OXFORD

Cerebral Cortex, 2023, **33**, 8300–8311

https://doi.org/10.1093/cercor/bhad115

Advance access publication date 1 April 2023

Original Article

Humans predict the forest, not the trees: statistical learning of spatiotemporal structure in visual scenes

Chuyao Yan[1,2,*], Benedikt V. Ehinger[1,3], Alexis Pérez-Bellido[1,4,5], Marius V. Peelen[1], Floris P. de Lange[1]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, Nijmegen 6525 EN, The Netherlands,
[2]School of Psychology, Nanjing Normal University, Nanjing 210098, China,
[3]Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart 70049, Germany,
[4]Department of Cognition, Development and Educational Psychology, University of Barcelona, Barcelona 17108035, Spain,
[5]Institute of Neurosciences, University of Barcelona, Barcelona 17108035, Spain

*Corresponding author: Chuyao Yan, School of Psychology, Nanjing Normal University, Nanjing 210098, China. Email: yanchuyao@gmail.com

The human brain is capable of using statistical regularities to predict future inputs. In the real world, such inputs typically comprise a collection of objects (e.g. a forest constitutes numerous trees). The present study aimed to investigate whether perceptual anticipation relies on lower-level or higher-level information. Specifically, we examined whether the human brain anticipates each object in a scene individually or anticipates the scene as a whole. To explore this issue, we first trained participants to associate co-occurring objects within fixed spatial arrangements. Meanwhile, participants implicitly learned temporal regularities between these displays. We then tested how spatial and temporal violations of the structure modulated behavior and neural activity in the visual system using fMRI. We found that participants only showed a behavioral advantage of temporal regularities when the displays conformed to their previously learned spatial structure, demonstrating that humans form configuration-specific temporal expectations instead of predicting individual objects. Similarly, we found suppression of neural responses for temporally expected compared with temporally unexpected objects in lateral occipital cortex only when the objects were embedded within expected configurations. Overall, our findings indicate that humans form expectations about object configurations, demonstrating the prioritization of higher-level over lower-level information in temporal expectation.

Key words: perception; temporal expectation; statistical learning; spatial arrangement; expectation suppression.

Introduction

A growing body of research has shown that the brain does not passively wait to be activated by sensory inputs but instead actively anticipates future input (Clark 2013). This process is achieved through continuous extraction of statistical regularities from the visual world, which in turn facilitate perceptual processing and object categorization (Biederman et al. 1982; Chun and Jiang 1999; Fiser and Aslin 2001; Green and Hummel 2006; Boettcher et al. 2020).

Whereas most studies on sensory expectation have focused on investigating how the brain predicts single objects (Meyer and Olson 2011; Manahova et al. 2018; Richter et al. 2018), objects rarely occur in isolation in the real world. They appear alongside other objects within specific spatial arrangements, posing a challenge for the visual system to understand complex cluttered visual scenes. For example, a keyboard and a mouse pad positioned in front of a monitor may be perceived as a single "desktop" percept, or a group of trees may be perceived as a "forest." Previous studies have shown that the visual system is sensitive to the familiar spatial arrangement of objects (Gronau et al. 2008; Kaiser and Peelen 2018). Furthermore, studies have demonstrated that statistical learning can extract the co-occurrence statistics of stimuli to represent multiple objects as a higher-order representation (Fiser and Aslin 2001, 2005; Orbán et al. 2008; Lengyel et al. 2019). Such hierarchical processing allows the same set of objects to be described both at the level of the whole (forest) and at the level of the individual parts (trees). These findings raise the question of at which level of the stimulus hierarchy temporal expectation takes place, specifically, whether the expectation of future input represents the collection of objects as a higher-order representation of the whole, or whether local information about individual items is predicted.

In the present study, we set out to investigate this issue. To this end, we exposed participants to 8 structured sets of 4 co-occurring objects. Initially, these objects were unrelated to each other and were learned by the participants in a fixed spatial arrangement. Based on prior research (Fiser and Aslin 2001; Orbán et al. 2008; Stansbury et al. 2013), we expected participants to learn the statistical regularities of the spatial arrangement and represent these objects as a scene-like representation. We then presented the displays in a predetermined sequence (Fig. 1) and manipulated the temporal and spatial aspects of the displays while recording the BOLD response. Specifically, keeping the temporal aspects at first constant, we introduced manipulations to spatial aspects so that familiar displays were followed by either (i) "familiar" displays; (ii) "mixed" displays (half of the items are from the temporally expected display, the other half from a different familiar display); or (iii) "shuffled" displays (items of the familiar display but at shuffled positions). To further manipulate the temporal aspect, we presented these displays at either a temporally expected or unexpected sequence position. We capitalized on the concept of expectation suppression to investigate the brain regions that

Received: December 16, 2022. **Revised:** March 11, 2023. **Accepted:** March 13, 2023

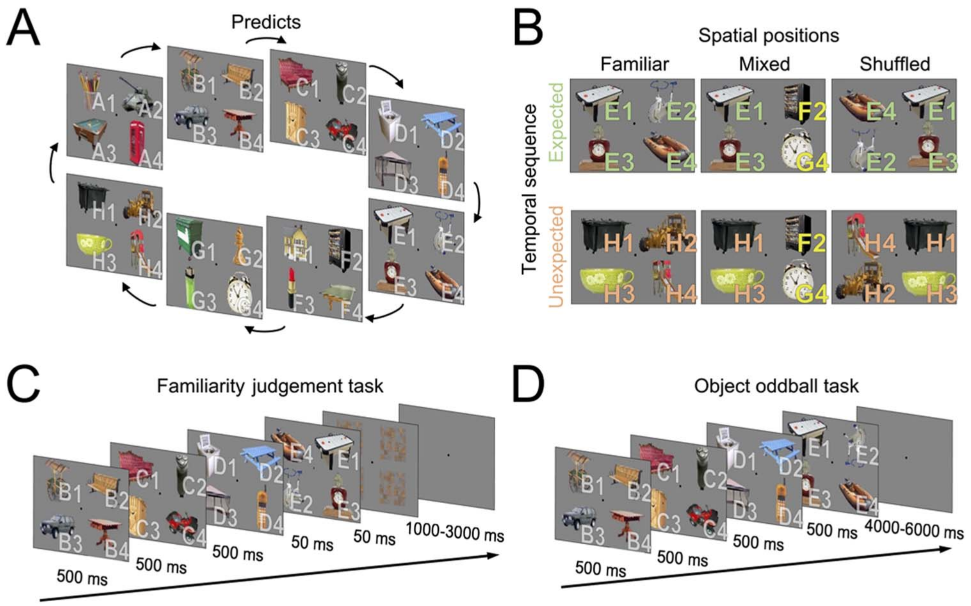

Fig. 1. A) Example stimuli. For each participant, 32 object images were (randomly) grouped into 8 specific displays. These displays were then arranged in a specific temporal sequence; thus, each display could predict the identity of the next. Note that each object is visualized here with an additional index which was not presented in the experiment. The displays are labeled as letters (from A to H), while the positions of objects are labeled as numbers (from 1 to 4). B) The 3-by-2 factorial design of an example target display that following the display F in a sequence. The target display could be familiar, mixed, or shuffled, and it could also be temporally expected (display E) or unexpected (display H), resulting in 6 conditions. C) A single trial of the familiarity judgment task. Four displays were presented sequentially. The first 3 displays were always familiar and could predict the next, lasting 500 ms each. The fourth display (the target) appeared for 50 ms, followed by a 50 ms mask. Each trial ends with a 1,000–3,000 ms ITI. Participants' task was to categorize if the target display was a familiar or novel display. In this example, the target is a novel (shuffled) display. D) A single trial of the object oddball task. Four displays were presented sequentially for 500 ms each, ending with a 4,000–6,000 ms ITI. The first 3 displays were always familiar and predictable. Participants' task was to detect an inverted object in the target display.

are sensitive to statistical irregularities in stimuli. Expectation suppression refers to a reduction in neural activity that typically follows the presentation of an input that is expected compared to the same input when it is not expected (Summerfield et al. 2008; Egner et al. 2010; Kaposvari et al. 2018; Richter et al. 2018).

We hypothesized that if temporal predictions rely on a high-level representation that retains information about spatial structure, we would only observe expectation suppression when comparing expected and unexpected "familiar" displays. Additionally, if expectation suppression is locally induced by each object, regardless of its spatial context, we would observe expectation suppression for "mixed" displays as well. Specifically, we expected to observe these suppression effects for the expected compared to unexpected "individual items" within the "mixed" displays. Finally, if the observer's temporal expectations only reflect the object identity, we would anticipate observing a similar level of expectation suppression effects when comparing expected and unexpected conditions in "familiar" and "shuffled" displays.

