# “The role of contextual semantic congruency on the facilitation of perceptual processing prior to recognition-level; an fMRI study”

# Abstract:

Recent research challenges the serial model of predictive coding, suggesting that prior knowledge can modulate and facilitate perception without the need for recognition and before recognition-level. However, these findings have been limited to simple stimuli and behavioral measures, leaving open questions about how predictive coding operates in complex, naturalistic contexts. Given evidence that predictive mechanisms differ across levels of stimulus complexity and behavioral tasks may be subject to interpretive bias, there is a need for neuroimaging studies using naturalistic stimuli. This fMRI study investigates whether semantic congruency between objects and scenes elicits expectation suppression during perception and facilitates the perception process. Using an intact/scrambled task, we will assess BOLD responses in V2, V3, and the fusiform gyrus. We hypothesize reduced activation for congruent stimuli, indicating minimized prediction error. Also, reduced neural response in the fusiform gyrus would be a demonstration of the inclusion of higher-level semantic priors in the perception shaping. Findings will support and extend revised predictive coding models by showing that high-level priors shape perception, even in complex settings.


# Introduction:

Visual perception is an active process influenced by prior experiences and predictive expectations. The predictive coding model, which is grounded in Bayesian inference, proposes that the brain hierarchically generates top-down predictions (“priors”) related to incoming sensory input. These priors are based on prior experiences and statistical regularities (probabilistic process related to the number of people encountering some stimuli more than others). These predictions are compared with the actual input, and the difference between expectations and actual value (“prediction error”) is used to update internal models for more accurate future perception (Friston, 2005; Rao & Ballard, 1999). By integrating top-down priors with bottom-up input, the brain constructs perception as a probabilistic inference process (Clark, 2013).

Up until now, this predictive model has been viewed as a serial model. Based on this perspective, predictions shape the recognition of the sensory input only after the completion of perceptual processing (Yang & Beck, 2023). Consequently, most studies have focused on recognition-based tasks, such as semantic categorization, object naming, or recognition memory (Kumar et al., 2021; Bonner & Epstein, 2021). Based on these studies, higher expectations result in facilitated recognition of visual stimuli.

Neuroimaging evidence has consistently supported these results with attenuated neural responses across the visual hierarchy in response to expected or semantically congruent stimuli, suggesting minimized prediction error. Functional MRI (fMRI) studies have demonstrated reduced BOLD activity for expected visual input not only in high-level areas such as the fusiform gyrus and lateral occipital complex (LOC) but also in early and mid-level visual regions, including V1, V2, and V3

(Richter et al., 2017); (Kok et al., 2012); (Summerfield et al., 2008). These reductions are called “expectation suppression”.

However, recent studies have begun to challenge this serial-model perspective. Yang and Beck (2023) demonstrated that prior knowledge can influence perception itself, even when explicit recognition is not utilized. By using an intact/scrambled discrimination task involving familiar and unfamiliar faces, they found that participants responded faster and more accurately to familiar compared to unfamiliar stimuli. Unlike previous methodologies, this task avoided semantic labeling or recognition, showing that the facilitation of perceptual processing happens without the utilization of recognition.

Following this study, Deng et al. (2024) by using the intact/scrambled task found that memorable images (high statistical regularities) elicited attenuated N300 and N400 ERP components, which are linked to template matching and semantic processing, respectively. These results align with the predictive coding model and provide neural-based evidence to support Yang and Beck’s new model. These findings suggest a dynamic interplay between top-down predictions and sensory input at the perceptual level.

Despite this progress, crucial aspects are still uncovered. These studies have focused on isolated stimuli (e.g., faces, shapes, or singular objects) without considering the contexts and semantic associations. Actually, in daily life, perception rarely involves isolated objects. In fact, contexts in which objects appear can significantly affect the way they are interpreted. This phenomenon is called “Object-Scene Semantic Congruency”. Based on this phenomenon, objects are recognized more easily in semantically congruent contexts (Nicholls et al, 2022). Nicholls et al (2022) found reduced N300/N400 for congruent object-scene recognition. For instance, a toaster is easier to recognize in a kitchen than in a gym. This facilitation is interpreted through the lens of predictive coding. Based on this model, a congruent object-scene would result in lower prediction error and reduced processing demands (El-Sourani et al, 2020).

