Spatial and Temporal Context Jointly Modulate the Sensory Response within the Ventral Visual Stream

Tao He[1], David Richter[2], Zhiguo Wang[3], and Floris P. de Lange[2]

Abstract

■ Both spatial and temporal context play an important role in visual perception and behavior. Humans can extract statistical regularities from both forms of context to help process the present and to construct expectations about the future. Numerous studies have found reduced neural responses to expected stimuli compared with unexpected stimuli, for both spatial and temporal regularities. However, it is largely unclear whether and how these forms of context interact. In the current fMRI study, 33 human volunteers were exposed to pairs of object stimuli that could be expected or surprising in terms of their spatial and temporal context. We found reliable independent contributions of both spatial and temporal context in modulating the neural response. Specifically, neural responses to stimuli in expected compared with unexpected contexts were suppressed throughout the ventral visual stream. These results suggest that both spatial and temporal context may aid sensory processing in a similar fashion, providing evidence on how different types of context jointly modulate perceptual processing. ■

INTRODUCTION

Humans are exquisitely sensitive to visual statistical regularities. Indeed, knowledge of both temporal and spatial context can facilitate visual perception and perceptual decision-making (Bar, 2004). Facilitatory effects of temporal context have been shown, for instance, during exposure to sequentially presented stimuli, with faster and more accurate responses to expected compared with unexpected stimuli (Richter & de Lange, 2019; Bertels, Franco, & Destrebecqz, 2012; Hunt & Aslin, 2001). At the same time, neural responses have been shown to be modulated by temporal context, with a marked suppression of sensory responses to expected compared with unexpected stimuli, reported in humans (Richter & de Lange, 2019; Richter, Ekman, & de Lange, 2018; Egner, Monti, & Summerfield, 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008) and nonhuman primates (Kaposvari, Kumar, & Vogels, 2018; Meyer & Olson, 2011). Studies have also shown that spatial expectations facilitate the recognition of objects (Quek & Peelen, 2020; Kaiser, Quek, Cichy, & Peelen, 2019; Munneke, Brentari, & Peelen, 2013; Davenport, 2007; Davenport & Potter, 2004). For instance, a foreground object is more easily identified when it appears on congruent backgrounds, compared with when it appears on incongruent backgrounds (Davenport & Potter, 2004). Thus, both spatial and temporal associations may result in expectations about stimulus identities, which can facilitate perception, such as object recognition.

On the one hand, spatial expectations can concern where a stimulus may occur, prompting allocation of spatial attention, such as during spatial cuing and similar paradigms (Castelhano & Witherspoon, 2016; Henderson, Malcolm, & Schandl, 2009; Torralba, Oliva, Castelhano, & Henderson, 2006). These studies show that positional regularities can guide attentional allocation during search. On the other hand, spatial expectations can predict arrangements of specific stimuli. That is, seeing your computer's keyboard predicts seeing your mouse. How such predictions about stimulus identity based on spatial arrangements modulate sensory processing has been investigated less, but nonetheless represents crucial information for an agent interacting with the sensory world.

Moreover, although temporal (sequence) predictions have been studied extensively (Richter & de Lange, 2019; Richter et al., 2018; Gheysen, Van Opstal, Roggeman, Van Waelvelde, & Fias, 2011; Meyer & Olson, 2011; Turk-Browne, Scholl, Johnson, & Chun, 2010; den Ouden et al., 2009; Turk-Browne, Scholl, Chun, & Johnson, 2009), it remains unclear how the sensory brain utilizes simultaneous sources of prior information to predict sensory input, such as spatial and temporal context. One human fMRI study suggests that a similar network of (sub)cortical areas is involved in learning spatial contexts as during learning of temporal sequences (Karuza et al., 2017). Thus, although the learning process of temporal and spatial regularities may share neural characteristics, it is unclear whether, following the acquisition of spatio-temporal regularities, they have similar effects on

[1]Peking University, China, [2]Radboud University, The Netherlands, [3]Zhejiang University, China

sensory processing. In particular, do predictions of spatial context result in a similar suppression of neural responses as temporal sequence predictions? Moreover, it is currently unclear if and how spatial and temporal context may interact in sharpening sensory processing.

In the current study, we set out to concurrently examine the neural and behavioral consequences of spatial and temporal contextual expectations following statistical learning. To this end, participants were exposed to leading image pairs, consisting of two object images presented left and right of fixation; these predicted the identity of trailing object image pairs, thus rendering the trailing images expected based on the temporal context. Moreover, the simultaneously presented images were also predictive of each other, thus generating a predictable spatial context (see Figure 1C). BOLD signals were recorded with fMRI while participants monitored the images for occasional target images (i.e., flipped object images) that occurred at unpredictable moments.

To preview our results, we show that spatial and temporal context both modulate sensory processing in the ventral visual stream, with pronounced reductions in neural responses to stimuli predicted by spatial and temporal context, compared with stimuli occurring in unexpected contexts. Interestingly, these modulations were evident in comparable cortical areas, suggesting that spatial and temporal context modulate sensory processing in a similar fashion.

METHODS

Data and Code Availability

All data and code used for stimulus presentation and analysis are freely available on the Donders Repository (https://webdav.data.donders.ru.nl/dccn/DSC_3018034.03_765_v1).

Participants

Thirty-three healthy, right-handed participants (13 women; age mean = 22.36 years, *SD* = 2.38 years) were recruited in exchange for monetary compensation (100 yuan/hr). All participants reported normal or corrected-to-normal vision, were prescreened for MRI compatibility, and had no history of epilepsy or cardiac problems. The experiments reported here were approved by the Institutional Review Board of Psychological Sciences at Hangzhou

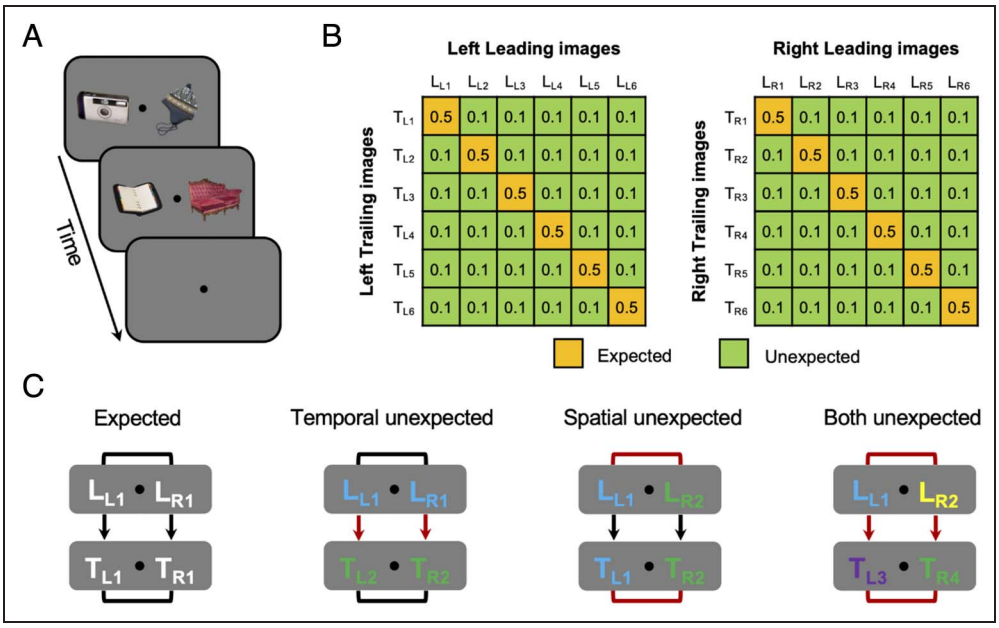

Figure 1. Experimental paradigm and design. (A) Experimental paradigm in both the behavioral learning and fMRI session. A trial starts with a 500-msec presentation of two leading images, presented to left and right of the central fixation dot. The two leading images are immediately followed by the trailing images, without ISI, at the same locations, also shown for 500 msec. Participants were asked to detect an infrequently presented upside down version of the images (~10% of trials). Trials were separated by a 2- to 6-sec (mean 3 sec) ITI period. (B) Shown are the image transition matrices determining the statistical regularities between leading and trailing images during MRI scanning. On the left, L_{L1} to L_{L6} represent the six leading images presented on the left of the fixation dot, and T_{L1} to T_{L6} represent the associated six left trailing images. Similarly, L_{R1} to L_{R6} represent the six right leading images, and T_{R1} to T_{R6} represent the six right trailing images. Yellow cells indicate image pairs that are expected by temporal context, and green cells denote unexpected image pairs. Numbers represent the probability of that cell during MRI scanning. Crucially, the left and right images were also associated with each other, constituting the spatial context. For instance, L_{L1} was associated with L_{R1}, and T_{L1} was associated with T_{R1}. In this case, L_{L1}, L_{R1}, T_{L1}, and T_{R1} composed two image pairs that were expected in both the temporal and spatial contexts (see Figure 1C, "Expected"). (C) Illustration of the four expectation conditions during MRI scanning. Black lines indicate expected associations, and red lines indicate unexpected pairings. *Expected condition*: the matched image configuration that was shown during the behavioral learning session. *Temporally unexpected context*: both the two leading images (L_{L1} and L_{R1}) and two trailing images (T_{L2} and T_{R2}) were expected in terms of spatial context (same as the expected condition); the temporal association was violated (i.e., $L_{L1} \rightarrow T_{L2}$ and $L_{R1} \rightarrow T_{R2}$). *Spatially unexpected context*: Although the leading image reliably predicted the identity of the trailing image on both the left ($L_{L1} \rightarrow T_{L1}$) and right ($L_{R2} \rightarrow T_{R2}$) side independently, thus retaining the expected temporal context, image pairs were not associated in terms of spatial context, neither during the leading images nor during the two trailing images (e.g., L_{L1} and L_{R2} occurring together). *Both unexpected*: Shown were four images that do not appear together in the expected condition. Therefore, the expectation violations occurred in both the temporal and spatial contexts.