To preview the results, we found that participants only showed a behavioral advantage of temporal regularities when the expected objects were arranged within the learned ("familiar") displays. Additionally, we only observed activity suppression for temporally expected displays in lateral occipital cortex (LOC) for the previously learned displays but not for rearranged ("mixed" and "shuffled") displays. Importantly, no expectation suppression was observed for temporally expected compared to unexpected individual objects when these objects were no longer part of a learned display. Overall, our findings suggest that expectations for future object sets reflect the higher-level spatial structure in which objects are encompassed.

Materials and methods

Preregistration and data availability

The current study was preregistered at Open Science Framework before any data were acquired. The preregistration form is available at DOI 10.17605/OSF.IO/8QCDX. All procedures in the preregistration document were followed unless specified otherwise in the sections below. All data and code are openly available at the Donders Institute for Brain, Cognition and Behaviour repository (https://data.donders.ru.nl/).

Participants and data exclusion

Thirty-five healthy human participants were recruited through the Radboud Research Participation System and received monetary compensation. One participant was excluded because of excessive head motion during scanning. All remaining participants ($n = 34$, 27 females, 1 left-handed, age = 24.2 ± 4.4 years) were included in all analyses, which is in line with our preregistered goal to achieve a sample size of $n = 34$ to detect an effect size of Cohen's $d \geq 0.5$ with 80% power using a 2-tailed within-subjects t-test. This study was approved by the local ethics committee (CMO Arnhem-Nijmegen, Radboud University Medical Center). Informed written consent was obtained before the experiment. All participants were prescreened for MRI compatibility and had normal or corrected-to-normal vision.

Stimuli

Stimuli were presented using PsychToolbox (Brainard 1997; Pelli 1997; Kleiner et al. 2007) running on MATLAB R2017b (The MathWorks, Natick, MA, United States). A subset of 32 natural object stimuli were used in the current study taken from

Richter et al. (2018), which were adapted from Brady et al. (2008). The 32 object images were randomly assigned to 8 groups for each participant (4 objects per group). These 8 groups were then arranged sequentially so that each group temporally predicts the next group (see Fig. 1A). The 4 object images (4° × 4°) forming each group were positioned at the corners of a squared template (6° × 6°), centered on the middle of the screen. We employed 3 types of displays to manipulate the spatial relationships between co-occurring objects: familiar, mixed, and shuffled (see Fig. 1B). In the familiar display, the 4 object images were positioned in the same spatial arrangement as in most of the presentations. In the mixed display, 2 objects from the familiar display were randomly replaced by objects from different familiar displays, with the object positions remaining unaltered. In the shuffled display, the 4 objects were identical to those in the familiar display, but their positions were shuffled. Importantly, these 3 display types could either be temporally predictable based on the preceding display or unpredictable. For example, as illustrated in Fig. 1B, the display following the previous display D could be either expected (display E) or unexpected (display H). In the familiar condition, the entire display could either be temporally expected or unexpected. In the mixed condition, half of the objects could either be temporally expected or unexpected (e.g. E1 and E3 were expected, while H1 and H3 were unexpected), and the other half is always unexpected. In the shuffled condition, the identities of all objects could either be temporally expected or unexpected regardless of their spatial structure. These manipulations enabled us to investigate whether the effects of temporal expectation are generated by higher-order representations that integrate information about both object identities and their spatial relationships by comparing the temporally expected and unexpected conditions within all 3 types of spatial displays. All object stimuli were presented in full color on a midgray background, with a centered fixation point (0.5° visual angle in size). The fixation point was formed by a bulls-eye combined with a crosshair to improve stable fixation (Thaler et al. 2013). For the behavioral training and testing, stimuli were presented on an LCD screen (BenQ XL2420T, 1,920 × 1,080 pixel resolution, 60 Hz refresh rate). For the MRI scanning, stimuli were rear-projected on an MRI-compatible screen using EIKI LC XL100 beamer (resolution: 1,024 × 768; refresh rate: 60 Hz), which was visible through an adjustable mirror.

Procedure and design

The experiment consisted of 2 sessions over 2 consecutive days. The first day consisted of a behavioral session, including a training and a familiarity judgment task. On the second day, participants underwent an MRI session, including an object oddball task and a functional localizer (also referred to as fixation brightness task). For each participant, the same stimuli groups were used throughout the different tasks.

Training

The aim of the training was to allow the participants to explicitly learn the scene-like displays and to implicitly learn the temporal sequence. The training consisted of 2 parts.

First, we familiarized the participants with 8 sequentially presented displays of objects (familiar displays) without assigning any task. These 8 familiar displays and sequential orders were randomly generated for each participant. Participants were not informed about the presence of any temporal statistical regularities. The task only required them to examine each display and move on to the next one with a key press in their own time. These displays were presented in sequential order and repeated consecutively, resulting in 200 trials in total, lasting from 15 to 30 min depending on participants' speed.

Second, we trained participants to explicitly recognize the familiar displays and differentiate them from novel ones (i.e. mixed and shuffled). We presented the familiar displays in the same sequential order as in the first part of the training, but we introduced a 13.33% chance to show a novel display instead of the familiar one: In some cases, we randomly replaced 2 objects within the display by 2 other objects from another 2 displays (mixed display); while in other cases, we randomly shuffled the position of 4 objects within a familiar display (shuffled display). Participants were instructed to indicate whether the display was familiar or novel using 2 alternative keys. After each catch trial (novel display), we restarted the sequence at a random starting point. Feedback on behavioral accuracy was provided at the end of each trial by changing the color of the fixation point to green (correct) or red (incorrect) depending on button responses. There were 900 trials split into 5 runs of equal length, which lasted approximately 40 min.

Familiarity judgment task

The training tasks were followed by a familiarity judgment task to test whether the participants had implicitly learned the temporal sequence and used it to form expectations of upcoming stimuli. In each trial, 4 displays were presented sequentially on the screen. The first 3 displays were always familiar, lasting for 500 ms each and always following the temporal sequence presented during the training phase, while the last display was visible for only 50 ms and followed by a phase-scrambled mask that was also presented for 50 ms (see Fig. 1C). Such a short presentation time and the mask were used to prevent a ceiling effect in accuracy and encourage the reliance on internal predictive processes. After each trial, there was an intertrial interval of 1,000–3,000 ms. During the experiment, the first 2 displays were always predictive of the identity of the next, allowing temporal expectations to be fulfilled. This minimized the risk of unlearning the supposedly learned temporal sequence. Importantly, the fourth and last displays could be manipulated in the spatial (familiar, mixed, and shuffled) and temporal dimensions (expected or unexpected in time), resulting in a 3 × 2 design (see Fig. 1B). In order to avoid repetition suppression effects in the mixed display trials, we pseudorandomized the selection of the 2 novel objects presented within the mixed display target to ensure that these were never repeated within the same trial sequence. The participant's task was to indicate whether the fourth display was familiar (as presented in the training task) or novel (mixed or shuffled displays). The number of familiar and novel targets was balanced with respect to button mappings (i.e. half of the targets were familiar and the other half were novel). The trials of the familiar responses were further divided into expected familiar (following the temporal sequence) and unexpected familiar displays, with 37.5% and 12.5% conditional probabilities, respectively. The higher probability for expected familiar display was used to support the implicit learning of the temporal sequence. The trials of novel responses were divided equally into 4 conditions with a probability of 12.5% each, with spatial (mixed and shuffled) and temporal (expected and unexpected) manipulation. Participants had 2 s after the onset of the last display to report if it was familiar or novel. Feedback on behavioral accuracy (percentage correct) was provided at the end of each run. In the next trial, the sequence was restarted with a random familiar starting-display. There were 240 trials for the expected familiar condition and 80 trials for each of the other conditions, resulting in 640 trials in total. All types of trials were presented in a pseudorandomized order and split into 8 runs that lasted approximately 10 min each.