Analyzing this new approach of predictive coding on the contextual level is vital for different reasons. First, analyzing predictive coding using Object-Scene stimuli is crucial for improving ecological validity. In real life, perception unfolds in semantically rich environments, not around isolated objects. To model real-life perception (not just lab-based), researchers must study predictive mechanisms in naturalistic contexts (Spaak et al., 2020). Furthermore, predictive processing may behave differently across different levels of complexity. It is shown that, in realistic settings, only the most informative and statistically reliable cues may drive predictions, which indicates a more selective and context-sensitive predictive system (Reuter et al., 2020). Therefore, to acquire reliable results related to the impact of predictive coding on the perceptual level, it is vital to analyze it in a more complex and naturalistic study.

Second, context-based study offers a more precise understanding of predictive coding by allowing a clearer distinction between semantic priming and prediction error minimization. Basic tasks that use familiar or memorable objects may reflect facilitated access to stored concepts (e.g., semantic priming) rather than actual minimization of perceptual surprise. So, Object-Scene congruency can provide a better setting for analyzing the influence of predictive coding theory on perceptual processing.

Based on previous studies and the nature of predictive coding, it can be predicted that a congruent context compared to their counterparts might cause facilitation in the process of perceiving visual stimuli. However, some findings and ideas can reduce our certainty regarding this prediction. Spaak et al (2020) introduced a phenomenon called "Congruency cost". They stated that congruent visual stimuli, due to their salience, would cause perception impairment, which was shown with increased reaction time for context-congruent stimuli. They offered that it might be a result of the attentional orienting. If this is the case, this pattern should not be seen in our study, since we would like to analyze this impact on the perceptual level. So, this study can be further used to see whether congruency cost happens on the perceptual level or not.

The other study is Reuter's study (2020), which demonstrated different levels of predictive coding at different complexity levels. So, while perceptual facilitation was seen for simple objects, that might not be the pattern that is seen in more complex and context-derived settings.

Altogether, this study goes beyond object familiarity to see whether this perceptual facilitation will be affected by semantic congruency (context) and whether this facilitation can be seen in more naturalistic and complex settings or not.

To investigate the outlined gaps, the present study will examine neural responses across the visual hierarchy using fMRI to provide information beyond the behavioral level. Mid-level visual regions (V2/V3) are important in this study since they mediate between early feature detection and high-level interpretation and have been involved in template matching and perceptual-level facilitation. These regions demonstrated sensitivity to predictive coding with suppressed BOLD amplitude for predictable stimuli (Schellekens et al., 2016). Yang and Beck (2023) suggested that these regions are critical for the facilitated perceptual processing. Supporting this notion, Deng et al. (2024) reported reduced N300 amplitudes, localized to mid-level regions, for high-statistical-regularity stimuli. These findings suggest V2 and V3 support early perceptual prediction based on prior experience.

At the higher level of the visual hierarchy, the fusiform gyrus is known for semantic integration and encoding object-context co-occurrence (Bonner & Epstein, 2021; Price et al., 2015). Also, it shows attenuated activity in response to stimuli with high statistical regularity (Pajani et al, 2017). Again, Deng et al (2024) found reduced N400 amplitude, which is related to semantical processing of visual input with high statistical regularity.

Overall, in this study, for the first time, these regions will be analyzed on the isolated perceptual level for more complex (object-scene semantic congruency) stimuli. Based on what has been mentioned, the hypothesis of this study would be as follows:

H1: Semantically congruent object-scene pairs will elicit reduced BOLD activation in V2, V3, and anterior fusiform gyrus compared to incongruent pairs (expectation suppression).

If this prediction is confirmed, reduced BOLD activation would suggest prediction error minimization. This would support the role of predictive coding at the perceptual level, even for more complex and naturalistic visual stimuli (Context-related) in both mid-level (V2/V3) and high-level (fusiform) visual areas, with the fusiform reflecting semantic-level integration. This would approve the new perspective regarding the predictive coding model in which the influence of priors and statistical regularities would be seen not after but during perception completion. Also, it would approve the study of Deng et al (2024), which showed that semantic aspects facilitate the perception of visual stimuli on the perceptual level. Null result in fusiform would suggest that semantic priors do not modulate perception before recognition.