Normal University and were carried out in accordance with the guidelines expressed in the Declaration of Helsinki. Written informed consent was obtained from all participants. Data from two participants were excluded: One participant's behavioral performance of the postscanning task was at chance level, and the other participant showed excessive head motion (i.e., a number of relative head motion events exceeding 1 mm, notably above the group mean).

Stimuli

The object images were a selection of stimuli from the study of Brady, Konkle, Alvarez, and Oliva (2008) and also previously used by Richter and de Lange (2019). A subset of 48 full-color object stimuli, composed of 24 electronic objects and 24 non-electronic objects, were shown during this study. For each participant, 24 objects (12 electronics and 12 non-electronics) were pseudorandomly selected, of which 6 (including 3 electronics) were pseudorandomly assigned as left leading images, 6 (including 3 electronics) were appointed as right leading images, another 6 (including 3 electronics) served as left trailing images, and the remaining 6 (including 3 electronics) acted as right trailing images. Therefore, each specific image could occur in any position or condition (left or right, leading or trailing) across participants, thereby minimizing potential biases by specific features of individual object stimuli. Image size was 5° × 5° visual angle presented on a mid-gray background. Stimuli and their association remained the same during the behavioral learning session, MRI scanning, and a postscanning object categorization task. During the behavioral learning session and postscanning test, object stimuli were presented on an LCD screen (ASUS VG278q, 1920 × 1080 pixel resolution, 60-Hz refresh rate). During MRI scanning, stimuli were displayed on a rear-projection MRI-compatible screen (SAMRTEC SA-9900 projector, 1024 × 768 pixel resolution, 60-Hz refresh rate), visible using an adjustable mirror mounted on the head coil.

Experimental Design

Each participant completed two sessions on two consecutive days. The first session comprised a behavioral learning task and the second session included an fMRI task and a postscanning object categorization task. Although the stimuli and their associations were identical during both sessions, different tasks were employed.

Day 1: Learning Session

Each trial began with a black fixation dot (diameter = 0.4° visual angle) in the center of the screen; participants were asked to maintain fixation on the fixation dot throughout the trial. Two leading images were presented 4° visual angle left and right from the central fixation dot for 500 msec, immediately followed by two trailing images, without ISI, at the same locations for 500 msec (Figure 1A). Participants were required to count the pairs of same category objects (electronic vs. non-electronic) shown during the leading and trailing images and respond within 2000 msec after trailing image onset by pressing one of three response buttons (corresponding to none, one, or both; see Pair Counting Task section for details). Finally, feedback was presented for 500 msec, followed by a 1000- to 2000-msec intertrial interval (ITI). Twenty-four object images (12 electronics and 12 non-electronics) were pseudorandomly preselected per participant from a pool of images, 12 of which were pseudorandomly combined into pairs, forming a total of 6 leading image pairs (i.e., the first two images on a trial), and the remaining 6 pairs were used as trailing image pairs (i.e., the second two images on a trial). Crucially, during the learning session, the leading image pair was perfectly predictive of the identity of the trailing image pair [P(trailing pair | leading pair) = 1]. At the same time, the left and right images within both the leading and trailing image pairs were 100% predictive of one another (i.e., pairs always occurred together), thus resulting in deterministic association in both spatial (co-occurrence) and temporal (sequence) contexts during the learning session (see the leftmost panel in Figure 1C). During the learning session, each participant performed five blocks, with each block composed of 216 trials, resulting in a total of 180 trials per pair during learning session. The learning session took approximately 60 min.

Day 2: fMRI Session

One day after the learning session, participants performed the fMRI session. This session started with one additional block identical to the behavioral learning session, including 216 trials, to renew the learned associations before MRI scanning. During MRI scanning, participants first performed 36 practice trials during acquisition of the anatomical image. The fMRI session was similar to the behavioral learning session, except for the following three modifications. First, a longer ITI of 2000–6000 msec (mean = 3000 msec) was used. Second, instead of counting pairs of the same category, participants were required to detect oddball images. Oddballs were the same object images, as shown before, but flipped upside down, occurring on 10% of trials. Participants were instructed to respond to these target images by pressing a button as quickly as possible; no response was required during trials without an oddball image. Crucially, whether an image was upside down was completely randomized and could not be predicted on the basis of the statistical regularities that were present in the image sequences. Third, although the association between images remained the same as

during the behavioral learning session, in the fMRI session unexpected image pairs were also shown. In particular, the transition matrices shown in Figure 1B determined how often images were presented together. In 50% of trials, a leading image pair was followed by its expected trailing image pair, identical to the learning session, thus constituting the expected condition. For instance, L_{L1} (leading image, left 1) and L_{R1} (leading image, right 1) served as leading image pair for T_{L1} (trailing image, left 1) and T_{R1} (trailing image, right 1). In the other half of trials, one of the three unexpected conditions (temporally unexpected context, spatially unexpected context, and both temporally and spatially unexpected context) occurred with equal possibilities, resulting in 16.67% per unexpected condition. Specifically, for the temporal unexpected context (Figure 1C left middle panel), after presenting a leading image pair, one of the other five unmatched trailing image pairs would appear. Thus, while the two images within both the leading and trailing image pair were still expected (i.e., no spatial expectation violation), the temporal sequence of images was unexpected. For example, in this condition L_{L1} and L_{R1} were followed by T_{L2} and T_{R2}. For the spatially unexpected context (Figure 1C right middle panel), each leading image was followed by its expected trailing image (e.g., $L_{L1} \rightarrow T_{L1}$ and $L_{R2} \rightarrow T_{R2}$). However, the two images presented during both the leading and trailing image period were not usually paired (e.g., $L_{L1} \times L_{R2}$ and $T_{L1} \times T_{R2}$). Thus, in this condition, spatial context expectations were violated and temporal context was expected, thus constituting the spatially unexpected condition. In a final condition, both spatial and temporal context were violated (Figure 1C rightmost panel). In particular, all four images shown during this condition did not appear together in the learning session. Crucially, the expectation status only depended on the usual association between the leading image pair and trailing image pair, rather than the frequency or identity of an object image per se. In other words, each object image occurred as expected object and in each unexpected condition. Therefore, all images occurred equally often throughout the experiment, ruling out potential confounds of stimulus frequency or familiarity. Feedback on behavioral performance (accuracy) was provided after each run.

During MRI scanning, each run consisted of 108 trials, including 54 expected trials, 18 temporal context violation trials, 18 spatial context violation trials, and 18 trials where both spatial and temporal context were violated. The order of trials was randomized within each run. In total, each participant performed five runs. Each run lasted ~12 min with five null events of 12 sec that were evenly distributed across the run, which also served as brief resting periods. The first 8 sec of fixation were discarded from analysis. Finally, after MRI scanning, a pair counting task, identical to the learning session, was performed outside of the MRI scanner room, which took approximately 20 min (see Pair Counting Task section for details).

Functional Localizer

Following the main task runs during the fMRI session, two functional localizer runs were scanned. These localizer runs were used to define object-selective lateral occipital complex (LOC) and to select voxels that were maximally responsive to the relevant object images. For each participant, the same 12 trailing images that were previously seen in the main task runs and their phase-scrambled version were presented during the localizer. Images were presented to the left and right of the center of the screen, corresponding to the location where the stimuli were shown during the main task runs. Each image was shown for 11 sec, alternating between the left and right side. Images flashed with a frequency of 2 Hz (300 msec on, 200 msec off). Throughout the localizer, participants were instructed to fixate the fixation dot, while monitoring for an unpredictable dimming of the stimulus (dimming period = 300 msec). Participants responded as quickly as possible by pressing a button. In each run, four null events of 11 sec were evenly inserted, and each trailing image and its phase-scrambled version was presented twice. The order of trials was fully randomized, except for excluding direct repetitions of the same image. Each participant completed two localizer runs, with each run lasting ~9.5 min. In total, each image and its phase-scrambled version was presented four times.

Pair Counting Task

Because the oddball detection performed during fMRI scanning does not relate to the underlying statistical regularities and, therefore, does not indicate whether statistical regularities were indeed learned, an additional pair counting task was performed after fMRI scanning. In this task, participants were asked to count the number of pairs of the same object category shown on each trial. Given that two successive image pairs (i.e., leading and trailing image pairs) were presented in each trial, participants need to make a three-alternative choice (corresponding to none, one, or both). Participants were further instructed to respond as quickly and accurately as possible. Thus, this task was the same as the task performed during the behavioral learning session, except that the three unexpected conditions were also included. The rationale of this task was to gauge the learning of the object pairs (i.e., statistical regularities) in terms of both temporal and spatial context. Participants could benefit from the knowledge of the associations between the image pairs, as both knowledge about the co-occurrence and temporal sequence would allow for faster responses. Therefore, the performance difference (e.g., accuracy and RT) between the expected condition and each unexpected condition could be considered as an indication for having learned the underlying statistical regularities. In total, participants performed 360 trials split into two

blocks, including 180 expected trials, 60 temporally unexpected context trials, 60 spatially unexpected context trials, and 60 trials in which both spatial and temporal context were unexpected. The pair counting task took approximately 20 min.

fMRI Parameters

Functional and anatomical images were acquired on a 3.0 T GE MRI-750 system (GE Medical Systems) at Hangzhou Normal University, using a standard 8-channel headcoil. Functional images were acquired in a sequential (ascending) order using a T2*-weighted gradient-echo EPI pulse sequence (repetition time/echo time = 2000/30 msec, voxel size 2.5 × 2.5 × 2.3 mm, 0.2-mm slice space, 36 transversal slices, 75° flip angle, field of view = 240 mm^2). Anatomical images were acquired using a T1-weighted inversion prepared 3-D spoiled gradient echo sequence (inversion time = 450 msec, repetition time/ echo time = 8.2/3.1 msec, field of view = 256 × 256 mm^2, voxel size 1 × 1 × 1 mm, 176 transversal slices, 8° flip angle, parallel acceleration = 2).

