Memory & Cognition
https://doi.org/10.3758/s13421-020-01050-4

When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention

Gwendolyn Rehrig[1] · Taylor R. Hayes[2] · John M. Henderson[1,2] · Fernanda Ferreira[1]

Abstract
The complexity of the visual world requires that we constrain visual attention and prioritize some regions of the scene for attention over others. The current study investigated whether verbal encoding processes influence how attention is allocated in scenes. Specifically, we asked whether the advantage of scene meaning over image salience in attentional guidance is modulated by verbal encoding, given that we often use language to process information. In two experiments, 60 subjects studied scenes (N_1 = 30 and N_2 = 60) for 12 s each in preparation for a scene-recognition task. Half of the time, subjects engaged in a secondary articulatory suppression task concurrent with scene viewing. Meaning and saliency maps were quantified for each of the experimental scenes. In both experiments, we found that meaning explained more of the variance in visual attention than image salience did, particularly when we controlled for the overlap between meaning and salience, with and without the suppression task. Based on these results, verbal encoding processes do not appear to modulate the relationship between scene meaning and visual attention. Our findings suggest that semantic information in the scene steers the attentional ship, consistent with cognitive guidance theory.

Keywords Scene processing · Visual attention · Meaning · Salience · Language

Introduction

Because the visual world is information-rich, observers prioritize certain scene regions for attention over others to process scenes efficiently. While bottom-up information from the stimulus is clearly relevant, visual attention does not operate in a vacuum, but rather functions in concert with other cognitive processes to solve the problem at hand. What influence, if any, do extra-visual cognitive processes exert on visual attention?

Two opposing theoretical accounts of visual attention are relevant to the current study: saliency-based theories and cognitive guidance theory. According to saliency-based theories (Itti & Koch, 2001; Wolfe & Horowitz, 2017), salient scene regions – those that contrast with their surroundings based on low-level image features (e.g., luminance, color, orientation) – pull visual attention across a scene, from the most salient location to the least salient location in descending order (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002). Saliency-based explanations cannot explain that physical salience does not determine which scene regions are fixated (Tatler, Baddeley, & Gilchrist, 2005) and that top-down task demands influence attention more than physical salience does (Einhäuser, Rutishauer, & Koch, 2008). Cognitive guidance theory can account for these findings: the cognitive system pushes visual attention to scene regions, incorporating stored knowledge about scenes to prioritize regions that are most relevant to the viewer's goals (Henderson, 2007). Under this framework, cognitive systems – for example, long- and short-term memory, executive planning, etc. – operate together to guide visual attention. Coordination of cognitive systems helps to explain behavioral findings where saliency-based attentional theories fall short. For example, viewers look preferentially at meaningful regions of a scene (e.g., those containing task-relevant objects), even when they are not visually salient (e.g., under shadow), despite the presence of a salient distractor (Henderson, Malcolm, & Schandl, 2009).

✉ Gwendolyn Rehrig
glrehrig@ucdavis.edu

[1] Department of Psychology, University of California, One Shields Ave., Davis, CA 95616-5270, USA

[2] Center for Mind and Brain, University of California, Davis, CA, USA

Published online: 19 May 2020

Recent work has investigated attentional guidance by representing the spatial distribution of image salience and scene meaning comparably (see Henderson, Hayes, Peacock, & Rehrig, 2019, for review). Henderson and Hayes (2017) introduced meaning maps to quantify the distribution of meaning over a scene. Raters on mTurk saw small scene patches presented at two different scales and judged how meaningful or recognizable each patch was. Meaning maps were constructed by averaging the ratings across patch scales and smoothing the values. Image salience was quantified using Graph-Based Visual Salience (GBVS; Harel et al., 2006). The feature maps were correlated with attention maps that were empirically derived from viewer fixations in scene memorization and esthetic judgement tasks. Meaning explained greater variance in attention maps than salience did, both for linear and semipartial correlations, suggesting that meaning plays a greater role in guiding visual attention than image salience does. This replicated when attention maps constructed from the same dataset were weighted on fixation duration (Henderson & Hayes 2018), when viewers described scenes aloud (Henderson, Hayes, Rehrig, & Ferreira, 2018; Ferreira & Rehrig, 2019), during free-viewing of scenes (Peacock, Hayes, and Henderson, 2019a, b), when meaning was not task-relevant (Hayes & Henderson, 2019a), and even when image salience was task-relevant (Peacock, Hayes, and Henderson, 2019a, b). In sum, scene meaning explained variation in attention maps better than image salience did across experiments and tasks, supporting the cognitive guidance theory of attentional guidance.

One question that remains unexplored is whether other cognitive processes indirectly influence cognitive guidance of attention. For example, it is possible that verbal encoding may modulate the relationship between scene meaning and visual attention: Perhaps the use of language, whether vocalized or not, pushes attention to more meaningful regions. While only two of the past experiments were explicitly linguistic in nature (scene description; Ferreira & Rehrig, 2019; Henderson et al., 2018), the remaining tasks did not control for verbal encoding processes.