Object oddball task

On the second day, participants performed an object oddball task in the MR scanner. During each trial, 4 displays were presented sequentially for 500 ms each, with an intertrial interval of 4,000–6,000 ms (see Fig. 1D). Similar to the familiarity judgment task, the first 3 displays were always familiar and presented sequentially to give rise to temporal expectations, while the last display could be any type of spatial display (familiar, mixed, or shuffled) or temporal expectation (expected or unexpected). The participant's task was to press a button as soon as they detected an inverted object in the last display (12.5% of trials) while maintaining their eyes on the fixation point. Due to the introduction of the oddball trials, the conditional probability of familiar expected displays was raised to 43.75%, while the probabilities of other conditions (familiar unexpected, mixed, and shuffled displays) were reduced to 8.75%. For each trial, the temporal sequence of displays was restarted randomly. No feedback was provided during the task. There were 320 trials split into 4 runs, lasting approximately 50 min in total. The trial order was pseudorandomized. At the beginning, participants familiarized themselves with the task via a brief practice lasting ∼5 min.

Functional localizer

After finishing the object oddball task, participants underwent a functional localizer. We aimed to identify those voxels overlapping with early visual cortex (V1), object-selective LOC, and scene-selective parahippocampal place area (PPA). We presented both single objects and displays using the same object images from the previous task. The 2 types of localizers were used to inspect neural activity at the individual object and display level, respectively. The task consisted of 3 runs in a block design. In the first 2 runs, each run included 32 stimulus blocks and 8 null-event blocks. During the stimuli block, each of 4 objects from a familiar display was present alone and flashed at 2 Hz (250 ms on, 250 ms off) for 11 s, while during the null-event block, only the fixation dot was presented for 11 s. In each run, each of 4 objects from 1 familiar display was present once. In total, there were 80 blocks within 2 runs per participant.

The third run was identical in structure, but instead of single objects, we showed the 3 types of displays to select voxels responding to displays: familiar, shuffled, and phase-scrambled. The familiar stimuli were the 8 familiar displays used in the previous tasks, while the shuffled stimuli were 8 randomly chosen shuffled displays. To remove stimulus unspecific activation for the LOC localizer, we additionally included displays with phase-scrambled objects. There were 16 blocks for each type of display and 8 null-event blocks, resulting in 56 blocks in total. The order of blocks was randomized. For all runs, participants were instructed to fixate their eyes on the fixation point and respond by pressing the button whenever the fixation point dimmed in brightness.

fMRI data acquisition

Functional and anatomical images were collected on a 3T Skyra MRI system (Siemens) using a 32-channel head coil. Functional images were acquired using a whole-brain T2*-weighted multiband-6 sequence (TR/TE = 1,000/34 ms, 66 slices, voxel size = 2 mm isotropic, 60° flip angle, A/P phase encoding direction). Anatomical images were acquired with a T1-weighted MP-RAGE (GRAPPA acceleration factor = 2, TR/TE = 2,300/3.03 ms, voxel size = 1 mm isotropic, 8° flip angle).

fMRI data preprocessing

fMRI data preprocessing was performed using FSL 5.0.9 (FMRIB Software Library; Oxford, United Kingdom; www.fmrib.ox.ac.uk/fsl; RRID:SCR_002823). The preprocessing pipeline included brain extraction, motion correction, temporal high-pass filtering (128 s), and spatial smoothing (Gaussian kernel with 5 mm FWHM). Functional images were registered to the anatomical image using boundary-based registration as implemented in FLIRT and were subsequently normalized to the MNI152 T1 2 mm template brain using linear registration with 12° of freedom). For every run, the first 8 volumes were discarded to allow for signal stabilization.

Region of interest (ROI) definition

To examine the expectation effects throughout the visual hierarchy, we defined 3 ROIs: V1, object-selective LOC, and scene-selective PPA for each participant. We used the preregistered V1 and LOC to investigate activity modulations by expectation at the low-level (feature) and object-selective regions, whereas the nonpreregistered localizer for PPA allowed us to further examine potential activity modulations by expectation in a scene-selective region. For V1, Freesurfer 6.0 (General Hospital Corporation, Boston, MA, United States, RRID:SCR_001847) was used to extract labels (left and right) per participant based on their anatomical image, which were transformed to native space using mri_label2vol and were combined into a bilateral mask. To select voxels that maximally responded to the displays, we modeled the third run of the functional localizer using a General Linear Model (GLM) performed in FSL FEAT. We modeled the familiar, shuffled, phase-scrambled stimuli, and null-events with corresponding duration (11 s). First-order temporal derivatives and motion regressors were added as nuisance regressors. To define the object-selective LOC and scene-selective PPA ROIs, we used the bilateral masks from Julian et al. (2012). The obtained bilateral masks were transformed to native space and were then used as spatial constraints to select the 200 more responsive voxels (with highest z-statistics) based on the contrast of interest (familiar + shuffled + phase-scrambled—null-trial × 3). We contrasted "familiar + shuffled – phase-scrambled × 2" to identify the LOC voxels more responsive to intact compared with phase-scrambles objects and contrasted "familiar – phase-scrambled" to identify the PPA voxels more responsive to "scenes" (here defined as familiar displays). As preregistered, we also generated the LOC anatomical mask from the Harvard-Oxford cortical atlas, but to keep it consistent with the PPA mask, we report our results using Julian et al.'s (2012) LOC mask. We replicated similar results using the Julian et al.'s and the Harvard-Oxford cortical atlas LOC masks (see Fig. S1).

In accordance with our preregistered exploratory analysis, we also obtained the ROIs of individual objects in V1 and LOC to explore how expectations between the expected and unexpected objects within mixed displays modulated the BOLD amplitude. To do so, we carried out an additional GLM analysis with the first 2 runs of the functional localizer, where we modeled the 4 individual object positions (upper left, upper right, lower left, and lower right) and null-event with an 11 s duration as regressors. First-order temporal derivatives and motion regressors were added as nuisance regressors.

Since any 2 objects within a mixed display can be expected items, there are a total of 6 possible object pairs (upper, lower, left, right, and both diagonals). To generate the V1 masks for each of these 6 possible object pairs, we used GLMs contrasting any 2 objects minus the other 2 and selecting the most active 100 voxels. For object-selective LOC, we used the conjunction contrast of object pairs (object localizer) and intact minus scrambled objects (display localizer). We first selected the voxels more activated by intact objects compared with phase-scrambles, using the same method mentioned before. Then, these voxels were further constrained to 100 voxels, resulting in 6 ROIs for object pairs in LOC. As a robustness check for the ROI size, we additionally repeated all ROI analyses with ROI sizes ranging from 50 to 300 voxels in steps of 50 to exclude the possibility of arbitrarily chosen mask size.

Behavioral data analysis

For the familiarity judgment task, we compared the percentage of images rated as familiar (familiarity rate [FR]) and mean reaction time (RT) for all participants between conditions. We defined the FR as the number of familiar responses divided by the number of trials per condition. The data were averaged across trials per participant and were submitted to a 3×2 repeated measures ANOVA (RM ANOVA) with spatial display (familiar, mixed, and shuffled) and temporal expectation (expected or unexpected) as factors for FR and RT, respectively. We then used paired *t*-tests for the planned main effects analyses of the difference between expected and unexpected within each spatial display. For RT analysis, the trials whose RTs exceeded 3 MAD of the median or were <200 ms were discarded. The partial eta-squared (η^2) and Cohen's d_z were calculated as effect size for the RM ANOVA and paired *t*-test, respectively. All standard errors of the mean (SEM) shown in the present paper were calculated as the within-subject normalized SEM (Cousineau 2005).

fMRI data analysis

We modeled BOLD signal responses to the different experimental conditions by voxel-wise fitting GLMs to each run's data and participant using FSL FEAT. For the object oddball task, 6 experimental conditions (familiar expected, familiar unexpected, mixed expected, mixed unexpected, shuffled expected, and shuffled unexpected) and the target trials (in which an upside-down object occurred) were modeled with 2 s duration as regressors. In object-level ROI analyses, we further modeled 6 possible paired positions of expected objects (upper, lower, left, right, and diagonal) within mixed displays, resulting in 12 regressors for both temporally expected and unexpected conditions. In addition, nuisance regressors were added for the first-order temporal derivatives for all modeled events and for the 6 motion parameters (FSL's standard set of motion parameters). Data were combined across runs using FSL's fixed-effects analysis.