Data Analysis

Behavioral Data Analysis

Behavioral data from the pair counting task were analyzed in terms of response accuracy and RT. RT was calculated relative to the onset of the trailing image objects. Only trials with correct responses were included in RT analysis. In addition, we excluded trials with RTs shorter than 200 msec (0.82%) or more than 3 *SD*s above the participant's mean response time (0.49%). RT and accuracy data for expected and unexpected trailing image trials were averaged separately per participant and across participants subjected to a paired *t* test. The effect size was calculated in terms of Cohen's d_z for all paired *t* tests, and partial eta-squared (η^2) was used for indicating effect sizes in the repeated-measures ANOVA (Lakens, 2013).

fMRI Data Preprocessing

fMRI data preprocessing was performed using FMRIB Software Library (FSL 6.0.1; www.fmrib.ox.ac.uk/fsl; Smith et al., 2004, RRID:SCR_002823). The preprocessing pipeline included brain extraction, motion correction (MCFLIRT), slice timing correction (regular up), temporal high-pass filtering (128 sec), and spatial smoothing for univariate analyses (Gaussian kernel with FWHM of 5 mm). Functional images were registered to the anatomical image using FSL FLIRT (boundary-based registration) and to the MNI152 T1 2-mm template brain (linear registration with 12 degrees of freedom). Registration to the MNI152 template brain was only applied for whole-brain analyses, whereas all ROI analyses were performed in each participant's native space in order to minimize data interpolation.

Whole-Brain Analysis

To estimate the BOLD response to expected and unexpected stimuli across the entire brain, FSL FEAT was used to fit voxel-wise general linear models (GLMs) to each participant's run data in an event-related approach. In the first-level GLMs, expected and three unexpected image object trials were modeled as four separate regressors with a duration of 1 sec (the combined duration of leading and trailing image pairs) and convolved with a double gamma hemodynamic response function. An additional nuisance regressor for oddball trials (upside down images) was added. In addition, first-order temporal derivatives for the five regressors and 24 motion regressors (FSL's standard + extended motion parameters) were also added to the GLM. To quantify the main effects of spatial and temporal expectation suppression (ES), we contrasted unexpected regressors and the expected regressors for spatial and temporal context separately (i.e., temporal context ES = $BOLD_{\text{Temporal unexpected}} + BOLD_{\text{Both unexpected}} - BOLD_{\text{Spatial unexpected}} - BOLD_{\text{Both expected}}$; spatial context ES = $BOLD_{\text{Spatial unexpected}} + BOLD_{\text{Both unexpected}} - BOLD_{\text{Temporal unexpected}} - BOLD_{\text{Both expected}}$). Data were combined across runs using FSL's fixed effects analysis. For the across-participants whole-brain analysis, FSL's mixed effect model (FLAME 1) was used. Multiple-comparison correction was performed using Gaussian random-field-based cluster thresholding. The significance level was set at a cluster-forming threshold of $z > 3.1$ (i.e., $p < .001$, two-sided) and a cluster significance threshold of $p < .05$.

ROI Analysis

ROI analyses were conducted in each participant's native space. Primary visual cortex (V1), object-selective LOC, and temporal occipital fusiform cortex (TOFC) were chosen as the three ROIs (see ROI definition section) for analysis, based on two previous studies that used a similar experimental design (Richter & de Lange, 2019; Richter et al., 2018). The mean parameter estimates were extracted from each ROI for the expected and unexpected conditions separately. For each ROI, these data were submitted to a two-way repeated-measures ANOVA with Temporal Context (expected vs. unexpected) and Spatial Context (expected vs. unexpected) as factors.

ROI definition. All ROIs were defined using independent data from the localizer runs. Specifically, V1 was defined based on each participant's anatomical image, using Freesurfer 6.0 to define the gray–white matter boundary and perform cortical surface reconstruction (*recon-all*; Dale, Fischl, & Sereno, 1999; RRID:SCR_001847). The resulting surface-based ROI of V1 was then transformed into the participant's native space and merged into one bilateral mask. Object-selective LOC was defined as bilateral clusters within anatomical LOC showing a

significant preference for intact compared to scrambled object stimuli during the localizer run (Haushofer, Livingstone, & Kanwisher, 2008; Kourtzi & Kanwisher, 2001). To achieve this, intact objects and scrambled objects were modeled as two separate regressors in each participant's localizer data. The temporal derivatives of all regressors and the 24 motion regressors were also added to fit the data. Finally, the contrast of interest, objects minus scrambles, was constrained to anatomical LOC. In order to create the TOFC ROI mask, the anatomical temporal-occipital fusiform cortex mask from the Harvard–Oxford cortical atlas (RRID:SCR_001476), distributed with FSL, was further constrained to voxels showing a significant conjunction inference of ES on the group level in the works of Richter et al. (2018) and Richter and de Lange (2019). The resulting mask was then transformed from MNI space to each participant's native space using FSL FLIRT. The 200 most active voxels in each of the three ROI masks were selected for further statistical analyses. To this end, the contrast of interest between the left and right hemispheres in V1 (including both the intact and scrambled images) was calculated, whereas in LOC and TOFC, the contrast of interest between the intact images and the scrambled images was calculated based on the localizer data. The resulting z-map of this contrast was then averaged across runs. Finally, we selected the 200 most responsive voxel from this contrast. In order to verify that our results did not depend on the a priori defined, but arbitrary, number of voxels in the ROI masks, we repeated all ROI analyses with masks ranging from 50 to 500 voxels in steps of 50 voxels.

Bayesian Analysis

In order to further evaluate any nonsignificant results and arbitrate between an absence of evidence and evidence for the absence of an effect, the Bayesian equivalents of the above outlined analyses were additionally performed. JASP 0.10.2 (JASP Team, 2019, RRID:SCR_015823) was used to perform all Bayesian analyses, using default settings. Thus, for Bayesian t tests, a Cauchy prior width of 0.707 was chosen. Qualitative interpretations of Bayes Factors are based on criteria by Lee and Wagenmakers (2014).

RESULTS

We exposed participants to statistical regularities by presenting two successive object image pairs in which the leading image pairs predicted the identity of the trailing image pairs. The identities of the image pairs were also predictable in terms of their spatial context; that is, simultaneously shown left and right images occurred together. Subsequently, in the MRI scanner, participants were shown the same predictable object image pairs (expected condition), but additional expectation violations were introduced. In particular, either the temporal context was violated, the spatial context was violated, or both contexts were violated (see Figure 1C).

Stronger Modulation of Spatial Context than Temporal Context on Sensory Processing throughout the Ventral Visual Stream

In order to assess the consequences of violating temporal and spatial context expectations, we performed a two-way repeated-measures ANOVA with Temporal Context (expected vs. unexpected) and Spatial Context (expected vs. unexpected) as factors, within our a priori defined ROIs: primary visual cortex (V1), object-selective LOC, and TOFC. In higher visual areas, LOC and TOFC, we observed a significant decrease in BOLD responses when stimuli were expected in terms of their spatial context (Figure 2A; LOC: $F(1, 32) = 31.389$, $p = 3.0e{-}6$, $\eta^2 = .495$; TOFC: $F(1, 32) = 23.083$, $p = 3.5e{-}5$, $\eta^2 = .419$). In other words, when two stimuli frequently co-occurred, thus making them expected in this pair, they elicited reduced sensory responses in ventral visual areas. Furthermore, we found a similar suppression of neural responses by temporal context expectations in TOFC, $F(1, 32) = 10.805$, $p = .0025$, $\eta^2 = .252$, but not in LOC, $F(1, 32) = 1.266$, $p = .2689$, $\eta^2 = .038$. That is, in TOFC, if a pair of stimuli was expected given the preceding stimulus pair, the elicited BOLD response was suppressed compared with the response to the same pair occurring in an unexpected temporal sequence. No interaction between Temporal and Spatial Context was found in either LOC or TOFC (LOC: $F(1, 32) = 0.111$, $p = .7412$, $\eta^2 = .003$; TOFC: $F(1, 32) = 0.064$, $p = .8013$, $\eta^2 = .002$). Thus, the suppression of neural responses induced by temporal expectations was not modulated by spatial context expectations, and vice versa.

In a post hoc analysis, we compared the magnitude of neural suppression induced by temporal and spatial context predictions. In LOC and TOFC, spatial context expectations resulted in a larger suppression than temporal expectations (LOC: $t(32) = 2.870$, $p = .0072$, Cohen's $d_z = 0.835$; TOFC: $t(32) = 2.575$, $p = .0149$, Cohen's $d_z = 0.691$), thus suggesting that spatial context may be a stronger modulator of visual responses than temporal context.