There is evidence that observers incidentally name objects silently during object viewing (Meyer, Belke, Telling, & Humphreys 2007; Meyer & Damian, 2007). Meyer et al. (2007) asked subjects to report whether a target object was present or not in an array of objects, which sometimes included competitors that were semantically related to the target or were semantically unrelated, but had a homophonous name (e.g., *bat* the tool vs. *bat* the animal). The presence of competitors interfered with search, which suggests information about the objects (name, semantic information) became active during viewing, even though that information was not task-relevant. In a picture-picture interference study, Meyer and Damian (2007) presented target objects that were paired with distractor objects with phonologically similar names, and instructed subjects to name the target objects. Naming latency was shorter when distractor names were phonologically similar to the name of the target object, suggesting that activation of the distractor object's name occurred and facilitated retrieval of the target object's name. Together, the two studies demonstrate a tendency for viewers to incidentally name objects they have seen.

Cross-linguistic studies on the topic of linguistic relativity employ verbal interference paradigms to demonstrate that performance on perceptual tasks can be mediated by language processes. For example, linguistic color categories vary across languages even though the visual spectrum of colors is the same across language communities (Majid et al., 2018). Winawer et al. (2007) showed that observers discriminated between colors faster when the colors belonged to different linguistic color categories, but the advantage disappeared with verbal interference. These findings indicate that language processes can mediate performance on perceptual tasks that are ostensibly not linguistic in nature, and a secondary verbal task that prevents task-incidental language use can disrupt the mediating influence of language. Similar influences of language on ostensibly non-linguistic processes, and the disruption thereof by verbal interference tasks, have been found for spatial memory (Hermer-Vazquez, Spelke, & Katsnelson, 1999), event perception (Trueswell & Papafragou, 2010), categorization (Lupyan, 2009), and numerical representations (Frank, Fedorenko, Lai, Saxe, & Gibson, 2012), to name a few (see Lupyan, 2012; Perry & Lupyan, 2013; Ünal & Papafragou, 2016, for discussion).

The above literature suggests we use internal language during visual processing, and in some cases those language processes may mediate perceptual processes. Could the relationship between meaning and visual attention observed previously (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019a,b) have been modulated by verbal encoding processes? To examine this possibility, we used an articulatory suppression manipulation to determine whether verbal encoding mediates attentional guidance in scenes.

In the current study, observers studied 30 scenes for 12 s each for a later recognition memory test. The scenes used in the study phase were mapped for meaning and salience. We conducted two experiments in which subjects performed a secondary articulatory suppression task half of the time in addition to memorizing scenes. In Experiment 1, the suppression manipulation was between subjects, and the articulatory suppression task was to repeat a three-digit sequence aloud during the scene viewing period. We chose this suppression task because we suspected subjects might adapt to and subvert simpler verbal interference such as a syllable repetition (e.g., Martin, Branzi, and Bar, 2018), and because digit sequence repetition imposes less cognitive load than n-back tasks (Allen, Baddeley, & Hitch, 2017). In Experiment 2, we implemented a within-subject design using two experimental blocks: one with the sole task of memorizing scenes, the other with an additional articulatory suppression task. Because numerical stimuli may be processed differently than

other verbal stimuli (Maloney et al., 2019; van Dijck & Fias, 2011), we instead asked subjects to repeat the names of a sequence of three shapes aloud during the suppression condition. In the recognition phase of both experiments, subjects viewed 60 scenes – 30 that were present in the study phase, 30 foils – and indicated whether or not they recognized the scene from the study phase.

We tested two competing hypotheses about the relationship between verbal encoding and attentional guidance in scenes. If verbal encoding indeed mediated the relationship between meaning and attentional guidance in our previous work, we would expect observers to direct attention to meaningful scene regions only when internal verbalization strategies are available to them. Specifically, meaning should explain greater variance in attention maps than saliency in the control condition, and meaning should explain less or equal variance in attention as salience when subjects suppressed internal language use. Conversely, if verbal encoding did not mediate attentional guidance in scenes, the availability of verbalization strategies should not affect attention, and so we would expect to find an advantage of meaning over salience whether or not subjects engaged in a suppression task.

Experiment 1: Methods

Subjects Sixty-eight undergraduates enrolled at the University of California, Davis participated for course credit. All subjects were native speakers of English, at least 18 years old, and had normal or corrected-to-normal vision. They were naïve to the purpose of the experiment and provided informed consent as approved by the University of California, Davis Institutional Review Board. Six subjects were excluded from analysis because their eyes could not be accurately tracked, one due to an equipment failure, and one due to experimenter error; data from the remaining 60 subjects were analyzed (30 subjects/condition).

Stimuli Scenes were 30 digitized (1,024 x 768) and luminance-matched photographs of real-world scenes used in a previous experiment (Henderson et al., 2018). Of these, ten depicted outdoor environments (five street views), and 20 depicted indoor environments (three kitchens, five living rooms, two desk areas, and ten different room types). People were not present in any scenes.

Another set of 30 digitized images of comparable scenes (similar scene categories and time period, no people depicted) were selected from a Google image search and served as memory foils. Because we did not evaluate attentional guidance for the foils, meaning and salience were not quantified for these scenes, and the images were not luminance-matched.