We carried out the group-level whole-brain analyses to characterize the spatial and temporal expectation suppression pattern across the brain. FSL's mixed effects analysis (FLAME 1) was used to combine data across participants. We also performed multiple-comparison correction through nonparametric tests (5,000 permutations) using the randomize function of the FSL. The statistical significance was assessed using the obtained corrected *P*-values. In the object oddball task, we used the contrasts "expected and unexpected shuffled - expected and unexpected familiar" and "expected and unexpected mixed - expected and unexpected familiar" to examine whether expectation suppression was sensitive to the spatial arrangement of the objects. We also contrasted "familiar expected–familiar unexpected" to test whether expectation suppression also takes place for temporal sequences. Finally, we explored whether spatial expectation suppression takes place automatically or depends on participants paying attention to the objects. To do so, we contrasted the shuffled displays versus the familiar displays using the independent data from the third localizer run in which participants performed a task in which the objects were irrelevant (i.e. the fixation brightness detection task).

ROI analysis

All reported ROI analyses were performed in each participant's native space by averaging all parameter estimates within a ROI and then comparing conditions within participants. To examine the main question whether temporal expectation depends on display, we extracted the parameter estimates for each condition separately from the whole-brain maps within each ROI (V1, LOC, and PPA; see Region of interest definition). These mean parameter estimates were then analyzed employing a 3×2 RM ANOVA with the factors display (familiar, mixed, and shuffled) and temporal expectation (expected and unexpected) for each ROI. Simple effects were calculated for temporal expectation in each display using paired *t*-tests. Additionally, a Bayesian *t*-test with a Cauchy prior width of 0.707 was used to assess any nonsignificant results.

To investigate whether expectations are formed at the object level, we further conducted ROI analyses using the ROIs of single objects (see Region of interest definition) in V1 and LOC, respectively. The 4 objects within a mixed display were then averaged to 2 object pairs in temporally expected and unexpected objects. The averaged parameter estimates within each ROI were then in turn subjected to a paired *t*-test to examine the expectation suppression for object pairs.

Software

PsychToolbox (Brainard 1997; Pelli 1997; Kleiner et al. 2007) running on MATLAB R2017b (The MathWorks, RRID:SCR_001622) was used for stimuli presentation. MRI data preprocessing and analysis were performed using FSL 5.0.9 (FMRIB Software Library; Oxford, United Kingdom; www.fmrib.ox.ac.uk/fsl; RRID:SCR_002823) and Freesurfer 6.0 (General Hospital Corporation, RRID:SCR_001847). We used a pipeline (Madan 2015) based on ITK-SNAP v 3.8.0 (Yushkevich et al. 2006) and ParaView v. 4.3.1 (Ayachit 2015) to visualize the MRI data in 3D. Python 3.7.4 (Python Software Foundation, RRID:SCR_008394) was used for data processing with following libraries. NumPy 1.17.2 (van der Walt et al. 2011) and Pandas 0.25.1 (McKinney 2010) were used for data handling; Matplotlib 3.1.1 (Hunter 2007) and Nanslice (Wood 2017, 2020) were used for data visualization. Pingouin 0.2.9 (Vallat, 2018) was used for statistical tests, including RM ANOVA, paired *t*-test, and Bayesian analyses.

Results

Temporal expectations facilitate behavior only for spatially structured arrangements

To examine whether the participants explicitly learned the displays and used the implicitly learned temporal sequences to improve their performance in the task, we analyzed FR and RTs during the object familiarity task. For both FR and RT, there were significant main effects of display (FR: $F_{(2, 66)} = 124.86$, $P < 0.001$, $\eta^2 = 0.79$; RT: $F_{(2, 66)} = 19.96$, $P < 0.001$, $\eta^2 = 0.38$), suggesting that participants performed differently as a function of the presented display. Indeed, Fig. 2A shows that participants correctly identified most of the familiar displays as familiar (familiar condition: ~86% of trials rated as familiar), while they predominantly categorized the shuffled displays as novel (shuffled condition: ~17% of trials rated as familiar). However, the familiarity ratings were

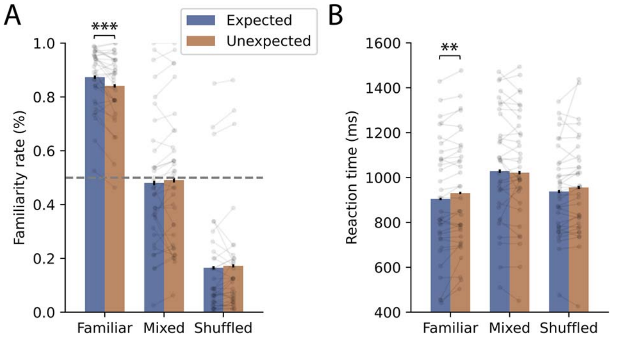

Fig. 2. Behavioral performance from the familiarity judgment task. A) FRs are higher for familiar than mixed or shuffled displays, indicating that participants could distinguish these 3 types of displays. The FRs were significantly higher for the temporal expected compared with unexpected displays in familiar display, whereas there were no differences in the mixed or shuffled displays; B) likewise, the RTs to expected items were faster than to unexpected items only in the familiar display. Gray dots represent individual participants. Error bars indicate within-subject SE. ** $P < 0.01$, *** $P < 0.001$.

more variable when object items from different displays were mixed (~49% of trials rated as familiar) compared to displays with the objects spatially shuffled. This indicates that participants' familiarity reports are highly determined by the correct display, but not the correct identity of the elements in each display. Briefly, participants could overall distinguish between familiar and novel displays, where unfamiliar object locations (shuffled condition) were easier to identify than unfamiliar mixing of objects between 2 displays.

Furthermore, the significant interactions between display and temporal sequences (FR: $F_{(2, 66)} = 5.66$, $P = 0.01$, $\eta^2 = 0.15$; RT: $F_{(2,66)} = 4.05$, $P = 0.02$, $\eta^2 = 0.11$) indicate that the effects of temporal expectation were modulated by the display. A post hoc *t*-test showed that the familiar expected items were more often correctly rated as familiar compared to the unexpected ones ($t_{(33)} = 3.70$, $P < 0.001$, $d_z = 0.65$). However, no FR differences were observed in the mixed and shuffled displays (mixed: $t_{(33)} = -0.86$, $P = 0.39$, $d_z = -0.15$; shuffled: $t_{(33)} = -0.82$, $P = 0.42$, $d_z = -0.14$). Bayesian analyses provided moderate evidence in favor of no differences between expected and unexpected items in the mixed and shuffled displays (mixed: $BF_{10} = 0.26$; shuffled: $BF_{10} = 0.25$). Similarly, participants categorized temporally expected familiar displays faster than temporally unexpected displays (26 ms [905 vs. 931 ms]; $t_{(33)} = -3.48$, $P < 0.01$, $d_z = 0.61$), whereas no RT benefit was observed for mixed and shuffled displays (mixed: $t_{(33)} = 0.65$, $P = 0.52$, $d_z = 0.11$; shuffled: $t_{(33)} = -1.84$, $P = 0.08$, $d_z = -0.32$). Bayesian analysis indicated moderate support for the absence of a difference in mixed display (1,028 vs. 1,021 ms, $BF_{10} = 0.23$) and inconclusive evidence in the shuffled display (938 vs. 956 ms, $BF_{10} = 0.83$). In sum, these results showed that participants used co-occurrence statistics of multiple objects to facilitate the recognition of upcoming displays.

Temporal expectation suppression was evident only in LOC

Next, we investigated whether temporal expectations are formed at the individual object level or relied on the spatial organization of the objects within each display. In order to examine how expectation suppression changes as a function of both display and temporal expectations, we ran 3 × 2 RM ANOVAs with the factors display (familiar, mixed, and shuffled) and temporal expectation (expected and unexpected) on 3 a priori-defined ROIs (Fig. 3A) in the ventral visual stream.

In all 3 ROIs, we found that BOLD signal to familiar displays was suppressed compared to BOLD signal to novel displays (Fig. 3B). This effect was revealed by the main effects of display (V1: $F_{(2, 66)} = 12.43$, $P < 0.001$, $\eta^2 = 0.27$; LOC: $F_{(2, 66)} = 35.56$, $P < 0.001$, $\eta^2 = 0.52$; PPA: $F_{(2, 66)} = 19.61$, $P < 0.001$, $\eta^2 = 0.37$). Paired *t*-tests confirmed the suppressed neural activity to the familiar compared with novel displays (all Ps < 0.001). This pattern of results could be mediated by the formation of the higher-order representation of the spatial arrangement of object items.