Perhaps surprisingly, we did not find any reliable modulation of neural responses by temporal or spatial context predictions in V1 (spatial context: $F(1, 32) = 2.448$, $p = .1275$, $\eta^2 = .071$; Temporal Context: $F(1, 32) = 0.154$, $p = .6971$, $\eta^2 = .005$; Spatial Context × Temporal Interaction: $F(1. 32) = .627$, $p = .4342$, $\eta^2 = .019$). Indeed, in V1, Bayesian analyses yielded moderate evidence for the absence of a modulation of neural responses by temporal context violations (temporally unexpected context vs. expected context: $BF_{10} = 0.141$), and anecdotal support for the absent of an effect when spatial context was violated (spatially unexpected context vs. expected context:

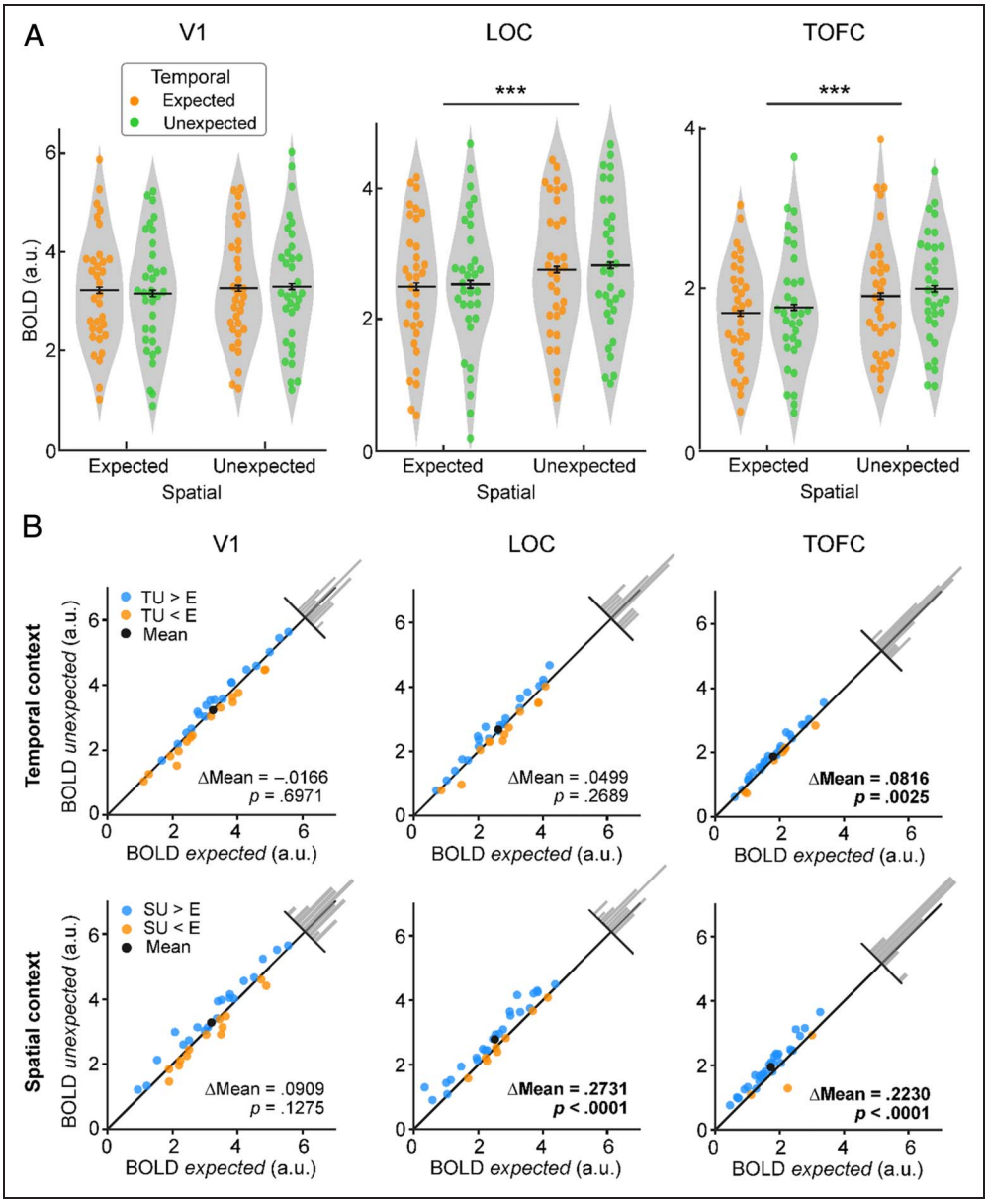

Figure 2. ES within V1, LOC, and TOFC. (A) Parameter estimates for responses to expected and unexpected images pairs. In both LOC and TOFC, BOLD responses to spatially expected image pairs were significantly attenuated compared with unexpected image pairs. Furthermore, a reliable suppression of responses by temporal context expectations was observed in TOFC. No modulation of BOLD responses by expectations was found in V1. Each dot denotes an individual participant, and the black line is the mean across participants. Error bars denote ±1 within-subject *SEM*. *p < .05, **p < .01, ***p < .001. (B) BOLD responses evoked by unexpected and expected context within V1 (left column), LOC (middle column), and TOFC (right column). The upper row represents the BOLD contrast between the temporally unexpected context and expected context, averaged across the spatially expected and unexpected context. The bottom row represents the BOLD contrast between the spatially unexpected context and expected context, averaged across the temporally expected and unexpected context. Blue and yellow dots represent individual participants. Blue indicates ES (unexpected > expected), yellow indicates expectation enhancement (unexpected < expected), and black indicates the mean of all participants. ΔMean is equal to the difference of BOLD response between the unexpected and expected conditions. The inset histogram shows the distribution of deviations from the unity line. a.u. = arbitrary units.

BF_{10} = 0.388). Thus, expectations, in terms of temporal or spatial context, did not appear to modulate sensory responses in V1. In contrast, in higher visual areas, a suppression of responses to expected stimuli was observed both for temporal and spatial contexts.

To ensure that our results were not dependent on the a priori but arbitrarily chosen mask sizes of the ROIs, we repeated the analyses for ROIs of sizes ranging from 50 to 500 voxels in steps of 50 voxels. Results, summarized in Appendix Figure A1, were qualitatively identical to those mentioned above (Figure 2A) for all ROI sizes within all three ROIs (V1, LOC, and TOFC), indicating that our results do not depend on ROI size, but represent results within the ROIs well.

A complementary whole-brain analysis was performed to investigate the effect of temporal context and spatial context outside of our predefined ROIs. Results are illustrated in Figure 3. In accordance with our ROI analysis, spatial expectations were associated with significantly suppressed neural responses throughout the ventral visual stream. Additional clusters of ES were evident outside the ventral visual stream, including bilateral frontal gyrus, bilateral precentral gyrus, bilateral frontal operculum and insular cortex, as well as the paracingulate gyrus. In contrast, no reliable

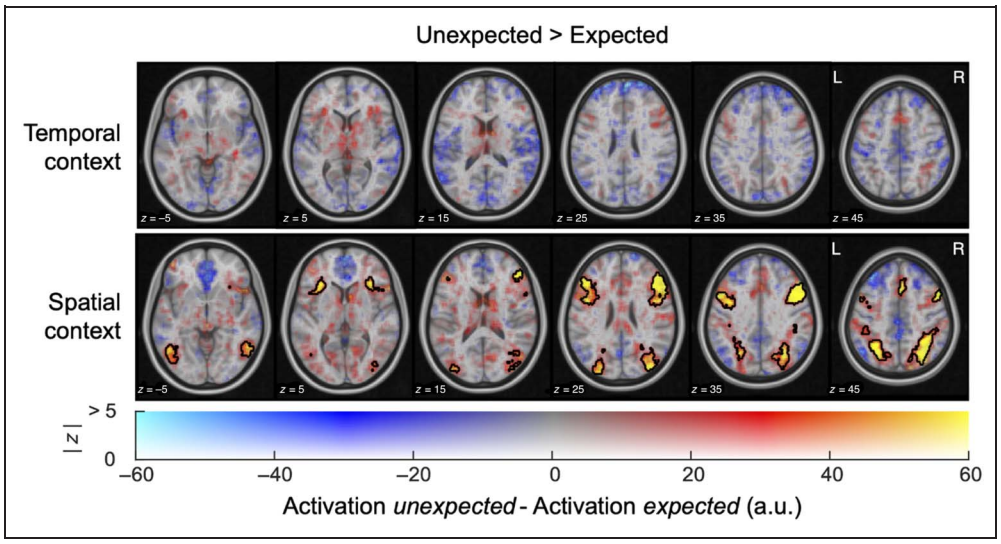

Figure 3. ES across cortex for temporal and spatial contexts. Displayed are parameter estimates for unexpected minus expected image pairs overlaid onto the MNI152 2-mm anatomical template. Color represents the unthresholded parameter estimates: red-yellow clusters denote ES, blue-cyan clusters indicate expectation enhancement; opacity indicates the z statistics of the contrast. Black contours outline statistically significant clusters (Gaussian random field cluster corrected). No significant clusters were found for the main effect of temporal context (upper row). The main effect of spatial expectation (bottom row) shows significant clusters of ES in parts of the ventral visual stream (LOC, TOFC), as well as bilateral frontal gyrus, bilateral precentral gyrus, bilateral frontal operculum and insular cortex, and paracingulate gyrus.

modulation by temporal context expectation was found outside of our predefined ROIs in the whole-brain analysis. Thus, temporal context expectations were only evident in the ROI analysis, but too small or hidden by interindividual variability to be detected in the whole-brain analysis (note: ROI masks were individually defined for each participant; also see Methods: ROI definition).