Digit sequences Digit sequences were selected randomly without replacement from all three-digit numbers ranging from 100 to 999 (900 numbers total), then segmented into 30 groups of 30 sequences each such that each digit sequence in the articulatory suppression condition was unique.

Apparatus Eye movements were recorded with an SR Research EyeLink 1000+ tower mount eye-tracker (spatial resolution 0.01) at a 1,000 Hz sampling rate. Subjects sat 83 cm away from a 24.5-in. monitor such that scenes subtended approximately 26° x 19° visual angle at a resolution of 1,024 x 768 pixels, presented in 4:3 aspect ratio. Head movements were minimized using a chin and forehead rest integrated with the eye-tracker's tower mount. Subjects were instructed to lean against the forehead rest to reduce head movement while allowing them to speak during the suppression task. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled using SR Research Experiment Builder software. Data were collected on two systems that were identical except that one subject computer operated using Windows 10, and the other used Windows 7.

Scene memorization procedure Subjects were told they would see a series of scenes to study for a later memory test. Subjects in the articulatory suppression condition were told each trial would begin with a sequence of three digits, and were instructed to repeat the sequence of digits aloud during the scene-viewing period. After the instructions, a calibration procedure was conducted to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error below 0.99°.

Following successful calibration, there were three practice trials to familiarize subjects with the task prior to the experimental trials. In the suppression condition, during these practice trials participants studied three-digit sequences prior to viewing the scene. Practice digit sequences were three randomly sampled sequences from the range 1–99, in three-digit format (e.g., "0 3 6" for 36). Subjects pressed any button on a button box to advance throughout the task.

Each subject received a unique pseudo-random trial order that prevented two scenes of the same type (e.g., kitchen) from occurring consecutively. A trial proceeded as follows. First, a five-point fixation array was displayed to check calibration (Fig. 1a). The subject fixated the center cross and the experimenter pressed a key to begin the trial if the fixation was stable, or reran the calibration procedure if not. Before the scene, subjects in the articulatory suppression condition saw the instruction "Study the sequence of digits shown below. Your task is to repeat these digits over and over out loud for 12 seconds while viewing an image of the scene" along with a sequence of three digits separated by spaces (e.g., "8 0 9"), and pressed a button to proceed (Fig. 1b). The scene was shown for 12 s, during which time eye-movements were recorded (Fig. 1c). After 12 s elapsed, subjects pressed a button to proceed to the next trial (Fig. 1d). The trial procedure repeated until all 30 trials were complete.

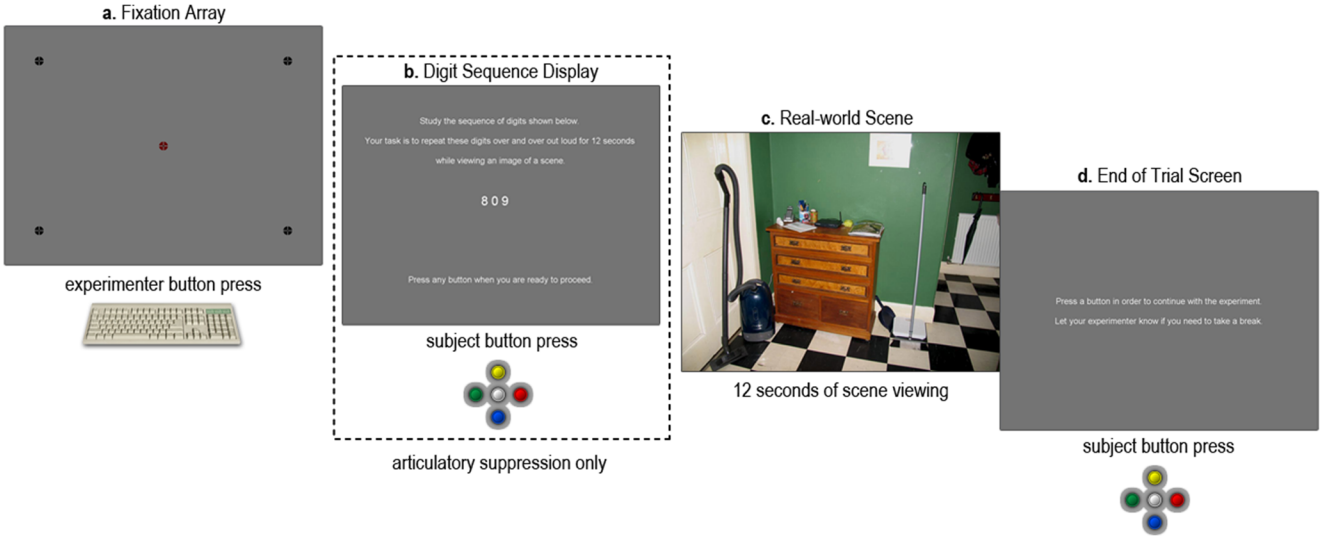

Fig. 1 Scene memorization trial procedure. (**a**) A five-point fixation array was used to assess calibration quality. (**b**) In the articulatory suppression condition only, the digit repetition task instructions were reiterated to subjects along with a three-digit sequence. (**c**) A real-world scene was shown for 12 s. (**d**) Subjects were instructed to press a button to initiate the next trial, at which point the trial procedure repeated (from **a**)

Memory test procedure A recognition memory test followed the experimental trials, in which subjects were shown the 30 experimental scenes and 30 foil scenes they had not seen previously. Presentation order was randomized without replacement. Subjects were informed that they would see one scene at a time and instructed to use the button box to indicate as quickly and accurately as possible whether they had seen the scene earlier in the experiment. After the instruction screen, subjects pressed any button to begin the memory test. In a recognition trial, subjects saw a scene that was either a scene from the study phase or a foil image. The scene persisted until a "Yes" or "No" button press occurred, after which the next trial began. Response time and accuracy were recorded. This procedure repeated 60 times, after which the experiment terminated.