Patterns related to predicted results can be seen in Figures 1 and 2.

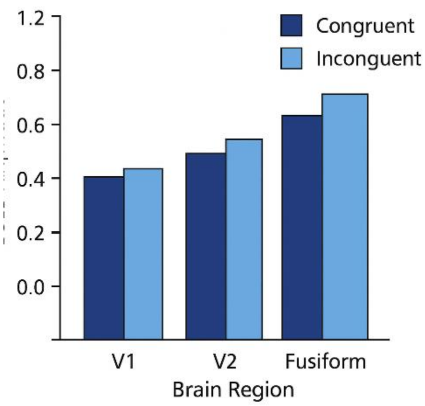


*Figure 1*

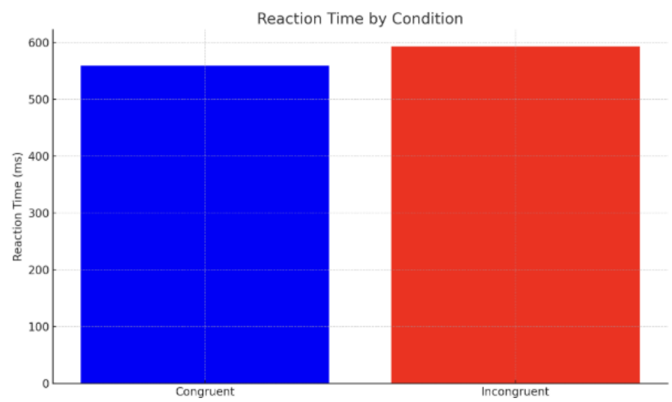


*Figure 2*

# Methods:

## Participants:

27 participants should be recruited for the fMRI study, based on a priori power analysis conducted using G*Power 3.1.9.7. Assuming a medium effect size (Cohen's d = 0.5), α = 0.05, and 80% power, this sample size was determined for a two-tailed paired-sample t-test. This estimate aligns with previous fMRI studies of expectation suppression and perceptual encoding (Geuter et al, 2018). However, due to expected exclusions during perceptual calibration (see section below) and general participant attrition, 40 participants should be recruited for the study.

Participants should be healthy, right-handed adults aged 18–35, with normal or corrected-to-normal vision and no history of neurological or psychiatric illness. Right-handedness will be confirmed using the Edinburgh Handedness Inventory (Oldfield, 1971). All participants will provide written informed consent, and all procedures will adhere to institutional ethical standards.

## Perceptual calibration:

To ensure consistent perceptual engagement across participants while maintaining feasibility in fMRI modeling, a group-level stimulus duration calibration should be done. Adapted from Yang and Beck (2023), thirty participants should complete an intact/scrambled discrimination task (Figure 3), with each stimulus briefly presented and followed by a backward mask created using grid scrambling. A QUEST adaptive staircase (Watson & Pelli, 1983) will be used to identify the individual duration yielding ~70.7% accuracy, which is an established threshold for perceptual access without full recognition (Kouider et al., 2007).

Threshold durations are gathered, and outliers are excluded using the 1.5×IQR rule. The mean of the remaining values was used as a fixed stimulus duration during fMRI scanning. This ensured uniform timing for GLM modeling, consistent with prior studies using group-averaged perceptual thresholds (Pessoa et al., 2002; Dehaene et al., 2001; Grill-Spector & Kanwisher, 2005).

To confirm that the fixed duration elicited comparable perceptual difficulty, participants completed a brief pre-scan validation task. Only those with discrimination accuracy between 60–75% were included, ensuring perceptual challenge without floor or ceiling effects. This threshold range is supported by prior research showing that 75% accuracy represents a standard Bayesian threshold for perceptual discrimination in 2AFC tasks (Kontsevich & Tyler, 1999), while 60% accuracy has been effectively used to maintain task difficulty during perceptual learning paradigms (Tsushima & Watanabe, 2009).