Expectations Facilitate Object Categorization

In addition to the neural effects of expectations, we also examined whether expectations facilitated behavioral responses. As shown in Figure 4A, during a postscanning object categorization task, participants were asked to count the number of object pairs of the same category shown as leading and trailing image pairs (i.e., zero, one, or two pairs could be of the same category). In order to fulfill this task as quickly and accurately as possible, participants could benefit from the knowledge of the underlying statistical regularities—both in terms of co-occurrence (spatial) and sequence (temporal) prediction. In line with our hypothesis, RTs and accuracy of responses (Figure 4B) were affected by expectations, in both temporal (RT: $t(32) = 4.891$, $p = 6.9e{-}6$, Cohen's d_z =

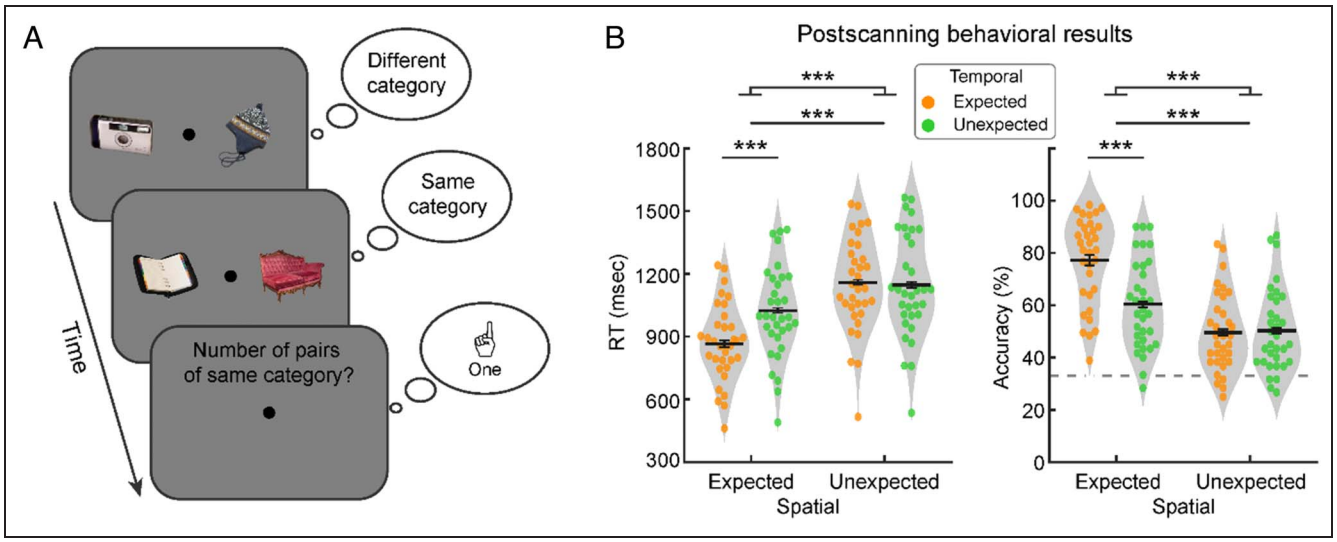

Figure 4. Behavioral paradigm and behavioral data indicate statistical learning. (A) In the postscanning task, participants performed an object categorization task; they were asked to count the number of object pairs of the same category (electronic vs. non-electronic) shown as leading and trailing image pairs (i.e., zero, one, or two pairs could be of the same category). (B) RT (left) and accuracy (right) are plotted for expected and unexpected conditions in temporal (dot color) and spatial contexts (abscissa), respectively. Behavioral responses in the spatially expected condition are significantly faster and more accurate than in the unexpected condition. Temporally expected stimulus pairs also result in faster and more accurate responses; however, this effect is only present when spatial expectations were met. Dashed horizontal gray line indicates chance level accuracy (33.33%). Dots represent single participant data. Black line is the mean across participants. Error bars denote ±1 within-subject *SEM*. ***$p < .001$.

0.851; accuracy: $t(32) = 4.924$, $p = 6.1e{-}6$, Cohen's $d_z = 0.857$) and spatial contexts (RT: $t(32) = 11.670$, $p = 1.3e{-}17$, Cohen's $d_z = 2.031$; accuracy: $t(32) = 10.224$, $p = 3.7e{-}15$, Cohen's $d_z = 1.780$). Thus, participants learned and benefitted from both spatial and temporal context predictions.

Interestingly, participants were faster and more accurate in response to objects predicted by the temporal sequence only when the spatial context was expected as well (RT: $t(32) = 9.329$, $p = 1.2e{-}10$, Cohen's $d_z = 1.624$; accuracy: $t(32) = 7.649$, $p = 1.0e{-}8$, Cohen's $d_z = 1.332$), but not when the spatial context was unexpected (RT: $t(32) = 0.269$, $p = .7898$, Cohen's $d_z = 0.047$, $BF_{10} = 0.193$; accuracy: $t(32) = 0.566$, $p = .5755$, Cohen's $d_z = 0.099$, $BF_{10} = 0.216$). The robustness of this distinct pattern of facilitation effect was statistically confirmed by an interaction analysis (RT: $F(1, 32) = 38.787$, $p = 5.6e{-}7$, $\eta^2 = .548$; accuracy: $F(1, 32) = 46.337$, $p = 1.1e{-}7$, $\eta^2 = .592$). Moreover, when a stimulus was expected by spatial context, participants showed faster and more accurate responses, irrespective of whether the temporal context was expected (RT: $t(32) = 13.977$, $p = 3.6e{-}15$, Cohen's $d_z = 2.433$; accuracy: $t(32) = 10.883$, $p = 2.7e{-}12$, Cohen's $d_z = 1.894$) or unexpected (RT: $t(32) = 5.838$, $p = 1.7e{-}6$, Cohen's $d_z = 1.016$; accuracy: $t(32) = 6.279$, $p = 4.9e{-}7$, Cohen's $d_z = 1.093$).

In summary, behavioral performance was reliably facilitated by spatial context, resulting in faster and more accurate responses. On the other hand, expected temporal sequences also aided in faster and more accurate responses, however, only if the spatial context was expected. These results may suggest that participants grouped pairs of objects and predicted the upcoming pair of objects, instead of individual sequences of objects on the left and right side separately.

DISCUSSION

Both spatial and temporal context play an important role in visual perception and behavior (Schwartz, Hsu, & Dayan, 2007). This study investigated the neural consequences of violations of expectations derived from spatial and temporal context, across the ventral visual stream. To this end, we exposed participants to two forms of statistical regularities, making stimuli predictable in terms of spatial context (co-occurrence of stimuli at specific locations) and temporal context (specific temporal sequence of stimuli). Although we measured brain activity to these stimuli, image transitions were not task relevant, and thus any neural modulations by spatial and temporal context were not dependent on task relevance of the underlying statistical regularities. We found a reliable and widespread activity modulation in the ventral visual stream, including LOC and TOFC, as a function of spatial context. In particular, when stimuli frequently co-occurred, neural responses were suppressed compared to the response to the same stimulus co-occurring with another stimulus, although all stimuli were equally familiar and always occurred at the same spatial location. Temporal context (i.e., predictability of stimulus sequence) also modulated neural responses in TOFC, again evident as a suppression of responses to expected stimuli. Thereby, our results extend previous studies (e.g., Richter & de Lange, 2019; Richter et al., 2018; Kok, Jehee, & de Lange, 2012; Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010; Summerfield et al., 2008) by demonstrating that spatial and temporal context priors may modulate neural responses in a similar fashion and within the same cortical network.

Spatial and Temporal Context Facilitate Behavior

Our data showed a substantial and robust facilitation of behavioral responses by both spatial and temporal context. During a postscanning test, requiring participants to count stimulus pairs of the same category (i.e., both electronic, or both non-electronic stimuli), spatial and temporal context strongly modulated behavioral performance (Figure 4B). Specifically, responses were faster and more accurate to stimuli presented in a spatially and temporally expected context, and the violation of either context increased RTs and decreased response accuracy —with larger decrements for spatial context violations. Crucially, the benefit of temporally expected contexts was only observed when the spatial context was expected. However, performance enhanced by spatially expected contexts was evident irrespective of whether temporal context expectations were confirmed or violated.

Thus, our data show that participants can, in principle, learn and benefit from both spatial and temporal statistical regularities. However, our results also suggest that our participants grouped simultaneously presented objects into image pairs, which together predicted the next image pair. That is, although object stimuli on the left and right side predicted the identity of the next stimulus independently from each other, even when the spatial configuration was unexpected, these (arguably simpler) statistical regularities may not have been learned, or the resulting predictions may not have been instantiated. This notion is in line with an earlier study that found robust statistical learning of stimulus pairs, particularly when the shapes were grouped by a visual cue (Baker, Olson, & Behrmann, 2004). Therefore, our results may suggest a dominance of spatial over temporal grouping in vision. However, it is important to note here that a strategy of grouping spatial pairs may have partially been induced by the same–different category counting task during learning, which specifically requires participants to make a judgment about the groups of objects.

Spatial and Temporal Context Modulate Sensory Processing in the Ventral Visual Stream

Our fMRI results show that sensory responses in object-selective visual areas (LOC and TOFC) are suppressed, if

stimuli occur in expected spatial contexts compared to unexpected spatial contexts. In other words, stimuli that frequently co-occur evoked reduced sensory responses relative to the same stimuli presented in less frequently co-occurring configurations. Note that the frequency of the individual stimuli occurring were equal, thereby excluding potentially confounding effects of stimulus frequency or familiarity. Moreover, during MRI scanning, predictions were task-irrelevant, thus suggesting that predictions were formed and modulated neural responses without participants being required to retrieve these predictions, and without the associations being helpful for task purposes.

The suppression of neural responses by spatial predictions matches key characteristics of ES, a phenomenon previously described in terms of suppressed sensory responses to stimuli expected by virtue of their temporal context; that is, a leading image predicting the identity of a trailing image (Richter & de Lange, 2019; Richter et al., 2018; Meyer & Olson, 2011; den Ouden et al., 2009). In line with previous studies, we also found a suppression of sensory responses by temporal context in TOFC. That is, stimuli in expected temporal sequences elicited suppressed BOLD responses compared with stimuli in unexpected temporal sequences.