Fixations and saccades were parsed with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2; SR Research, 2017). Eye-movement data were imported offline into Matlab using the Visual EDF2ASC tool packaged with SR Research DataViewer software. The first fixation was excluded from analysis, as were saccade amplitude (> 20°) and fixation duration outliers (< 50 ms, > 1,500 ms).

Attention maps Attention maps were generated by constructing a matrix of fixation counts with the same x,y dimensions as the scene, and counting the total fixations corresponding to each coordinate in the image. The fixation count matrix was smoothed with a Gaussian low-pass filter with circular boundary conditions and a frequency cutoff of -6 dB. For the scene-level analysis, all fixations recorded during the viewing period were counted. For the fixation analysis, separate attention maps were constructed for each ordinal fixation.

Meaning maps We generated meaning maps using the context-free rating method introduced in Henderson and Hayes (2017). Each 1,024 x 768 pixel scene was decomposed into a series of partially overlapping circular patches at fine and coarse spatial scales (Fig. 2b and c). The decomposition resulted in 12,000 unique fine-scale patches (87-pixel diameter) and 4,320 unique coarse-scale patches (205-pixel diameter), totaling 16,320 patches.

Raters were 165 subjects recruited from Amazon Mechanical Turk. All subjects were located in the USA, had a HIT approval rating of 99% or more, and participated once. Subjects provided informed consent and were paid $0.50.

All but one subject rated 300 random patches extracted from the 30 scenes. Subjects were instructed to rate how informative or recognizable each patch was using a 6-point Likert scale ("very low", "low", "somewhat low", "somewhat high", "high", "very high"). Prior to rating patches, subjects were given two examples each of low-meaning and high-meaning patches in the instructions to ensure they understood the task. Patches were presented in random order. Each patch was rated three times by three independent raters totaling 48,960 ratings per scene. Because there was high overlap across patches, each fine patch contained data from 27 independent raters, and each coarse patch from 63 independent raters (see Fig. 2d for patch examples).

Meaning maps were generated from the ratings for each scene by averaging, smoothing, and combining the fine and coarse scale maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The fine and coarse maps were then averaged [(fine map + coarse map)/2]. Because subjects in the eye-

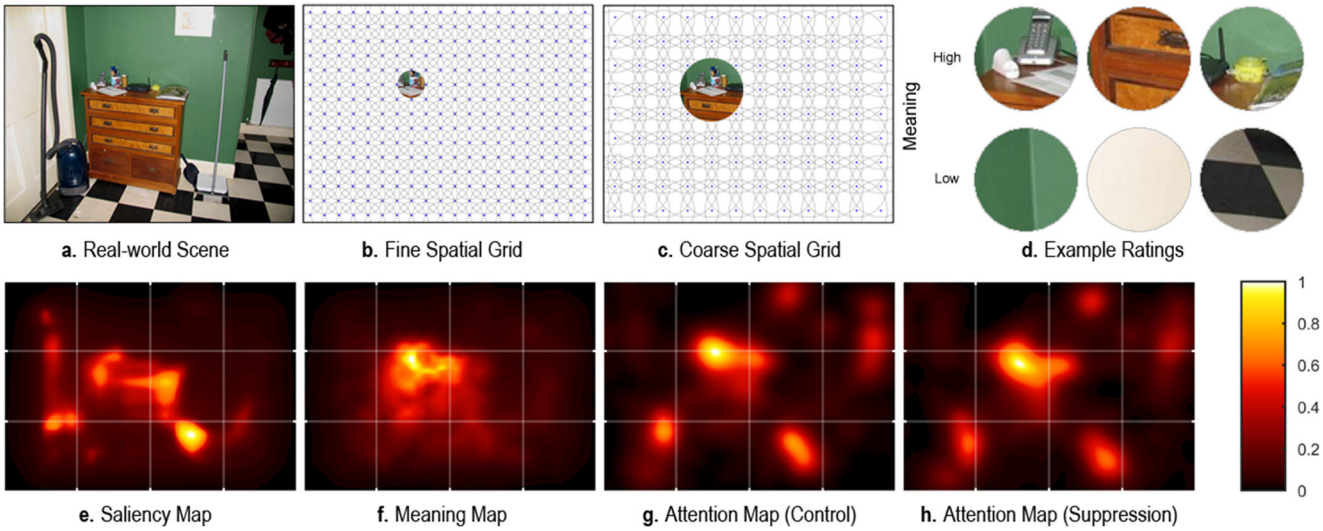

Fig. 2 (**a-d**). Meaning map generation schematic. (**a**) Real-world scene. (**b, c**) Fine scale (**b**) and coarse scale (**c**) spatial grids used to deconstruct the scene into patches. (**d**) Examples of scene patches that were rated as either low or high in meaning. (**e**–**h**) Examples of saliency (**e**), meaning (**f**), and attention (**g, h**) maps for the real-world scene shown in (**a**). Attention maps were empirically derived from viewer fixations in the control condition (**g**) and the articulatory suppression condition (**h**). For the purpose of visualization, all maps were normalized to the same attention map (**g**)