By contrast, the ANOVA neither showed any significant modulations by temporal expectation (V1: $F_{(1, 33)} = 0.81$, $P = 0.37$, $\eta^2 = 0.02$; LOC: $F_{(1, 33)} = 1.09$, $P = 0.30$, $\eta^2 = 0.03$; PPA: $F_{(1, 33)} = 0.18$, $P = 0.67$, $\eta^2 = 0.01$) nor an interaction between display and temporal expectations in all ROIs (V1: $F_{(2, 66)} = 0.15$, $P = 0.86$, $\eta^2 < 0.01$; LOC: $F_{(2, 66)} = 2.91$, $P = 0.06$, $\eta^2 = 0.08$; PPA: $F_{(2, 66)} = 0.31$, $P = 0.73$, $\eta^2 = 0.01$). Due to the interaction being close to our alpha threshold in LOC, we performed paired *t*-tests to examine the modulation by temporal expectation focusing on LOC ROI. We found that temporal expectation suppression was present when we compared the temporal expected to unexpected familiar displays ($t_{(33)} = -3.44$, $P < 0.01$, $d_z = -0.60$) but was absent when the displays were spatially mixed or shuffled (mixed: $t_{(33)} = -0.25$, $P = 0.81$, $d_z = -0.04$; shuffled: $t_{(33)} = 0.79$, $P = 0.44$, $d_z = 0.14$). Bayesian analyses were carried out, showing moderate support for no expectation suppression in these novel displays (mixed: $BF_{10} = 0.19$; shuffled: $BF_{10} = 0.25$).

In a nonpreregistered exploratory analysis, a 2-way RM ANOVA was performed to test whether the temporal expectation effects (unexpected minus expected) changed as a function of the factors ROI (V1, LOC, and PPA) and display (familiar, mixed, and shuffled) across the 3 brain regions (Fig. 3C). A significant interaction between the 3 brain regions and displays ($F_{(4, 132)} = 2.97$, $P = 0.02$, $\eta^2 = 0.08$) indicated that expectation suppression was modulated over brain regions. Further *t*-test contrasts showed that temporal expectation effects were significantly larger in LOC than in V1 ($t_{(33)} = 2.62$, $P = 0.01$, $d_z = 0.46$) and PPA ($t_{(33)} = 3.67$, $P < 0.001$, $d_z = 0.64$) for the familiar displays, while there was no difference between the ROIs for the mixed and shuffled displays (all Ps > 0.10). These results are congruent with our previous analysis in showing that temporal expectation suppression only occurs for familiar displays and only in LOC.

In addition to ROI analyses, we also performed a preregistered whole-brain analysis to investigate whether temporal expectations modulated neural activity outside the a priori-defined ROIs. Figure 4A shows that, for familiar displays, the temporally expected versus unexpected displays resulted in lower brain activity only in the left LOC, while no significant cluster was observed in mixed or shuffled displays (not shown). Taken together, these results showed that temporal expectation suppression was only evident in LOC for the familiar display.

Temporal expectations rely on spatially structured arrangements

Our previous analyses indicate that temporal expectation effects are only present for familiar displays, suggesting that participants predict upcoming displays rather than individual objects. However, analyzing the BOLD signal modulations evoked by 4 objects together might obscure the individual contributions of the locally expected and unexpected objects in the mixed displays. For example, it could be that, whereas the 2 locally expected objects generated strong expectation suppression, the 2 unexpected items produced the same effect but in the opposite direction (i.e. surprise enhancement), leading on average to a null effect. Hence, to test whether locally predictable individual objects produce local

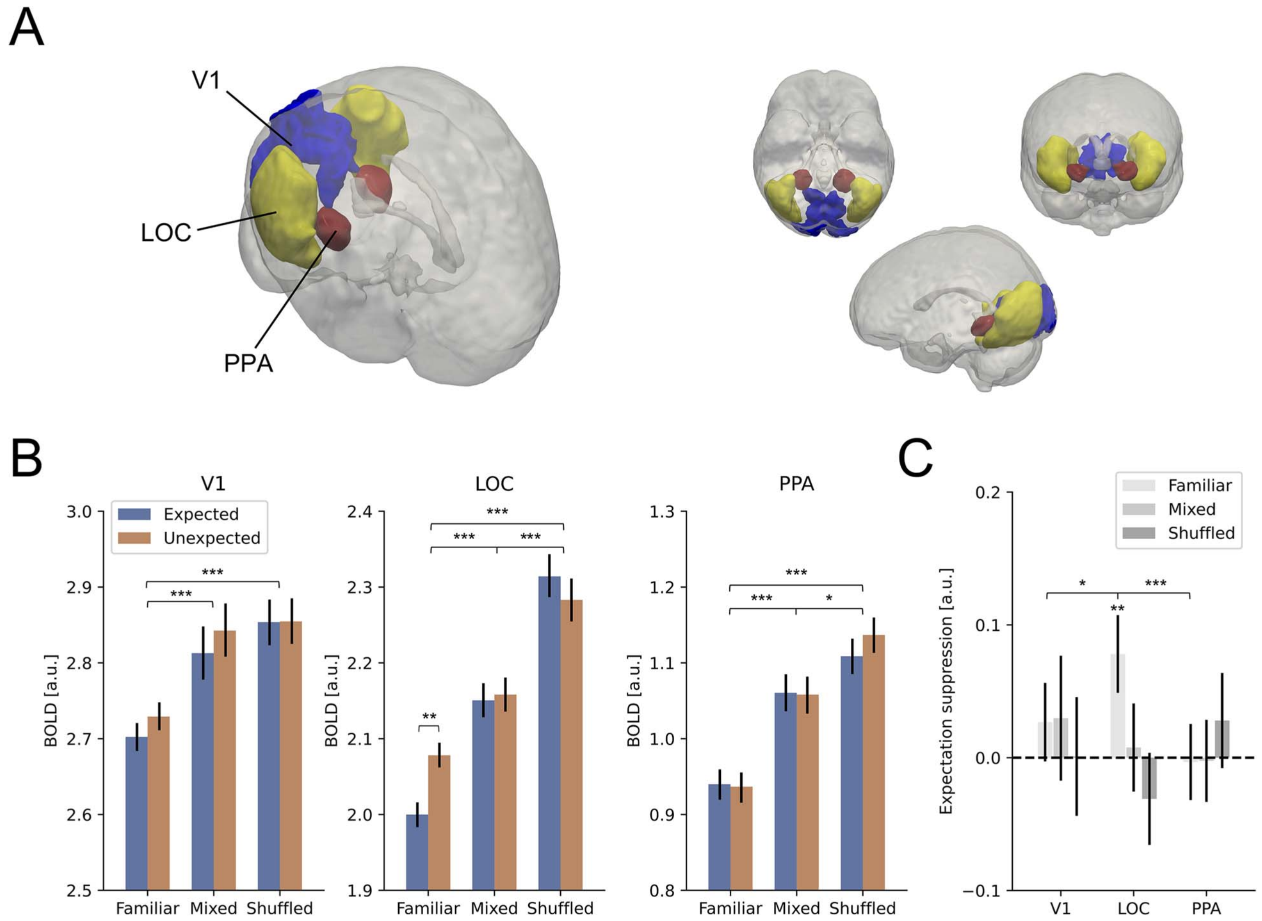

Fig. 3. A) Three anatomically defined masks in the ventral visual pathway overlaid onto a 3D glass brain from a representative participant: early visual cortex, object-selective LOC, and scene-selective PPA. B) Averaged parameter estimates within V1 (left), LOC (middle), and PPA (right). In all 3 ROIs, BOLD responses were significantly suppressed to the familiar display. In LOC, the BOLD signals showed significant suppression to the temporal expected items compared to unexpected in familiar displays. No difference was found between BOLD responses to temporally expected and unexpected items in mixed or shuffled displays in LOC, or in all 3 types of displays in V1 and PPA. C) Temporal expectation suppression across all brain regions. Error bars indicate within-subject SE. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

expectation suppression effects, we generated spatially specific ROIs to separate the BOLD responses to each pair of objects in the mixed displays depending on whether they were temporally predictable or unpredictable (Fig. 5A). To distinguish these ROIs from the ROIs used in previous analyses, the term "paired ROIs" was used here to refer to the ROIs for object pairs (see Region of interest definition).

We compared the BOLD responses to temporally expected and unexpected pairs within the expected mixed displays (Fig. 5B). Paired *t*-tests revealed no significant difference between the temporally expected and unexpected objects within the expected mixed display (V1: $t_{(33)} = 0.92$, $P = 0.36$, $d_z = 0.16$; LOC: $t_{(33)} = 1.55$, $P = 0.13$, $d_z = 0.27$), indicating no detectable object-specific expectation effect to the partial familiar displays. Likewise, Bayesian analyses indicated weak-to-mixed evidence against expectation suppression in the expected mixed displays (V1: $BF_{10} = 0.27$; LOC: $BF_{10} = 0.54$). Taken together, these results suggest that there was no local expectation suppression for temporally expected compared to unexpected pairs (for validation see Fig. S2–S4 and Table S1).