In addition, the observed suppression of neural responses by spatial expectations were evident in ventral visual areas similar to those previously reported by studies investigating temporal context expectations (e.g., Richter & de Lange, 2019; Richter et al., 2018; Gheysen et al., 2011; Meyer & Olson, 2011; Turk-Browne et al., 2009, 2010; den Ouden et al., 2009). Interestingly, this overlap in cortical regions was not limited to object-selective visual cortex, but also included several nonsensory areas (e.g., frontal gyrus, precentral gyrus). This finding is consistent with those previous studies that observed a temporal context ES in similar frontal and parietal cortex (Richter & de Lange, 2019; Fletcher et al., 2001). A critical difference between those studies and this study is the manipulation of the task during MRI scanning. Whereas predictions in their studies were task-relevant, predictions were task-irrelevant in this study (detection of unpredictable oddballs). This suggests that predictions induce or even depend on modulations of neural activity in frontal areas, such as inferior frontal and bilateral precentral gyrus, independent of task relevance. Indeed, in line with this idea, Christiansen, Kelly, Shillcock, and Greenfield (2010) have shown that agrammatic aphasics with damage to the left frontal areas were unable to discriminate between test items in an AGL task, further suggesting that statistical learning may depend on computations in frontal areas (Wang, Uhrig, Jarraya, & Dehaene, 2015; Karuza et al., 2013; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006). Combined, these results suggest that spatial and temporal context have overlapping modulatory effects on neural processing, thereby implying that the neural mechanism underlying contextual prediction effects may be independent of the type of prediction—temporal or spatial context. In agreement with this suggestion, Karuza et al. (2017) reported similar neural modulations, and comparable correlations of these modulations with behavior, during learning of spatial regularities as previously reported for statistical learning of temporal (sequence) regularities (Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014; Gheysen et al., 2011; Gheysen, Van Opstal, Roggeman, Van Waelvelde, & Fias, 2010; Turk-Browne et al., 2009, 2010). Thus, the available data suggest that the neural architecture and computations underlying different types of context predictions may largely overlap, evident in similar modulations of both behavioral and neural responses.

A recent fMRI study demonstrated that spatial context associations are reflected in BOLD activity in three scene-selective brain regions: parahippocampal place area, retrosplenial complex, and occipital place area (Aminoff & Tarr, 2015). We did not find modulations in these areas by expectations. There are some notable differences in the design in the two studies, which may account for the apparent discrepancy in results. First, in their Spatial & Identity Associations condition, Aminoff and Tarr (2015) contrast learned scenes with random (not learned) arrangements, whereas we contrast expected with unexpected appearances of paired images. An expectation violation is crucially different from a random arrangement (not predictable), which might be reflected in differential cortical activation maps. Moreover, in the study of Aminoff and Tarr (2015), participants concurrently learned the spatial position and identity of the stimuli during the training phase, forming a small "scene." In contrast, in our design, violations in spatial context were defined by presenting a pair of images that are usually not shown together, but at the same spatial location. In other words, the location (left vs. right) was never violated for any image, only the co-occurrence of images differed. Thus, the spatial arrangement ("scene") was still intact, with only the identity of a stimulus being different. Therefore, our expectation violations may not recruit "scene" representations in the same way as the different spatial arrangements in the study of Aminoff and Tarr (2015).

Finally, it is an interesting question whether and how ES relates to behavioral facilitation induced by prediction. However, our present data are not ideally suited for assessing this question, given that predictions were task-irrelevant during MRI scanning, and the behavioral data were acquired in a separate postscanning session.

Identity- versus Location-Based Spatial Context ES

In this study, spatial context violations originated from changing the identity of objects while their locations were fixed—that is, spatial context refers to the spatial pairing of two objects. Such violations of spatial context,

relying on identity-based associations between concurrently presented objects, have also been investigated in previous studies (e.g., Munneke et al., 2013; Davenport, 2007; Davenport & Potter, 2004; Biederman, Mezzanotte, & Rabinowitz, 1982). However, in addition to identity-based associations, locations of objects are another sense of spatial context (e.g., airplanes in the sky) or relative positions (e.g., keyboards in front of computer monitors). Such location-based associations (for a review, see Kaiser et al., 2019) were not manipulated in this study; hence, we cannot draw conclusions whether such conceptually related but distinct spatial contexts modulate sensory processing in a similar fashion. Of note, a recent study showed that both types of spatial context associations jointly modulate object processing (Quek & Peelen, 2020), because both are needed for creating a coherent group representation. In addition, other studies showed several brain regions, including inferior prefrontal cortex, parahippocampal cortex, and LOC, may be involved in this interaction (Kaiser & Peelen, 2018; Gronau, Neta, & Bar, 2008).

Thus, although we did not manipulate object position, it is interesting to speculate what would happen if we flipped the left and right leading images or trailing images during the test phase. Given that both location and identity-dependent spatial context modulations have shown to interact, we believe that a surprising reversal of the position of two simultaneously presented objects will elicit stronger neural responses in similar cortical areas as shown in this study. Study of the interaction between different types of spatial context predictions as well as temporal predictions may pose an interesting avenue for future research.

Stronger Modulations of Neural Responses by Spatial Context than Temporal Context

Although the current data showed that a comparable neural mechanism may underlie both spatial and temporal context predictions, the modulation by temporal context was relatively modest compared to the modulation by spatial context. Initially, these results may be surprising given the multitude of previous studies reporting strong and extensive modulations of sensory responses by temporal context predictions across the ventral visual stream (Richter & de Lange, 2019; Richter et al., 2018; Plante et al., 2015; Tremblay, Baroni, & Hasson, 2013; Tobia, Iacovella, Davis, & Hasson, 2012; Tobia, Iacovella, & Hasson, 2012; Meyer & Olson, 2011; Gheysen et al., 2010; Turk-Browne et al., 2009, 2010). These previous studies, however, lacked spatial context, presenting single stimuli in isolation.

Vision is particularly apt to handle simultaneous inputs and the spatial structure between these stimuli (Saffran, 2002). Audition on the other hand shows a remarkable sensitivity to the temporal structure of inputs (Conway & Christiansen, 2009; Kubovy, 1988). Indeed, such modality-specific constraints can affect the manner in which stimuli are processed (Repp & Penel, 2002; Mahar, Mackenzie, & McNicol, 1994), maintained in working memory (Collier & Logan, 2000; Penney, 1989), and learned (Conway & Christiansen, 2009; Saffran, 2002; Handel & Buffardi, 1969). Thus, modality-specific biases in the visual system may result in an emphasis on spatial configurations and hence a stronger modulation of neural responses by spatial than temporal context predictions.

Our behavioral results also support the notion that spatial predictions were more readily acquired and utilized than temporal predictions. In particular, only when spatial configurations were expected, temporal predictions facilitated behavioral responses. Thus, in the present data, and possibly vision in general, spatial regularities appear to take precedence over temporal statistical regularities, resulting in a larger magnitude of behavioral and neural modulations by spatial compared to temporal context.

Although there may indeed be an increased sensitivity toward spatial regularities in vision, we also note that the specific task used here during learning may have promoted stronger learning of spatial than temporal relationships. Such bias during learning may therefore provide an alternative account of the stronger effect of spatial predictions in the present data. This concern could be addressed by future studies by utilizing a task that does involve task-irrelevant predictions already during training (e.g., oddball task), as at least temporal expectations are also formed by task-irrelevant associations (Richter et al., 2018).

No Modulation of Neural Responses by Prediction in Primary Visual Cortex

Surprisingly, we found no modulation by predictions in V1, unlike in some previous studies (e.g., Richter & de Lange, 2019; Kok et al., 2012). It is possible that, because expectations constitute a top–down modulation, likely originating from beyond visual cortex (Hindy, Avery, & Turk-Browne, 2019), its effect might be less pronounced in V1 compared to higher visual areas. Indeed, in previous studies, prediction effects appear to reduce in magnitude in lower visual areas (e.g., see Figure 1A in Richter & de Lange, 2019). Moreover, it is possible that spatial arrangements of object stimuli were too complex to yield specific predictions relevant to the response properties of neural assemblies in V1. That is, predictions in our study constitute arrangements and sequences of full-color object images, thus particularly depending on object-selective cortical areas. Hence, arrangements of stimuli exploiting the neural tuning in V1, for example, pairs of oriented grating stimuli, may result in prediction-induced modulations in V1. Thus, the absence of ES in V1 observed here may be a consequence of the utilized stimuli and experimental design.

Familiarity or Expectation?

Earlier in the paper, we interpreted the reduced response to spatially expected image pairs as stemming from a

confirmed expectation based on spatial context. That is, frequently co-occurring objects predict one another, thereby facilitating processing and resulting in attenuated sensory responses. However, one could also view the effect of spatial context as a consequence of increased familiarity of the scene. The visual scene of two spatially paired (expected) objects becomes more familiar than a scene of two rarely co-occurring (unexpected) object pairs. Familiarity has been shown to result in attenuated sensory responses (Manahova, Mostert, Kok, Schoffelen, & de Lange, 2018; Hayworth, Lescroart, & Biederman, 2011; Rossion, Schiltz, & Crommelinck, 2003; Baker, Behrmann, & Olson, 2002; Li, Miller, & Desimone, 1993; for a review, see Quek & Peelen, 2020). Therefore, it is plausible that scene familiarity may account for the reduced response to spatially expected object pairs. A similar reasoning applies to the temporal domain, given that participants become familiar with, and therefore learn to expect, certain transitions between objects. Although predictions are often formed by exposing an agent to certain regularities that occur more frequently, leading to an inseparable bond between familiarity (of spatial or temporal contingencies) and expectation (of these contingencies), these two factors can be empirically dissociated (Marcovitch & Lewkowicz, 2009). This may be an interesting avenue for future research.