tracking task showed a consistent center bias[1] in their fixations, we applied center bias to the maps using a multiplicative down-weighting of scores in the map periphery (Hayes & Henderson, 2019b). The final map was blurred using a Gaussian filter via the Matlab function "imgaussfilt" with a sigma of 10 (see Fig. 2f for an example meaning map).

Saliency maps Image-based saliency maps were constructed using the Graph-Based Visual Saliency (GBVS) toolbox in Matlab with default parameters (Harel et al., 2006). We used GBVS because it is a state-of-the-art model that uses only image-computable salience. While there are newer saliency models that predict attention better (e.g., DeepGaze II: Kümmerer, Wallis, & Bethge, 2016; ICF: Kümmerer, Wallis, Gatys, & Bethge, 2017), these models incorporate high-level image features through training on viewer fixations (DeepGaze II and ICF) and object features (DeepGaze II), which may index semantic information. We used GBVS to avoid incorporating semantic information in image-based saliency maps, which could confound the comparison with meaning (see Henderson et al., 2019, for discussion).

Map normalization Prior to analysis, feature maps were normalized to a common scale using image histogram matching via the Matlab function "imhistmatch" in the Image Processing Toolbox. The corresponding attention map for each scene served as the reference image (see Henderson & Hayes, 2017). Map normalization was carried out within task conditions: for the map-based analysis of the control condition, feature maps were normalized to the attention map derived from fixations in the control condition only, and likewise for the suppression condition. Results did not differ between the current analysis and a second analysis using feature maps normalized to the same attention map (from fixations in the control condition).

We computed correlations (R^2) across the maps of 30 scenes to determine the degree to which saliency and meaning overlap with one another. We excluded the peripheral 33% of the feature maps when determining overlap between the maps to control for the peripheral downweighting applied to both, which otherwise would inflate the correlation between them. On average, meaning and saliency were correlated (R^2 = 0.48), and this relationship differed from zero (meaning and saliency: $t(29) = 17.24$, $p < 0.001$, 95% CI = [.43 .54]).

Experiment 1: Results

To determine what role verbal encoding might play in extracting meaning from scenes, we asked whether the advantage of meaning over salience in explaining variance in attention would hold in each condition. To answer this question, we conducted two-tailed paired t-tests within task conditions.

Sensitivity analysis To determine whether we obtained adequate effect sizes for the primary comparison of interest, we conducted a sensitivity analysis using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009). We computed the effect size index d_z – a standardized difference score (Cohen, 1988) – and the critical t statistic for a two-tailed paired t-test with 95% power

[1] "Center bias" is the tendency for fixations to cluster around the center of the scene and to be relatively absent in the periphery of the image (Tatler, 2007).

and a sample size of 30 scenes. The analysis revealed a critical t value of 2.05 and a minimum d_z of 0.68.

Attention: Scene-level analysis We correlated meaning and saliency maps with attention maps to determine the degree to which meaning or salience guided visual attention (Fig. 3). Squared linear and semipartial correlations (R^2) were computed within each condition for each of the 30 scenes. The relationship between meaning and salience, respectively, and visual attention was analyzed using t-tests. Cohen's d was computed to estimate effect size, interpreted as small (d = 0.2 - 0.49), medium (d = 0.5 - 0.79), or large (d = 0.8+) following Cohen (1988).

Linear correlations. In the control condition, when subjects were only instructed to memorize scenes, meaning accounted for 34% of the average variance in attention (M = 0.34, SD = 0.14) and salience accounted for 21% (M = 0.21, SD = 0.13). The advantage of meaning over salience was significant ($t(29)$ = 6.07, p < .001, 95% CI = [0.09 0.17], d = 0.97, d 95% CI = [0.58 1.36], d_z = 1.10). In the articulatory suppression condition, when subjects additionally had to repeat a sequence of digits aloud, meaning accounted for 37% of the average variance in attention (M = 0.37, SD = 0.17) whereas salience accounted for 23% (M = 0.23, SD = 0.12). The advantage of meaning over salience was also significant when the task prevented verbal encoding ($t(29)$ = 6.04, p < .001, 95% CI = [0.09 0.19], d = 0.88, d 95% CI = [0.53 1.22], d_z = 1.12).