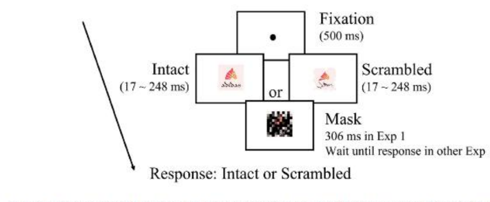


*Figure 3 / Yang and Beck (2023)*

## Stimuli:

The stimulus set will be constructed by combining real-world background scenes with foreground objects following the method introduced by Davenport and Potter (2004). Each object will be semantically either congruent or incongruent with its background scene. Backgrounds and objects will be sourced from publicly available high-resolution image databases. Unlike the original Davenport et al (2004) set, we will use multiple exemplars for each scene category (e.g., several jungle scenes). This prevents familiarity with objects or backgrounds (Figure 4).

Each composite image will be sized to 512 × 512 pixels and contrast-normalized across conditions to minimize low-level visual differences. Using diffeomorphic scrambling (Stojanoski & Cusack,

2014) and grid scrambling, scrambled formats and masks are generated, respectively (Figure 5). To assess and control for visual salience, the MATLAB-based Saliency Toolbox will be used (Walther & Koch, 2006). Any image showing extreme deviation from the mean salience will be excluded. Additionally, salience metrics will be included as parametric regressors in the fMRI analysis to account for potential low-level confounds in BOLD activity (Rice et al., 2013).

Finally, to validate congruency assignments, a separate pilot group (n = 15) will rate the semantic fit of each object-background combination on a 7-point Likert scale. The finalized set will include 320 trials (160 intact, 160 scrambled), with intact trials evenly divided between congruent and incongruent conditions.

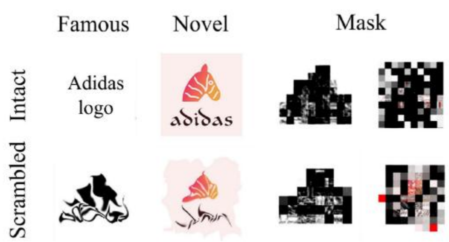


*Figure 5 / Yang et al (2023)*

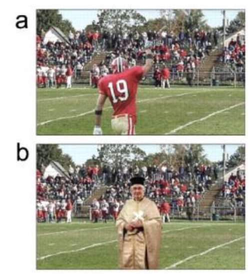


*Figure 4 / Davenport et al (2004)*

## Design:

This study uses a within-subjects, 2-condition design with an event-related structure. The independent variable is semantic congruency, with two categorical levels including congruent and incongruent object-scene pairs. This variable is manipulated across short in-scanner trials randomized within each run. Each trial consists of either a congruent or an incongruent intact image, intermixed with scrambled images. Only correct trials are included for further analysis to ensure consistent perceptual processing across conditions.

The dependent variable is BOLD activation in univariate analyses, particularly in V2, V3, and the fusiform gyrus, to assess expectation suppression.

## Procedure:

The in-scanner procedure is adapted from the intact/scrambled paradigm used by Deng et al. (2024) and the rapid event-related fMRI design implemented by Kok et al. (2012).

Participants will first complete a brief practice session (20 trials) to familiarize themselves with the task, using object-scene stimuli not included in the main experiment. No feedback will be provided during this session or the main experiment.

During scanning, participants will complete four event-related runs, each consisting of 80 trials: 40 scrambled, 20 semantically congruent, and 20 incongruent object-scene images. So, overall, this study takes around 30 minutes to complete all 4 runs. Trials will be pseudorandomized within each run. Each trial begins with a central fixation cross (500 ms), followed by a target image shown with the duration derived from the group mean (see above), and then a perceptual mask for 494 ms. This is followed by a 1500 ms response window, during which participants will indicate whether the image was intact or not using an MR-compatible button box. Intertrial intervals will be jittered between 2000 and 4000 ms (mean = 3000 ms) to optimize design efficiency and decorrelate event-related BOLD responses (Figure 6).  Participants will be unaware of the congruency manipulation and instructed only to perform the intact/scrambled detection task.