Repetition or Expectation?

In addition to ES, repetition of the same stimulus is also associated with a reduction in neural responses, a phenomenon known as repetition suppression (RS) or adaptation (for a review, see Grill-Spector, Henson, & Martin, 2006). Although they share some characteristics, ES and RS are both theoretically and experimentally distinguishable. RS is commonly understood as a bottom–up effect; the reduced signal may be driven by the fatigue of a neuronal population responding to a particular stimulus, or sharpening of sensory representations (Grill-Spector et al., 2006). Conversely, ES is usually regarded as a top–down process—for example, in hierarchical predictive coding models (Wacongne et al., 2011; Friston, 2005; Rao & Ballard, 1999). Indeed, also experimentally, RS and ES have been dissociated. For example, Todorovic and de Lange (2012) reported that RS and ES occurred at distinct time windows, with an early RS effect and a later ES effect (see also Grotheer & Kovács, 2015). Moreover, Summerfield et al. (2008) showed that neural responses of RS can be modulated by ES. That is, BOLD signal differences between repeated and unrepeated (i.e., RS effects) stimuli were larger when stimulus repetitions are expected than when they are unexpected.

How does RS relate to the results presented here? First, it is important to note that all stimuli appeared equally often in this study and images were not repeated on adjacent trials. Thus, simple stimulus repetition cannot account for the results observed here. However, participants formed expectations about stimulus pairs by statistical learning—that is, unexpected image pairs appeared less often. Because image pairs were not repeated on adjacent trials, repetitions of the same image pairs would be separated by a substantial amount of time and, on average, many intervening trials. Therefore, any repetition effects would be distinct from classical short-term RS or adaptation, requiring long-term modulations. Although there is some evidence that repetitions of simple stimuli can result in sensory modulations across long timespans in early visual cortex (Fritsche, Solomon, & de Lange, 2021), it remains unclear whether similar modulations may hold for repetitions of complex stimulus pairs across extended periods, particularly in higher visual areas.

Conclusion

In conclusion, our data suggest that temporal and spatial statistical regularities jointly facilitate behavioral responses, leading to faster and more accurate responses. At the same time, predictions based on both forms of context modulate sensory responses, resulting in a suppression of responses to expected stimuli in a similar cortical network, including object-selective visual cortex.

APPENDIX

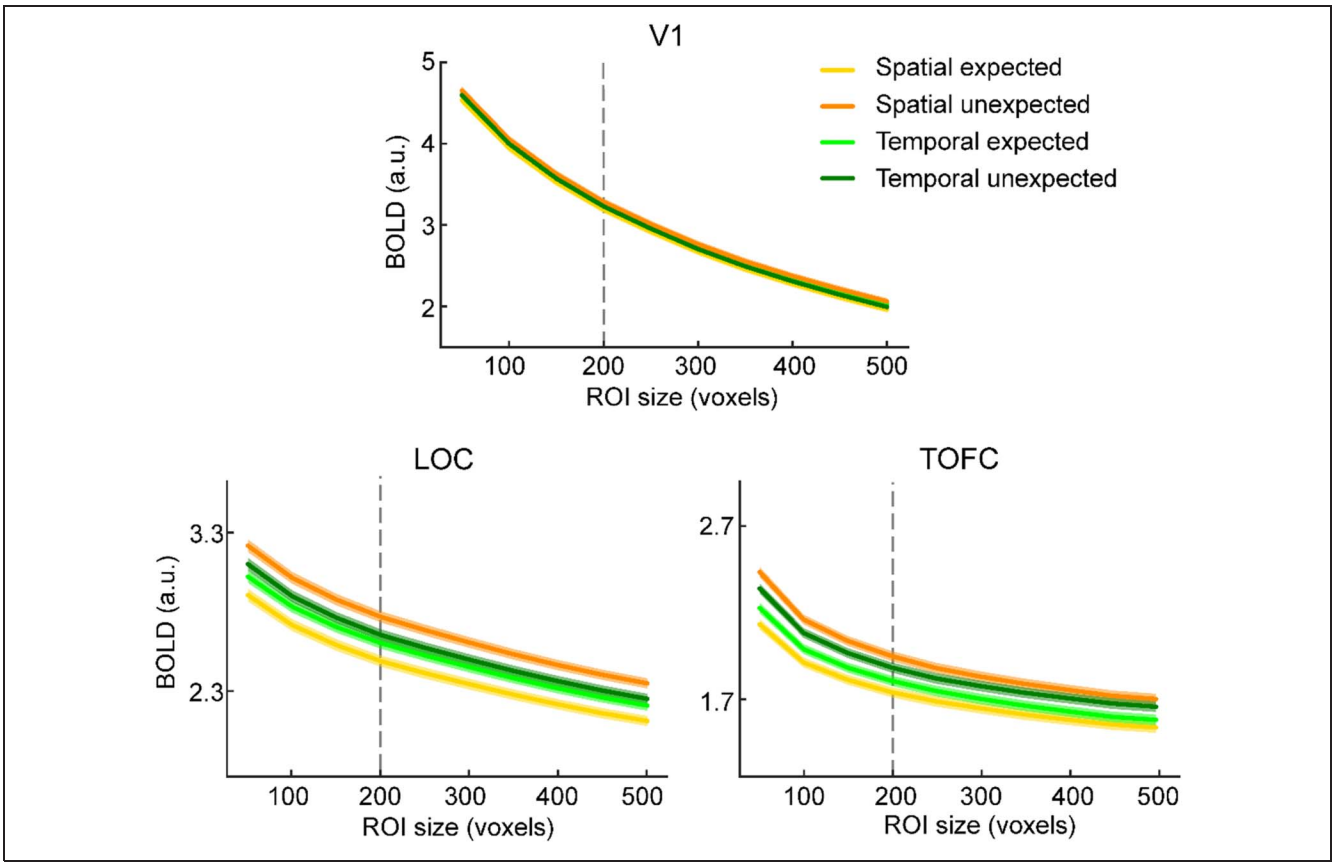

Figure A1. The differences of the BOLD responses between the unexpected and expected condition in our a priori defined ROIs (V1, LOC, and TOFC) are stable over a range of ROI sizes. The dashed vertical line is at the predefined ROI of 200 voxels. Same analysis as in Figure 2A, but performed over a wide range of ROI sizes, from 50 to 500 voxels, with steps of 50. Obviously, BOLD responses gradually decreased with the increase of the mask size of the ROI in all areas. However, the spatially unexpected image pairs elicited significantly stronger BOLD responses than expected image pairs in LOC and TOFC regions, all $F(1, 32) > 12.177$, all $p < .0014$, all $\eta^2 > .276$. No significant difference between the temporally unexpected and expected image pairs within V1 and LOC, all $F(1, 32) < 1.953$, all $p > .1719$, all $\eta^2 < .058$. The shaded areas denote ±1 within-subject *SEM*. a.u. = arbitrary unit.

Acknowledgments

We thank Wanlu Fu and Zehao Huang for assistance with data collection.

Reprint requests should be sent to Tao He, School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China, or via e-mail: t.he@pku.edu.cn or Zhiguo Wang, Center for Psychological Sciences, Zhejiang University, Hangzhou, 310058, China, or via e-mail: zhiguo@zju.edu.cn.

Author Contributions

T. H., D. R., Z. W., and F. P. d. L. designed research; T. H. performed research; T. H. and D. R. analyzed data; T. H. and D. R. wrote the first draft of the paper; T. H., D. R., Z. W., and F. P. d. L. edited the paper.

Funding Information

This work was supported by the National Natural Science Foundation of China (https://dx.doi.org/10.13039/501100001809), grant number: 31371133 to Z. W., The Netherlands Organisation for Scientific Research Vidi (https://dx.doi.org/10.13039/501100003246), grant number: 452-13-016 to F. P. d. L., the EC Horizon 2020 Program ERC Starting (https://dx.doi.org/10.13039/100010663), grant number: 678286 "Contextvision" to F. P. d. L., the James S. McDonnell Foundation (https://dx.doi.org/10.13039/100000913), grant number: 220020373 to F. P. d. L. and the China Scholarship Council (CSC; https://dx.doi.org/10.13039/501100004543), grant number: 201608330264 to T. H.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during

this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Alink, A., Schwiedrzik, C., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, *30*, 2960–2966. https://doi.org/10.1523/jneurosci.3730-10.2010, PubMed: 20181593

Aminoff, E. M., & Tarr, M. J. (2015). Associative processing is inherent in scene perception. *PLoS One*, *10*, e0128840. https://doi.org/10.1371/journal.pone.0128840, PubMed: 26070142

Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, *5*, 1210–1216. https://doi.org/10.1038/nn960, PubMed: 12379864

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science*, *15*, 460–466. https://doi.org/10.1111/j.0956-7976.2004.00702.x, PubMed: 15200630

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–629. https://doi.org/10.1038/nrn1476, PubMed: 15263892

Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1425–1431. https://doi.org/10.1037/a0027210, PubMed: 22329789

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143–177. https://doi.org/10.1016/0010-0285(82)90007-X, PubMed: 7083801

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, U.S.A.*, *105*, 14325–14329. https://doi.org/10.1073/pnas.0803390105, PubMed: 18787113

Castelhano, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science*, *27*, 606–621. https://doi.org/10.1177/0956797616629130, PubMed: 27022016