Semipartial correlations. Because meaning and salience are correlated, we partialled out the shared variance explained by both meaning and salience. In the control condition, when the shared variance explained by salience was accounted for, meaning explained 15% of the average variance in attention (M = 0.15, SD = 0.10), while salience explained only 2% of the average variance once the variance explained by meaning was accounted for (M = 0.02, SD = 0.02). The advantage of meaning over salience was significant ($t(29)$ = 6.07, p < .001, 95% CI = [0.09 0.18], d = 1.98, d 95% CI = [0.86 3.10], d_z = 1.15). In the articulatory suppression condition, meaning explained 16% of the average unique variance after shared variance was partialled out (M = 0.16, SD = 0.11), while salience explained only 2% of the average variance after shared variance with meaning was accounted for (M = 0.02, SD = 0.03), and the advantage was significant ($t(29)$ = 6.05, p < .001, 95% CI = [0.09 0.19], d = 1.95, d 95% CI = [0.85 3.04], d_z = 1.09).

To summarize, we found a large advantage of meaning over salience in explaining variance in attention in both conditions, for both linear and semipartial correlations. For all comparisons, the value of the t statistic and d_z exceeded the thresholds obtained in the sensitivity analysis.

Attention: Fixation analysis Following our previous work (Henderson & Hayes, 2017; Henderson et al., 2018), we examined early fixations to determine whether salience influences early scene viewing (Parkhurst et al., 2002; but see Tatler et al., 2005). We correlated each feature map (meaning, salience) with attention maps at each fixation (Fig. 3b). Squared linear and semipartial correlations (R^2) were computed for each fixation, and the relationship between meaning and salience with attention, respectively, was assessed for the first three fixations using paired t-tests.

Linear correlations. In the control condition, meaning accounted for 37% of the average variance in attention during the first fixation, and 14% and 13% during the second and third fixations, respectively (1: M = 0.37, SD = 0.19; 2: M = .14, SD = .11; 3: M = .13, SD = .10). Salience accounted for 9% (1: M = .09, SD = .11), 8% (2: M = 0.08, SD = 0.09), and 7% of the average variance (3: M = 0.07, SD = 0.09) during the first, second, and third fixations, respectively. The advantage of meaning was significant for all three fixations (1: $t(29)$ = 8.59, p < .001, 95% CI = [0.21 0.34], d = 1.70, d 95% CI = [1.08 2.31]; 2: $t(29)$ = 3.40, p = .002, 95% CI = [0.03 0.11], d = 0.66, d 95% CI = [0.23 1.08]; 3: $t(29)$ = 4.21, p < .001, 95% CI = [0.03 0.08], d = 0.60, d 95% CI = [0.29 0.90]). For subjects in the suppression condition, meaning accounted for 42% of the average variance during the first fixation (M = 0.42, SD = 0.18), 21% during the second (M = 0.21, SD = 0.15), and 17% during the third fixation (M = 0.17, SD = 0.13). Salience accounted for 10% of the average variance during the first fixation (M = 0.10, SD = 0.10) and 9% during the second and third fixations (2: M = 0.09, SD = 0.09; 3: M = 0.09, SD = 0.09). The advantage of meaning over salience was significant for all three fixations (1: $t(29)$ = 10.27, p < .001, 95% CI = [0.26 0.38], d = 2.12, d 95% CI = [1.39 2.92]; 2: $t(29)$ = 5.49, p < .001, 95% CI = [0.08 0.17], d = 0.90, d 95% CI = [0.51 1.29]; 3: $t(29)$ = 4.49, p < .001, 95% CI = [0.04 0.12], d = 0.71, d 95% CI = [0.35 1.06]).

Semipartial correlations. To account for the correlation between meaning and salience, we partialled out shared variance explained by both meaning and salience, then repeated the fixation analysis on the semipartial correlations. In the control condition, after the shared variance explained by both meaning and salience was partialled out, meaning accounted for 30% of the average variance at the first fixation (M = 0.30, SD = 0.16), 10% of the variance during the second fixation (M = 0.10, SD = 0.09), and 8% during the third fixation (M = 0.08, SD =