Expectation suppression for familiar displays throughout the ventral visual stream

We performed an exploratory whole-brain analysis to investigate the neural topography of the spatial expectation suppression for familiar displays outside the a priori-defined ROIs within the ventral visual stream. We contrasted BOLD responses to the familiar displays versus novel displays (mixed and shuffled displays, respectively). As illustrated in Fig. 4B, the whole-brain analysis confirmed the ROI analysis, revealing extensive clusters of suppressed neural activity throughout the visual ventral stream, including the early visual cortex, bilateral LOC, bilateral fusiform gyrus, and bilateral PPA. Outside the ventral visual stream, additional clusters of expectation suppression were observed in middle and inferior frontal gyri, caudate, insula gyrus, and cingulate gyrus. Furthermore, in addition to the suppression activity, we also observed enhanced activity in ventromedial prefrontal cortex (vmPFC), inferior parietal cortex, posterior cingulate cortex, and precuneus.

Next, we assessed whether the previously reported suppression effects for the familiar displays took place automatically when the objects were unattended. For this analysis, we used the independent data from the third localizer run. Note that the displays presented in this localizer run were fully task-irrelevant, as the participants performed a demanding task at the fixation point. Thus, it allowed us to investigate whether expectation suppression takes place when the participants had already learned the spatial associations between the objects, but the displays were task-irrelevant. We contrasted the familiar displays versus shuffled displays, and we found similar clusters of spatial

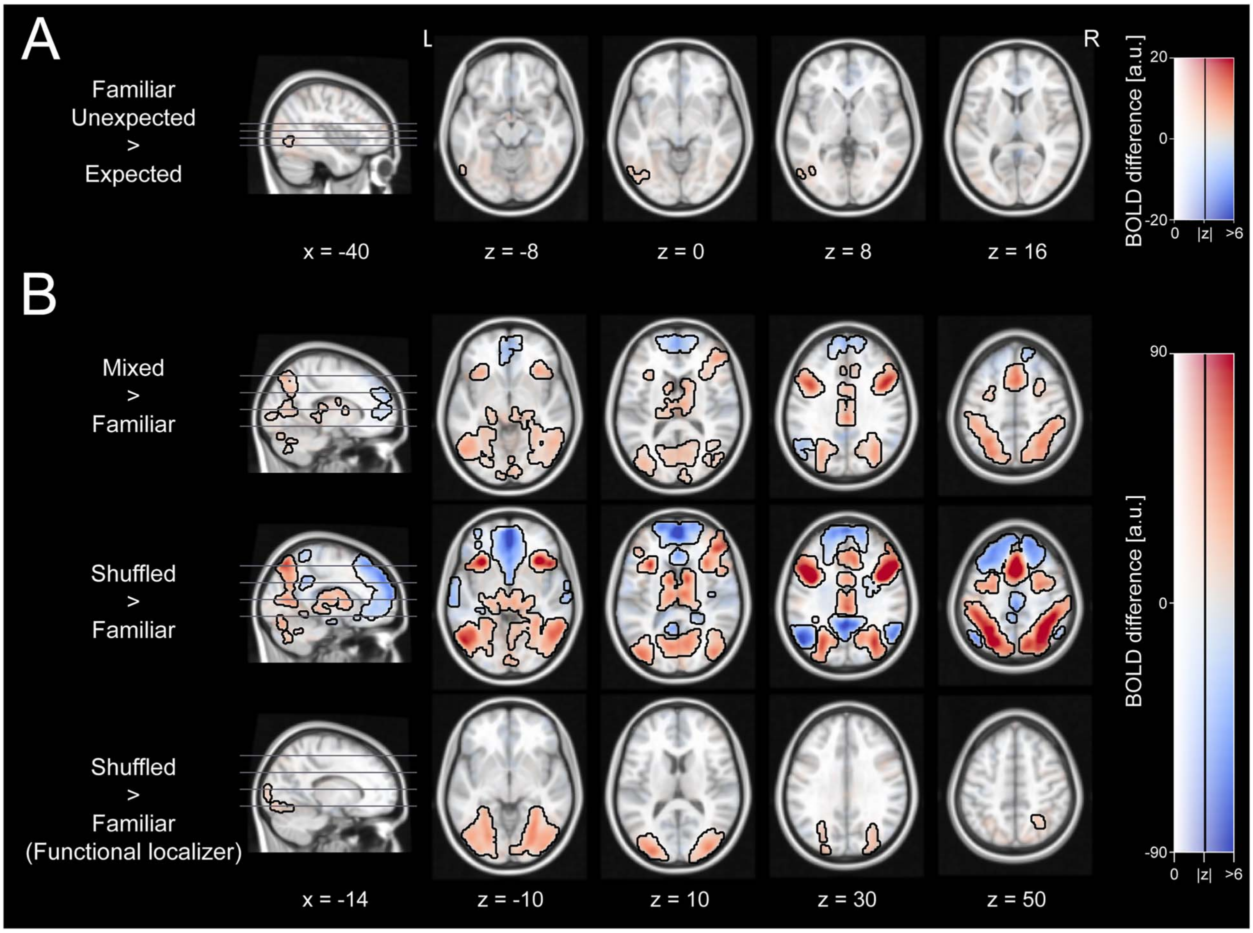

Fig. 4. Expectation suppression revealed by whole-brain analyses. Color represents the parameter estimates for unexpected minus expected displays on the MIN 152 2 mm template brain: Red clusters represent increased activity (compared to the expected display), while blue clusters represent decreased activity (compared to the expected display). Opacity indicates the z-statistics of the contrasts. Black contours outline statistically significant clusters, associated z-statistics were shown as black lines in the color bars. A) Temporal expectation suppression for the expected compared to unexpected familiar displays. The significant clusters were in the left LOC; B) spatial expectation suppression for the familiar compared to novel displays. When participants attended to the objects in the object oddball task (upper and middle rows), the significant clusters were present in the visual ventral stream (early visual cortex, bilateral LOC, bilateral fusiform gyrus, and bilateral PPA), middle and inferior frontal gyri, anterior insula, and inferior cingulate gyrus. When participants attended to the fixation point in the localizer run (lower row), significant clusters were less extensive compared to those of the oddball task only in the visual ventral stream.

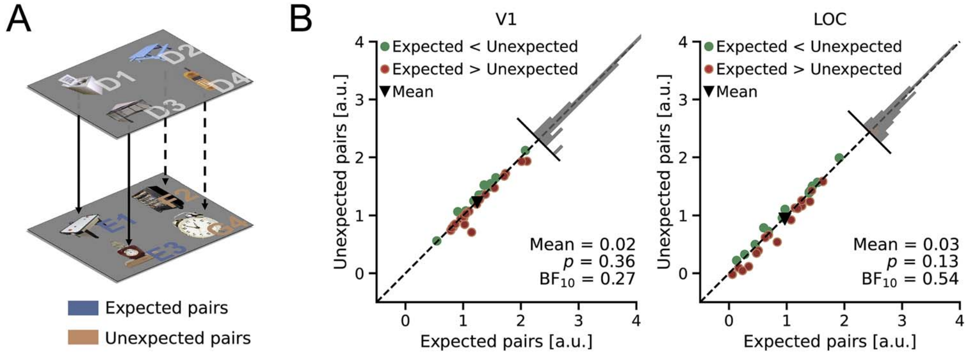

Fig. 5. A) Illustration of the object pairs within expected mixed displays. After presentation of a familiar display labeled with letter D, participants would expect to see the familiar display E next. In the mixed display conditions, we randomly replaced 2 objects in display E, resulting in an expected mixed display. Thus, only 2 objects were temporally expected (solid arrows), while the other 2 were unexpected (dashed arrows). B) The BOLD responses to temporally expected (x-axes) and unexpected (y-axes) object pairs within the expected mixed displays. The dashed line indicates no difference between the BOLD response to expected and unexpected object pairs. The histograms represent the distribution of deviations from the unity line.

expectation suppression in the early visual cortex, bilateral LOC, bilateral fusiform gyrus, and bilateral PPA (Fig. 4B, lower row). However, no significant clusters were observed outside the visual ventral stream. In summary, our results show that spatial expectation suppression to familiar displays was evident in the ventral visual stream regardless of whether the stimuli were attended or unattended, suggesting that spatial expectations along the ventral visual stream unfold automatically.