Christiansen, M. H., Kelly, L. M., Shillcock, R. C., & Greenfield, K. (2010). Impaired artificial grammar learning in agrammatism. *Cognition*, *116*, 382–393. https://doi.org/10.1016/j.cognition.2010.05.015, PubMed: 20605017

Collier, G. L., & Logan, G. (2000). Modality differences in short-term memory for rhythms. *Memory & Cognition*, *28*, 529–538. https://doi.org/10.3758/BF03201243, PubMed: 10946536

Conway, C. M., & Christiansen, M. H. (2009). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology*, *21*, 561–580. https://doi.org/10.1080/09541440802097951

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395, PubMed: 9931268

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*, 393–401. https://doi.org/10.3758/BF03193280, PubMed: 17691140

Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science*, *15*, 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x, PubMed: 15271002

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*, 1175–1185. https://doi.org/10.1093/cercor/bhn161, PubMed: 18820290

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, *30*, 16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010, PubMed: 21147999

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, *4*, 1043–1048. https://doi.org/10.1038/nn733, PubMed: 11559855

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences, U.S.A.*, *103*, 2458–2463. https://doi.org/10.1073/pnas.0509389103, PubMed: 16461904

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *360*, 815–836. https://doi.org/10.1098/rstb.2005.1622, PubMed: 15937014

Fritsche, M., Solomon, S. G., & de Lange, F. P. (2021). Brief stimuli cast a long-term trace in visual cortex. *bioRxiv*. https://doi.org/10.1101/2021.02.10.430579

Gheysen, F., Van Opstal, F., Roggeman, C., Van Waelvelde, H., & Fias, W. (2010). Hippocampal contribution to early and later stages of implicit motor sequence learning. *Experimental Brain Research*, *202*, 795–807. https://doi.org/10.1007/s00221-010-2186-6, PubMed: 20195849

Gheysen, F., Van Opstal, F., Roggeman, C., Van Waelvelde, H., & Fias, W. (2011). The neural basis of implicit perceptual sequence learning. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00137, PubMed: 22087090

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*, 14–23. https://doi.org/10.1016/j.tics.2005.11.006, PubMed: 16321563

Gronau, N., Neta, M., & Bar, M. (2008). Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, *20*, 371–388. https://doi.org/10.1162/jocn.2008.20027, PubMed: 18004950

Grotheer, M., & Kovács, G. (2015). The relationship between stimulus repetitions and fulfilled expectations. *Neuropsychologia*, *67*, 175–182. https://doi.org/10.1016/j.neuropsychologia.2014.12.017, PubMed: 25527870

Handel, S., & Buffardi, L. (1969). Using several modalities to perceive one temporal pattern. *Quarterly Journal of Experimental Psychology*, *21*, 256–266. https://doi.org/10.1080/14640746908400220, PubMed: 5347011

Haushofer, J., Livingstone, M. S., & Kanwisher, N. (2008). Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biology*, *6*, e187. https://doi.org/10.1371/journal.pbio.0060187, PubMed: 18666833

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2011). Neural encoding of relative position. *Journal of Experimental*

Psychology: Human Perception and Performance, *37*, 1032–1050. https://doi.org/10.1037/a0022338, PubMed: 21517211

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850–856. https://doi.org/10.3758/PBR.16.5.850, PubMed: 19815788

Hindy, N. C., Avery, E. W., & Turk-Browne, N. B. (2019). Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nature Communications*, *10*, 3989. https://doi.org/10.1038/s41467-019-12016-9, PubMed: 31488845

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*, 658–680. https://doi.org/10.1037/0096-3445.130.4.658, PubMed: 11757874

JASP Team. (2019). *JASP* (Version 0.10.2)[Computer software].

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage*, *169*, 334–341. https://doi.org/10.1016/j.neuroimage.2017.12.065, PubMed: 29277645

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, *23*, 672–685. https://doi.org/10.1016/j.tics.2019.04.013, PubMed: 31147151

Kaposvari, P., Kumar, S., & Vogels, R. (2018). Statistical learning signals in macaque inferior temporal cortex. *Cerebral Cortex*, *28*, 250–266. https://doi.org/10.1093/cercor/bhw374, PubMed: 27909007

Karuza, E. A., Emberson, L. L., Roser, M. E., Cole, D., Aslin, R. N., & Fiser, J. (2017). Neural signatures of spatial statistical learning: Characterizing the extraction of structure from complex visual scenes. *Journal of Cognitive Neuroscience*, *29*, 1963–1976. https://doi.org/10.1162/jocn_a_01182, PubMed: 28850297

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*, 46–54. https://doi.org/10.1016/j.bandl.2012.11.007, PubMed: 23312790

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*, 265–270. https://doi.org/10.1016/j.neuron.2012.04.034, PubMed: 22841311

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science*, *293*, 1506–1509. https://doi.org/10.1126/science.1061133, PubMed: 11520991

Kubovy, M. (1988). Should we resist the seductiveness of the space:time::vision:audition analogy? *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 318–320. https://doi.org/10.1037/0096-1523.14.2.318

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863, PubMed: 24324449

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A practical course*. Cambridge University Press.

Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology*, *69*, 1918–1929. https://doi.org/10.1152/jn.1993.69.6.1918, PubMed: 8350131

Mahar, D., Mackenzie, B., & McNicol, D. (1994). Modality-specific differences in the processing of spatially, temporally, and spatiotemporally distributed information. *Perception*, *23*, 1369–1386. https://doi.org/10.1068/p231369, PubMed: 7761246

Manahova, M. E., Mostert, P., Kok, P., Schoffelen, J.-M., & de Lange, F. P. (2018). Stimulus familiarity and expectation jointly modulate neural activity in the visual ventral stream. *Journal of Cognitive Neuroscience*, *30*, 1366–1377. https://doi.org/10.1162/jocn_a_01281, PubMed: 29762101

Marcovitch, S., & Lewkowicz, D. J. (2009). Sequence learning in infancy: The independent contributions of conditional probability and pair frequency information. *Developmental Science*, *12*, 1020–1025. https://doi.org/10.1111/j.1467-7687.2009.00838.x, PubMed: 19840056

Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 19401–19406. https://doi.org/10.1073/pnas.1112895108, PubMed: 22084090

Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00552, PubMed: 23970878

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition*, *17*, 398–422. https://doi.org/10.3758/BF03202613, PubMed: 2668697

Plante, E., Patterson, D., Gómez, R., Almryde, K. R., White, M. G., & Asbjørnsen, A. E. (2015). The nature of the language input affects brain activation during learning from a natural language. *Journal of Neurolinguistics*, *36*, 17–34. https://doi.org/10.1016/j.jneuroling.2015.04.005, PubMed: 26257471

Quek, G. L., & Peelen, M. V. (2020). Contextual and spatial associations between objects interactively modulate visual processing. *Cerebral Cortex*, *30*, 6391–6404. https://doi.org/10.1093/cercor/bhaa197, PubMed: 32754744

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87. https://doi.org/10.1038/4580, PubMed: 10195184

Repp, B. H., & Penel, A. (2002). Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1085–1099. https://doi.org/10.1037/0096-1523.28.5.1085, PubMed: 12421057

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *eLife*, *8*, e47869. https://doi.org/10.7554/eLife.47869, PubMed: 31442202

Richter, D., Ekman, M., & de Lange, F. P. (2018). Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *Journal of Neuroscience*, *38*, 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018, PubMed: 30030402

Rossion, B., Schiltz, C., & Crommelinck, M. (2003). The functionally defined right occipital and fusiform "face areas" discriminate novel from visually familiar faces. *Neuroimage*, *19*, 877–883. https://doi.org/10.1016/s1053-8119(03)00105-8, PubMed: 12880816

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*, 172–196. https://doi.org/10.1006/jmla.2001.2839

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience*, *26*, 1736–1747. https://doi.org/10.1162/jocn_a_00578, PubMed: 24456393

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, *8*, 522–535. https://doi.org/10.1038/nrn2155, PubMed: 17585305

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051, PubMed: 15501092

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*, 1004–1006. https://doi.org/10.1038/nn.2163, PubMed: 19160497

Tobia, M. J., Iacovella, V., Davis, B., & Hasson, U. (2012). Neural systems mediating recognition of changes in statistical regularities. *Neuroimage*, *63*, 1730–1742. https://doi.org/10.1016/j.neuroimage.2012.08.017, PubMed: 22906790

Tobia, M. J., Iacovella, V., & Hasson, U. (2012). Multiple sensitivity profiles to diversity and transition structure in non-stationary input. *Neuroimage*, *60*, 991–1005. https://doi.org/10.1016/j.neuroimage.2012.01.041, PubMed: 22285219

Todorovic, A., & de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *Journal of Neuroscience*, *32*, 13389–13395. https://doi.org/10.1523/JNEUROSCI.2227-12.2012, PubMed: 23015429

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766–786. https://doi.org/10.1037/0033-295X.113.4.766, PubMed: 17014302

Tremblay, P., Baroni, M., & Hasson, U. (2013). Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure. *Neuroimage*, *66*, 318–332. https://doi.org/10.1016/j.neuroimage.2012.10.055, PubMed: 23116815

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*, 1934–1945. https://doi.org/10.1162/jocn.2009.21131, PubMed: 18823241

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*, 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010, PubMed: 20720125

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 20754–20759. https://doi.org/10.1073/pnas.1117807108, PubMed: 22147913

Wang, L., Uhrig, L., Jarraya, B., & Dehaene, S. (2015). Representation of numerical and sequential patterns in Macaque and human brains. *Current Biology*, *25*, 1966–1974. https://doi.org/10.1016/j.cub.2015.06.035, PubMed: 26212883