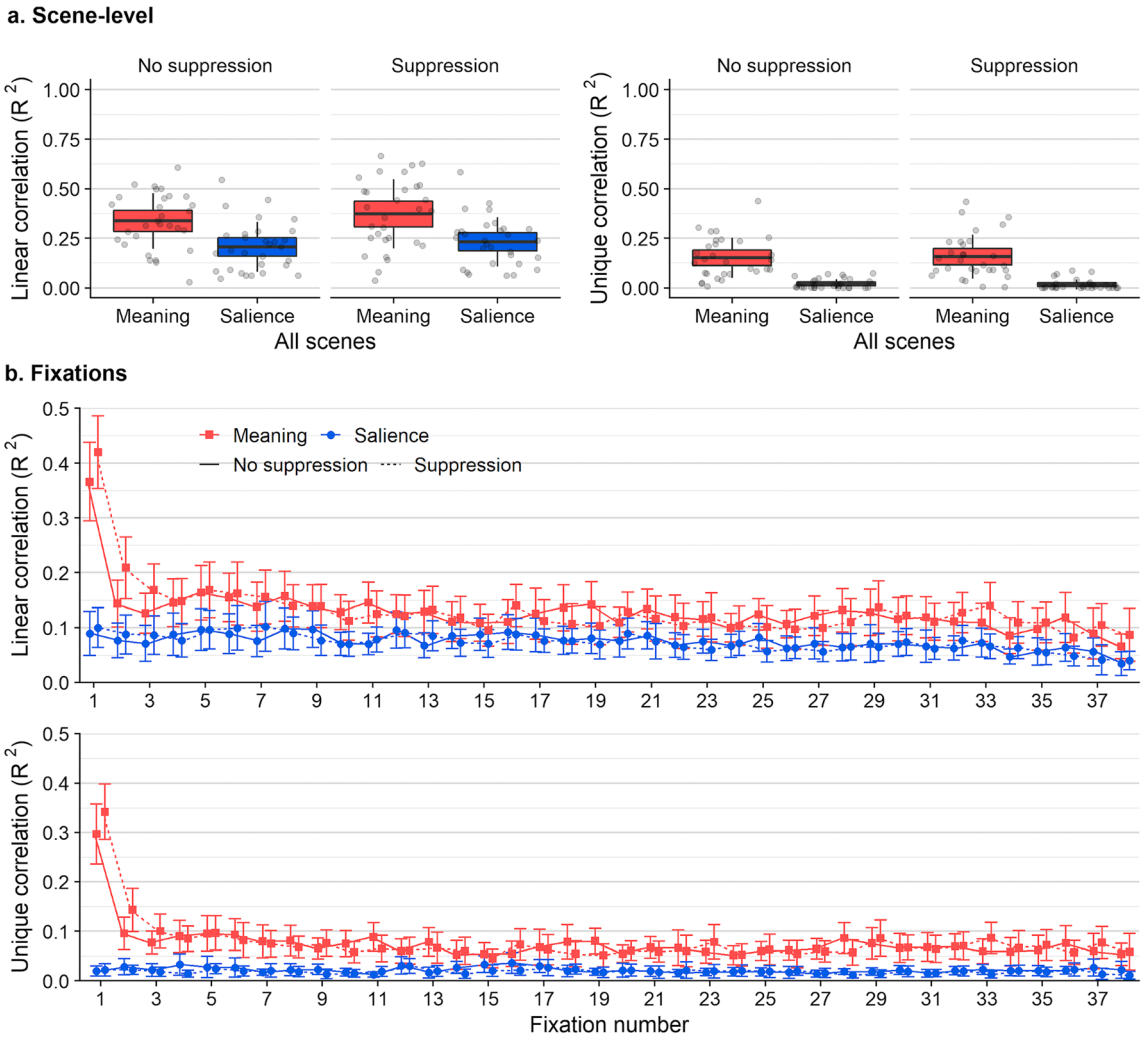

Fig. 3 (**a**) Box plots showing linear correlations (left) and semipartial correlations (right) between feature maps (meaning, saliency) and attention maps. The scatter box plots show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning (red box) and salience (blue box) across 30 scenes. (**b**) Line graphs showing linear correlations (top) and semipartial correlations (bottom) between feature maps and attention maps for each fixation (1–38) when subjects engaged in a memorization task only (solid lines) or additionally an articulatory suppression task (dashed lines). Error bars indicate 95% confidence intervals

0.06). After shared variance with meaning was partialled out, salience accounted for only 2% of the average unique variance at the first and third fixations (1: $M = 0.02$, $SD = 0.03$; 3: $M = 0.02$, $SD = 0.03$) and 3% at the second fixation ($M = 0.03$, $SD = 0.04$). The advantage of meaning was significant for all three fixations (1: $t(29) = 8.58$, $p < .001$, 95% CI = [0.21 0.34], $d = 2.66$, d 95% CI = [1.34 3.97]; 2: $t(29) = 3.40$, $p < .001$, 95% CI = [0.03 0.11], $d = 0.99$, d 95% CI = [0.28 1.70]; 3: $t(29) = 4.21$, $p < .001$, 95% CI = [0.03 0.08], $d = 1.10$, d 95% CI = [0.44 1.76]). In the articulatory suppression condition, after the shared variance with salience was partialled out, meaning accounted for 34% of the average variance during the first fixation ($M = 0.34$, $SD = 0.15$), 14% at the second fixation ($M = 0.14$, $SD = 0.12$), and 10% during the third fixation ($M = 0.10$, $SD = 0.09$). After the shared variance with meaning was partialled out, on average salience accounted for 2% of the variance at all three fixations (1: $M = 0.02$, $SD = 0.03$; 2: $M = 0.02$, $SD = 0.02$; 3: $M = 0.02$, $SD = 0.03$). The advantage of meaning was significant for all three fixations (1: $t(29) = 10.27$, $p < .001$, 95% CI = [0.26 0.38], $d = 3.25$, d 95% CI = [1.67 4.85]; 2: $t(29) = 5.49$, $p < .001$, 95% CI = [0.08 0.17], $d = 1.46$, d 95% CI = [0.69 2.22]; 3: $t(29) = 4.49$, $p < .001$, 95% CI = [0.04 0.12], $d = 1.25$, d 95% CI = [0.51 1.99]).

In sum, early fixations revealed a consistent advantage of meaning over salience, counter to the claim that salience influences attention during early scene viewing (Parkhurst et al., 2002). The advantage was present for the first three fixations in

both conditions, when we analyzed both linear and semipartial correlations, and all effect sizes were medium or large.

