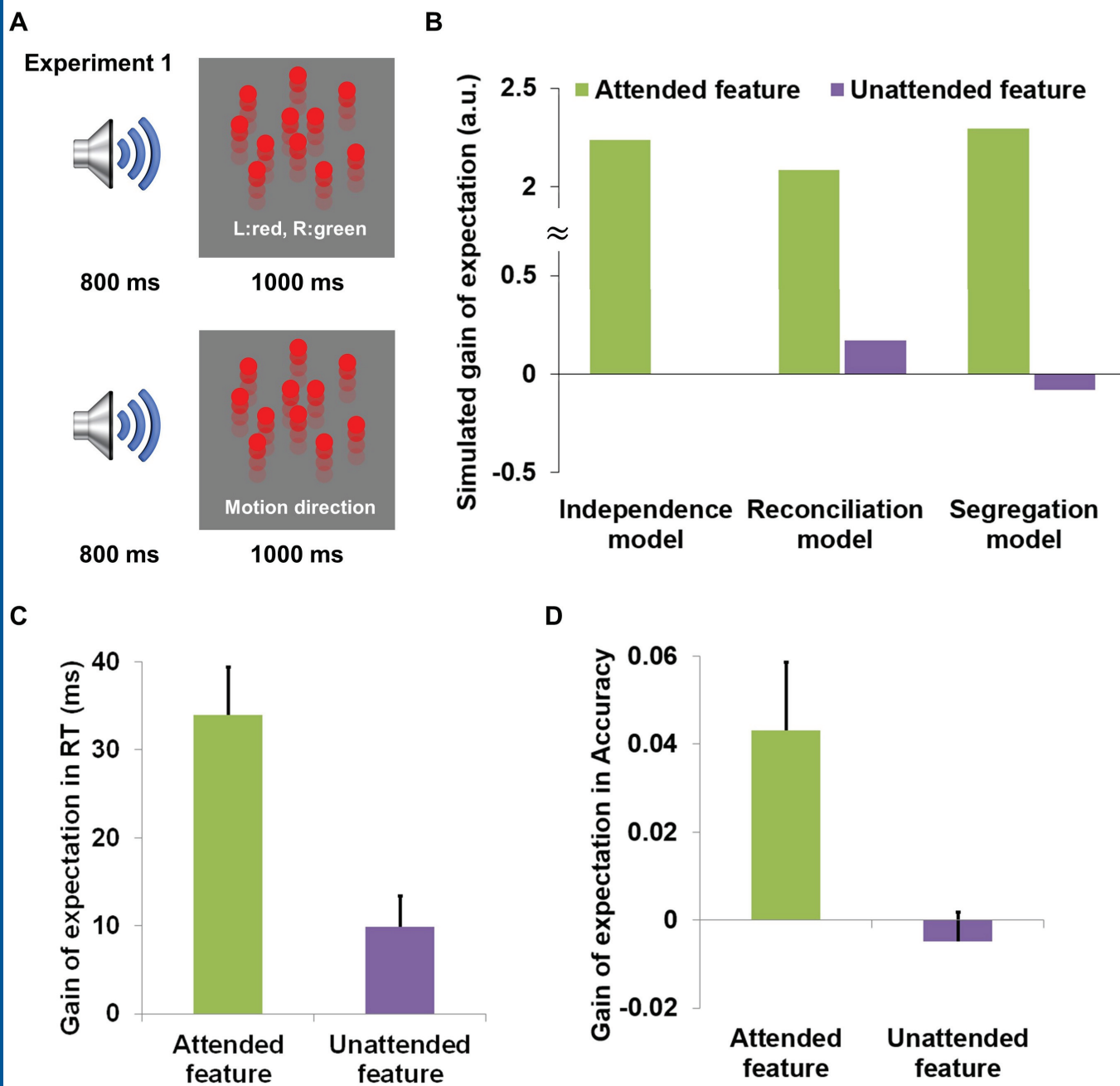
Figure 4. Joint effects of color- and motion-expectation in simulation and fMRI data. (A-C) Simulation results of the distances of r unit activity within expectation consistent and expectation inconsistent conditions at model level 1 (EVC, see Fig. 1A), using the independence model, reconciliation model, and segregation model, respectively. (D) From 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) An 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- \times 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- \times 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- \times motion-expectation. The panels, from left to right, show 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.