Discussion

In the current study, we investigated whether expectations about upcoming input incorporate global information about the spatial arrangement of co-occurring objects, or whether expectations are formed locally for each item present in the visual field. Our results showed that, after participants learned statistical regularities to associate objects into spatially structured arrangement as a higher-level representation, that were presented in fixed temporal sequences, participants could utilize the learned co-occurrence statistics of multiple objects to guide their expectations about upcoming arrangements of objects. This was confirmed by the behavioral benefits in RTs and discrimination of temporally expected familiar displays. Moreover, fMRI analyses revealed the presence of expectation suppression effects to the temporally expected displays in LOC but only when the displays were familiar. On the contrary, no expectation suppression was observed for temporally expected compared to temporally unexpected individual objects when these objects were no longer part of a familiar display. These findings suggest that the brain forms expectations at a global level. Furthermore, we shuffled the relative positions of objects within a learned display and showed that temporal expectation represents the associated objects in a spatially structured unit. In sum, our findings provide evidence that the construction of temporal expectations is predominantly driven by a higher-level representation of co-occurring objects that operates holistically rather than focusing on individual parts.

Temporal expectations benefit from co-occurrence statistics of stimuli

Our behavioral data showed that participants were able to differentiate the learned displays from the novel displays (mixed and shuffled) after training. Surprisingly, mixed displays were more likely perceived as familiar compared to shuffled displays during the familiarity judgment task. This suggests that participants placed more weight on the position of each individual item within each display as a cue to decide whether a display is familiar or novel than on the identity of items composing it. An alternative explanation is that participants may use heuristics during the familiarity judgment task, for example, they might only examine 1 of the 4 objects to determine if a display is familiar or novel. In this case, there is a 1 in 2 chance of selecting a familiar item in mixed displays, resulting in a FR at chance level. However, if participants only evaluated 1 object across all displays, we would expect to see no difference in RT between them. By contrast, our results showed slower RTs for mixed displays compared to familiar and shuffled displays, indicating that participants needed more time to process the mixed displays. Thus, we believe the lower FR for mixed displays is likely due to the difficulty in recognizing it.

In general, our results clearly show that participants learned the temporal regularities and could use it to guide their expectation of upcoming stimuli. Crucially, the benefit of temporal expectation was only observed when the objects were presented as a structured display. Thus, our results suggest that participants grouped simultaneously presented objects into displays, which were used to facilitate the processing of the next display.

Co-occurring objects are represented as a unified display in space

By taking a quick look at objects in the real world, we can notice that objects are rarely isolated from each other and therefore global processing may be especially ecological and efficient. In this view, the human brain should hold a global representation of the objects' spatial structure. In fact, our results showed that the neural responses to the familiar display in the ventral visual stream (V1, LOC, and PPA) was reduced compared to the novel (mixed and shuffled) displays. Notably, these results cannot be explained by differences in frequency or familiarity, as we controlled the presentation frequency of individual objects. We speculate that the differences in BOLD response are due to the violation of 2 types of regularities for representing co-occurring objects: co-occurrence statistics, in which certain objects are more likely to appear together (e.g. keyboard and mouse), and positional regularities, in which objects tend to appear at typical locations (e.g. keyboard on the left of a mouse). Therefore, the enhanced BOLD response to the mixed display compared to the familiar display indicated the violation of the object identities, while the enhanced BOLD response to the shuffled display compared to the familiar display indicated the violation of the objects' positions. Interestingly, we observed larger neural responses to shuffled than mixed displays in LOC and PPA (Fig. 3B), suggesting that the shuffled displays were more surprising than the mixed displays. This is consistent with our behavioral results, where the shuffled displays were easier to categorize as novel than the mixed display. These suppression effects in LOC and PPA may reflect the expectation of the positional properties rather than the identity properties of co-occurring objects (Epstein 2008; Hayworth et al. 2011).

Moreover, the whole-brain analyses indicated that the suppressed neural activity for familiar displays involved similar brain regions as in previously reported expectation suppression for temporally expected stimuli (Richter et al. 2018; Ferrari et al. 2022). The suppression effects overlapped not only in the ventral visual stream but also in nonsensory areas, such as inferior frontal gyrus and anterior insula, revealing that expectations of spatial and temporal context might rely on similar neural mechanisms. In addition, we observed stronger BOLD responses to familiar compared to shuffled displays in several nonsensory higher-order regions, such as the vmPFC, which play a crucial role in representing superordinate knowledge structures (Gilboa and Marlatte 2017). Our findings are also consistent with earlier research (van Kesteren et al. 2010, 2013; Bein et al. 2014),which has demonstrated that prior knowledge leads to enhanced activity in medial prefrontal cortex (mPFC), as well as enhanced connectivity between the mPFC and visual areas, suggesting that the mPFC may serve as a potential source of perceptual predictions. Finally, we found that the suppression effects to the familiar displays were evident in the ventral visual stream even with attention diverted from the stimuli by using independent data from the localizer run. This suggests that the suppression for familiar displays is preattentive and therefore is an automatic process. Taken together, our results provide evidence that participants were able to exploit the spatial regularities to represent multiple objects as a structured unit in space.

Co-occurring objects are predicted in time as a structural unit

Our visual system needs to constantly predict new objects dealing with a limited capacity of a finite number of objects at once.

One strategy to overcome this visual system limitation is to predict sets of objects at a higher level, representing objects as a structured unit rather than as individuals (Kaiser et al. 2019). Our results showed temporal expectation suppression when sets of objects were represented as wholes, but this was absent when the objects were represented individually, suggesting that temporal expectations convey global information for sets of objects instead of separately predicting individual objects. Nevertheless, our results do not exclude the possibility that the visual system might predict multiple objects simultaneously and independently if the object sequences are learned separately from one another (Rosenthal et al. 2018).

A surprising result is that we only observed expectation suppression for temporally expected displays in LOC but not in V1 or PPA. This may be because LOC is more sensitive to the stimuli we used in the present study. Besides visual object recognition (Allen et al. 2012), it has been shown that LOC is also sensitive to the positions of objects (Gronau et al. 2008; Hayworth et al. 2011) and important in coding the scene-like relations between multiple objects (Kim and Biederman 2011; Kaiser and Peelen 2018). Another possible explanation for the null expectation effect in V1 and PPA might be that in addition to temporal expectation, the neural responses to the familiar spatial displays were also modulated by their spatial structure. This is confirmed by robust activity suppression in learned (familiar) compared to novel displays in all ROIs. Therefore, the strong suppression effects for spatial displays may have induced a ceiling effect, masking the temporal expectation effects.

An alternative strategy to represent associated objects at a higher abstraction level is to represent collections of objects using ensemble statistics (Ariely 2001; Alvarez and Oliva 2009; Alvarez 2011). It has been proposed that the visual system can calculate the statistical summary of a set of objects to provide a compressed, accurate "gist" representation. For instance, Brady and Alvarez (2011) found that the remembered size of individual items in a display was biased toward the mean size of all items in the display. These findings imply that the visual system might generate expectations based on these ensemble representations. Thus, the expectation effects should be observed in the shuffled display, which contained the same ensemble statistics as the familiar display. However, our results showed that expectation suppression was absent when the relative position of objects was shuffled, suggesting participants predicted a structured but not ensemble representation. Only ensemble information is not sufficient for activating temporal expectations. This is compatible with the view that positional regularities play a key role in optimally representing complex visual scenes (Kaiser et al. 2019).

Conclusion

Most previous research on predictive mechanisms has focused on investigating temporal expectations of individual objects, neglecting the more realistic situation where multiple objects are present in a scene and can be predicted collectively. Here, we sought to explore this scenario and found that the human brain generates temporal expectations for co-occurring objects on a global level. Specifically, when co-occurring objects are grouped into a scene-like representation, we observed expectation suppression effects for these scenes, rather than individual objects, thus revealing that perceptual anticipation operates predominantly on the global level. These findings have important implications for our understanding of how humans predict complex scenes by showing that expectations act upon high-level representations of the visual input. Such bound multiobject representations might help to free up processing resources by reducing information redundancy when the scene is already known.

Authors' contributions

CY, BVE, AP-B and FPdL designed research; CY performed experiments; CY and BVE analyzed data; CY and BVE wrote the first draft of the paper; CY, BVE, AP-B, MVP, and FPdL edited and revised the paper.