Stimuli presentation will be controlled using MATLAB R2021a with Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and displayed via a luminance-calibrated EIKI projector (1024 × 768 resolution, 60 Hz refresh rate) onto a rear-projection screen viewed through a mirror mounted on the head coil.

In addition to functional runs, a high-resolution T1-weighted anatomical scan will be acquired to allow for accurate co-registration, normalization, and ROI localization.

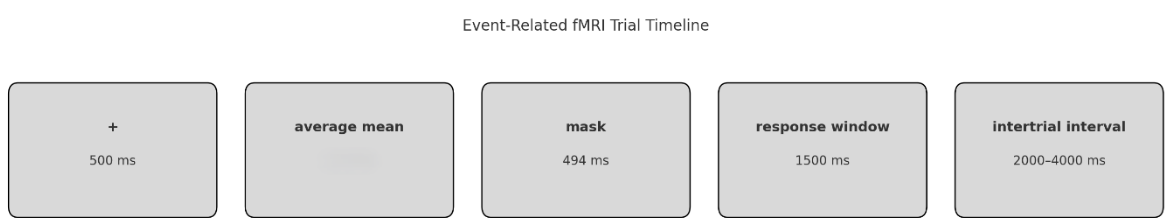


*Figure 6*

# Analyze:

**Preprocessing:**

All functional and anatomical MRI data will be preprocessed and denoised using SPM12 and the CONN toolbox implemented in MATLAB R20204a. Preprocessing will follow established pipelines for event-related designs (Kok et al., 2012) with the following steps applied in the specified order: Functional volumes will first undergo slice timing correction based on the MRI device's slice acquisition parameters. All functional images will then be realigned to the first image of each run and motion-corrected using six motion parameters; volumes exhibiting >3 mm displacement or >3° rotation will be flagged for exclusion. The mean functional image will be co-registered to each participant's anatomical T1-weighted scan. Anatomical images will be segmented into gray matter, white matter, and CSF via SPM's unified segmentation, and the resulting deformation fields applied to normalize functional images to MNI152 space, resampled to 2 × 2 × 2 mm voxels. To improve signal-to-noise ratio, normalized images will be smoothed using a 6 mm FWHM Gaussian kernel. Finally, confound regression and band-pass filtering (0.008–0.09 Hz) will be performed using the CONN toolbox.

**First-level analysis:**

First-level (subject-level) analysis will be conducted using the general linear model (GLM) in SPM12. Only sessions that demonstrated an accuracy rate within the threshold will be used for analysis (0.60-0.75). The design matrix will include three regressors of interest: congruent intact images, incongruent intact images, and scrambled images. These regressors will be modeled as impulse functions since we are not following a block-wise study, corresponding to stimulus onset, and convolved with the canonical hemodynamic response function (HRF). Only correct trials will be included in the model to ensure perceptual reliability across conditions.

Six head motion parameters obtained from realignment (three translations, three rotations) will be included as nuisance regressors. An additional regressor for each run will be included to model session-specific baseline differences. Also, visual salience scores may be added as a parametric modulator on intact conditions (Groen et al, 2013).

In our contrast design matrices, we would only use regressors of interest:

1) **Congruent > Incongruent:**

   This contrast tests the prediction suppression hypothesis by identifying regions with greater BOLD response to incongruent stimuli. The contrast weights will be:

   Congruent = -1, Incongruent = +1, Scrambled = 0.

The scrambled condition will not be directly analyzed in the main hypothesis tests but is modeled to capture shared low-level visual processing variance. The beta values of these contrasts will be calculated for each subject for the second-level analysis.

**Second-level analysis:**

At the group level, contrast images from each subject will be entered into a second-level and a one-sample t-test will be applied. Instead of mass univariate analysis, region-specific analysis would be done in this part, which enhances statistical power and control over multiple comparisons (prediction of type-1 error). This pattern is ideal for detecting subtle activation due to predictive coding. ROI masks will be defined using the Harvard-Oxford cortical atlas.

Statistical significance will be assessed using a family-wise error (FWE) correction at $p < 0.05$ within the ROI masks. In exploratory whole-brain analyses, multiple comparisons will be corrected using FWE at the cluster level ($p < 0.05$, cluster-forming threshold $p < 0.001$ uncorrected), as recommended in recent neuroimaging guidelines (Eklund et al., 2016).