Figure 6. The effects of multi-feature expectation on the accuracy of attention classifiers in visual cortex. (A) Left: A cluster of searchlights in the right anterior collateral sulcus 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: A 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: A 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 (\pm 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

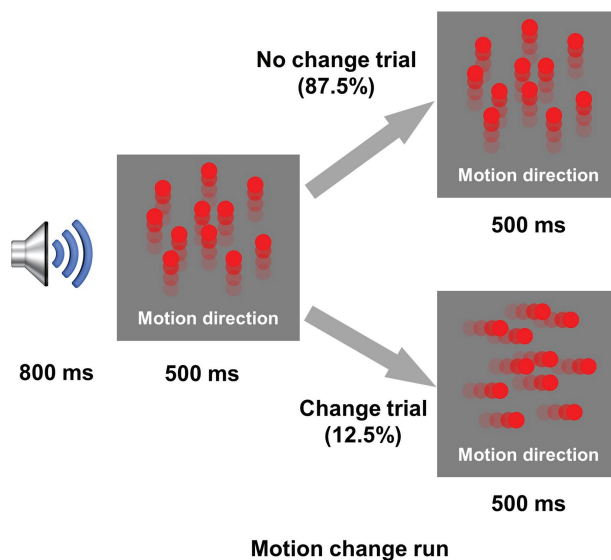
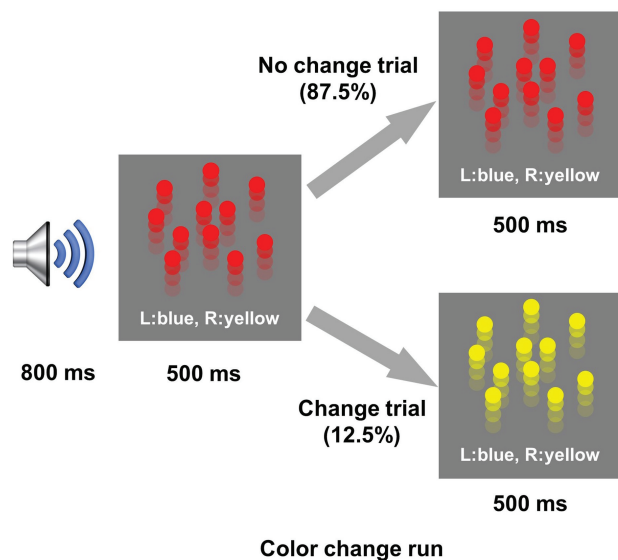
target feature, and the expectation of preferred and non-preferred features. (E) Mean fMRI activation level (\pm MSE) of the area shown in (C), plotted as a function of target feature, color-, and motion-expectation.

Figure 7. The simulation results are only sensitive to the sign of β . Each heatmap visualizes an effect that has divergent model predictions (from left to right: the expectation effect on the unattended feature shown in Fig. 1B; the distance between expectation-consistent conditions minus the distance between expectation-inconsistent conditions shown in Fig. 4A-C; and the interaction of the two features' expectation on attentional effect shown in Fig. 5C). 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 heatmap represents the result from a model, whose 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 indicates the magnitude of the effect. In all three heatmaps, 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.

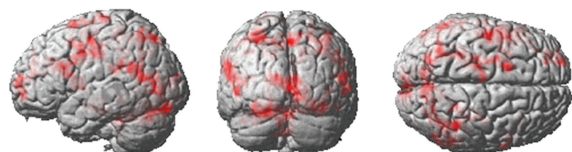


A

Experiment 2

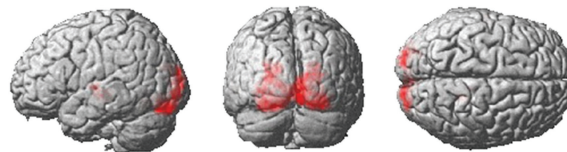


B



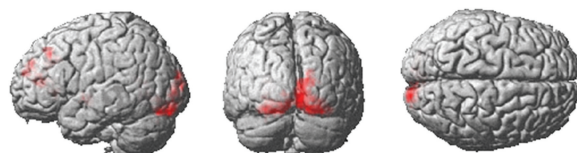
Attended to color vs. attended to motion

E



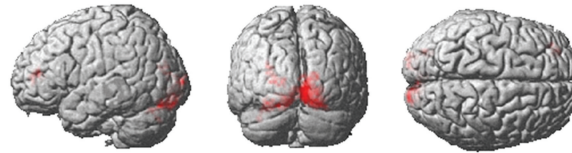
Expected vs. unexpected of the attended feature

C



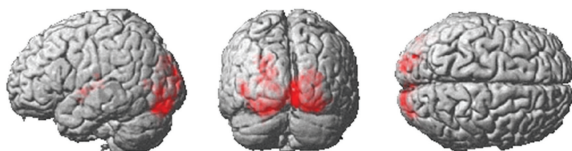
Color expected vs. color unexpected

F



Expected vs. unexpected of the unattended feature

D



Motion expected vs. motion unexpected