CRediT authors statement

Chuyao Yan (Conceptualization, Data curation, Formal analysis, Visualization, Writing—original draft, Writing—review & editing), Benedikt V. Ehinger (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Alexis Pérez Bellido (Conceptualization, Writing—review & editing), Marius V. Peelen (Writing—review & editing), and Floris P. de Lange (Conceptualization, Funding acquisition, Writing—review & editing)

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

CY is supported by a grant from the China Scholarship Council (201708330238). BVE is supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy-EXC 2075-390740016. AP-B is supported by RTI2018-100977-J-I00 from MINECO (Spain). MVP is supported by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725970). FPdL is supported by a grant from the Horizon 2020 Framework Programme (ERC Starting Grant 678286). The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement: The authors declare no competing financial interests.

References

Alvarez GA. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn Sci*. 2011:15(3):122–131. https://doi.org/10.1016/j.tics.2011.01.003.

Alvarez GA, Oliva A. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proc Natl Acad Sci*. 2009:106(18):7345–7350. https://doi.org/10.1073/pnas.0808981106.

Ariely D. Seeing sets: representation by statistical properties. *Psychol Sci*. 2001:12(2):157–162. https://doi.org/10.1111/1467-9280.00327.

Ayachit U. *The ParaView guide: a parallel visualization application*. Kitware, Inc; 2015. ISBN 9781930934306.

Bein O, Reggev N, Maril A. Prior knowledge influences on hippocampus and medial prefrontal cortex interactions in subsequent memory. *Neuropsychologia*. 2014:64:320–330. https://doi.org/10.1016/j.neuropsychologia.2014.09.046.

Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: detecting and judging objects undergoing relational violations. *Cogn Psychol*. 1982:14(2):143–177. https://doi.org/10.1016/0010-0285(82)90007-X.

Boettcher SEP, Stokes MG, Nobre AC, van Ede F. One thing leads to another: anticipating visual object identity based on associative-memory templates. *J Neurosci*. 2020:40(20):4010–4020. https://doi.org/10.1523/JNEUROSCI.2751-19.2020.

Brady TF, Alvarez GA. Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychol Sci* 2011:22(3):384–392. https://doi.org/10.1177/0956797610397956.

Brady TF, Konkle T, Alvarez GA, Oliva A. Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci*. 2008:105(38):14325–14329. https://doi.org/10.1073/pnas.0803390105.

Brainard DH. The psychophysics toolbox. *Spat Vis*. 1997:10(4): 433–436.

Chun MM, Jiang Y. Top-down attentional guidance based on implicit learning of visual covariation. *Psychol Sci*. 1999:10(4):360–365. https://doi.org/10.1111/1467-9280.00168.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci*. 2013:36(3):181–204. https://doi.org/10.1017/S0140525X12000477.

Cousineau D. Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method. *Tutor Quant Methods Psychol*. 2005:1(1):42–45.

Egner T, Monti JM, Summerfield C. Expectation and surprise determine neural population responses in the ventral visual stream. *J Neurosci*. 2010:30(49):16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010.

Epstein RA. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends Cogn Sci*. 2008:12(10):388–396. https://doi.org/10.1016/j.tics.2008.07.004.

Ferrari A, Richter D, Lange FP de. Updating contextual sensory expectations for adaptive behavior. *J Neurosci*. 2022:42(47):8855–8869. https://doi.org/10.1523/JNEUROSCI.1107-22.2022.

Fiser J, Aslin RN. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol Sci*. 2001:12(6): 499–504. https://doi.org/10.1111/1467-9280.00392.

Fiser J, Aslin RN. Encoding multielement scenes: statistical learning of visual feature hierarchies. *J Exp Psychol Gen*. 2005:134:521–537. https://doi.org/10.1037/0096-3445.134.4.521.

Gilboa A, Marlatte H. Neurobiology of schemas and schema-mediated memory. *Trends Cogn Sci*. 2017:21(8):618–631. https://doi.org/10.1016/j.tics.2017.04.013.

Green C, Hummel JE. Familiar interacting object pairs are perceptually grouped. *J Exp Psychol Hum Percept Perform*. 2006:32(5): 1107–1119. https://doi.org/10.1037/0096-1523.32.5.1107.

Gronau N, Neta M, Bar M. Integrated contextual representation for objects' identities and their locations. *J Cogn Neurosci*. 2008:20(3): 371–388.

Hayworth KJ, Lescroart MD, Biederman I. Neural encoding of relative position. *J Exp Psychol Hum Percept Perform*. 2011:37(4):1032–1050. https://doi.org/10.1037/a0022338.

Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007:9(3):90–95. https://doi.org/10.1109/MCSE.2007.55.

Julian JB, Fedorenko E, Webster J, Kanwisher N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*. 2012:60(4):2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055.

Kaiser D, Peelen MV. Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage*. 2018:169:334–341. https://doi.org/10.1016/j.neuroimage.2017.12.065.

Kaiser D, Quek GL, Cichy RM, Peelen MV. Object vision in a structured world. *Trends Cogn Sci*. 2019:23(8):672–685. https://doi.org/10.1016/j.tics.2019.04.013.

Kaposvari P, Kumar S, Vogels R. Statistical learning signals in macaque inferior temporal cortex. *Cereb Cortex*. 2018:28(1): 250–266. https://doi.org/10.1093/cercor/bhw374.

Kim JG, Biederman I. Where do objects become scenes? *Cereb Cortex (New York, NY)*. 2011:21(8):1738–1746. https://doi.org/10.1093/cercor/bhq240.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. What's new in psychtoolbox-3. *Perception*. 2007:36(14):1–16.

Lengyel G, Žalalytė G, Pantelides A, Ingram JN, Fiser J, Lengyel M, Wolpert DM. Unimodal statistical learning produces multimodal object-like representations. *elife*. 2019:8: e43942.

McKinney W. *Data Structures for Statistical Computing in Python*. 2010:56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Madan CR. Creating 3D visualizations of MRI data: a brief guide. *F1000Research*. 2015:4:466.

Manahova ME, Mostert P, Kok P, Schoffelen JM, De Lange FP. Stimulus familiarity and expectation jointly modulate neural activity in the visual ventral stream. *J Cogn Neurosci*. 2018:30(9): 1366–1377.

Meyer T, Olson CR. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc Natl Acad Sci*. 2011:108(48): 19401–19406. https://doi.org/10.1073/pnas.1112895108.

Orbán G, Fiser J, Aslin RN, Lengyel M. Bayesian learning of visual chunks by human observers. *Proc Natl Acad Sci U S A*. 2008:105(7):2745–2750. https://agris.fao.org/agris-search/search.do?recordID=US201300862339.

Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis*. 1997:10(4):437–442. https://doi.org/10.1163/156856897X00366.

Richter D, Ekman M, de Lange FP. Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *J Neurosci*. 2018:38(34):7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018.

Rosenthal CR, Mallik I, Caballero-Gaudes C, Sereno MI, Soto D. Learning of goal-relevant and -irrelevant complex visual sequences in human V1. *NeuroImage*. 2018:179:215–224. https://doi.org/10.1016/j.neuroimage.2018.06.023.

Stansbury DE, Naselaris T, Gallant JL. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*. 2013:79(5):1025–1034. https://doi.org/10.1016/j.neuron.2013.06.034.

Summerfield C, Trittschuh EH, Monti JM, Mesulam M-M, Egner T. Neural repetition suppression reflects fulfilled perceptual expectations. *Nat Neurosci*. 2008:11(9):1004–1006. https://doi.org/10.1038/nn.2163.

Thaler L, Schütz AC, Goodale MA, Gegenfurtner KR. What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vis Res*. 2013:76:31–42. https://doi.org/10.1016/j.visres.2012.10.012.

Vallat R. Pingouin 0.2.9: statistics in python. *J Open Source Softw*. 2018: 3(31):1026.

van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng*. 2011:13(2):22–30. https://doi.org/10.1109/MCSE.2011.37.

van Kesteren MTR, Rijpkema M, Ruiter DJ, Fernández G. Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *J Neurosci*. 2010:30(47):15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010.

van Kesteren MTR, Beul SF, Takashima A, Henson RN, Ruiter DJ, Fernández G. Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia*. 2013:51(12):2352–2359. https://doi.org/10.1016/j.neuropsychologia.2013.05.027.

Wood T. *Spinicist/nanslice [python]*. 2020. https://github.com/spinicist/nanslice. (Original work published 2017).

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*. 2006:31(3):1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.