## References:


Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24368-2

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897x00357

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/s0140525x12000477

Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758. https://doi.org/10.1038/89551

Deng, W., Beck, D. M., & Federmeier, K. D. (2024). Image memorability is linked to facilitated perceptual and semantic processing. *Imaging Neuroscience*, *2*, 1–13. https://doi.org/10.1162/imag_a_00281

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Geuter, S., Boll, S., Eippert, F., & Büchel, C. (2017). Functional dissociation of stimulus intensity encoding and predictive coding of pain in the insula. *ELife*, *6*. https://doi.org/10.7554/elife.24770

Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition. *Psychological Science*, *16*(2), 152–160. https://doi.org/10.1111/j.0956-7976.2005.00796.x

Kok, P., Jehee, Janneke F. M., & de Lange, Floris P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. https://doi.org/10.1016/s0042-6989(98)00285-5

Kouider, S., Gardelle, V. de, Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301–307. https://doi.org/10.1016/j.tics.2010.04.006

Kumar, N., van Vugt, F. T., & Ostry, D. J. (2021). Recognition memory for human motor learning. *Current Biology*. https://doi.org/10.1016/j.cub.2021.01.097

Nadiya El-Sourani, Trempler, I., Wurm, M. F., Fink, G. R., & Schubotz, R. I. (2020). Predictive Impact of Contextual Objects during Action Observation: Evidence from Functional Magnetic Resonance Imaging. *Journal of Cognitive Neuroscience*, *32*(2), 326–337. https://doi.org/10.1162/jocn_a_01480

Nicholls, V. I., Alsbury-Nealy, B., Krugliak, A., & Clarke, A. (2022). Context effects on object recognition in real-world environments: A study protocol. *Wellcome Open Research*, *7*, 165–165. https://doi.org/10.12688/wellcomeopenres.17856.2

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Pajani, A., Kouider, S., Roux, P., & de Gardelle, V. (2017). Unsuppressible Repetition Suppression and exemplar-specific Expectation Suppression in the Fusiform Face Area. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-00243-3

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://pubmed.ncbi.nlm.nih.gov/9176953/

Pessoa, L., Gutierrez, E., Bandettini, P. A., & Ungerleider, L. G. (2002). Neural Correlates of Visual Working Memory. *Neuron*, *35*(5), 975–987. https://doi.org/10.1016/s0896-6273(02)00817-6

Price, C. J., Noppeney, U., Phillips, J., & Devlin, J. T. (2003). HOW IS THE FUSIFORM GYRUS RELATED TO CATEGORY-SPECIFICITY? *Cognitive Neuropsychology*, *20*(3-6), 561–574. https://doi.org/10.1080/02643290244000284

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*(1), 79–87. https://doi.org/10.1038/4580

Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. *Journal of Neuroscience*, *34*(26), 8837–8844. https://doi.org/10.1523/jneurosci.5265-13.2014

Schellekens, W., van Wezel, R. J. A., Petridou, N., Ramsey, N. F., & Raemaekers, M. (2014). Predictive coding for motion stimuli in human early visual cortex. *Brain Structure and Function*, *221*(2), 879–890. https://doi.org/10.1007/s00429-014-0942-2

Spaak, E., Peelen, M. V., & de Lange, F. P. (2022). Scene Context Impairs Perception of Semantically Congruent Objects. *Psychological Science*, *33*(2), 299–313. https://doi.org/10.1177/09567976211032676

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12), 6–6. https://doi.org/10.1167/14.12.6

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M-Marsel., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004–1006. https://doi.org/10.1038/nn.2163

Tsushima, Y., & Watanabe, T. (2009). Roles of attention in perceptual learning from perspectives of psychophysics and animal learning. *Learning & Behavior*, *37*(2), 126–132. https://doi.org/10.3758/lb.37.2.126

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407. https://doi.org/10.1016/j.neunet.2006.10.001

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. https://doi.org/10.3758/bf03202828

Yang, P.-L., & Beck, D. M. (2023). Familiarity influences visual detection in a task that does not require explicit recognition. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-023-02703-7