Memory: Recognition To confirm that subjects took the memorization task seriously, we totaled the number of hits, correct rejections, misses, and false alarms on the recognition task for each subject, each of which ranged from 0 to 30 (Fig. 4a). Recognition performance was high in both conditions. On average, subjects in the control condition correctly recognized scenes shown in the memorization task 95% of the time (M_{hits} = 0.95, SD_{hits} = 0.06), while subjects who engaged in the suppression task during memorization correctly recognized scenes 90% of the time (M_{hits} = 0.90, SD_{hits} = 0.09). Subjects in the control conditions falsely reported that a foil scene had been present in the memorization scene set 3% of the time on average ($M_{false\ alarms}$ = 0.03, $SD_{false\ alarms}$ = 0.03), and those in the suppression condition false alarmed an average of 4% of the time ($M_{false\ alarms}$ = 0.04, $SD_{false\ alarms}$ = 0.07). Overall, subjects in the control condition had higher recognition accuracy, though the difference in performance was small.

We then computed d' with log-linear correction to handle extreme values (ceiling or floor performance) using the dprime function from the psycho package in R, resulting in 30 data points per condition (1 data point/subject; Fig. 4b). On average, d' scores were higher in the control condition (M = 3.30, SD = 0.55) than the articulatory suppression condition (M = 2.99, SD = 0.74). The difference in performance was not significant, and the effect size was small ($t(58)$ = 1.83, p = 0.07, 95% CI = [-0.03 0.64], d = 0.47, d 95% CI = [-0.05 1.00]).

In sum, recognition was numerically better for subjects who were only instructed to study the scenes as opposed to those who additionally completed an articulatory suppression task, but the difference was not significant.

Experiment 1: Discussion

The results of Experiment 1 suggest that incidental verbalization does not modulate the relationship between scene meaning and visual attention during scene viewing. However, the experiment had several limitations. First, we implemented the suppression manipulation between-subjects rather than within-subjects out of concern that subjects might infer the hypothesis in a within-subject paradigm and skew the results. Second, because numerical cognition is unique (Maloney et al., 2019; van Dijck & Fias, 2011), it is possible that another type of verbal interference would affect the relationship between meaning and attention. Third, we tested relatively few scenes (N=30).

We conducted a second experiment to address these limitations and replicate the advantage of meaning over salience despite verbal interference. In Experiment 2, the verbal interference consisted of sequences of common shape names (e.g., square, heart, circle) rather than digits, and the interference paradigm was implemented within-subject using a blocked design. We added 30 scenes to the Experiment 1 stimulus set, yielding 60 experimental items total.

We tested the same two competing hypotheses in Experiments 1 and 2: If verbal encoding mediates the relationship between meaning and attentional guidance, and the use of numerical interference in Experiment 1 was insufficient to disrupt that mediation, then the relationship between meaning and attention should be weaker when incidental verbalization is not available, in which case meaning and salience may explain comparable variance in attention. If verbal encoding does not mediate attentional guidance in scenes and our Experiment 1 results cannot be explained by numerical interference specifically, then we expect meaning to explain

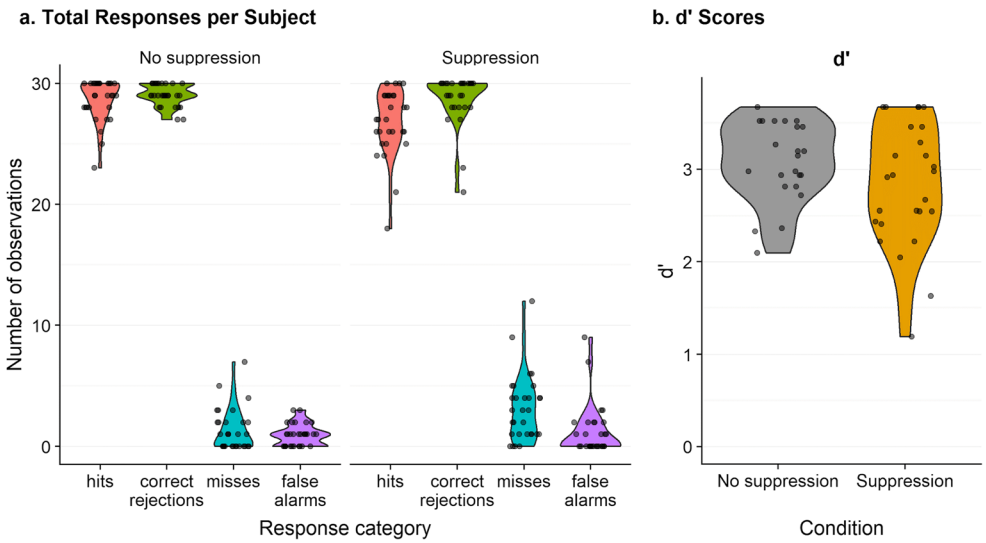

Fig. 4 (**a**) Violin plot showing the total number of recognition task responses for each subject (individual points), broken into hits, correct rejections, misses, and false alarms. (**b**) Violin plot showing d' values for each subject

greater variance in attention both when shape names are used as interference and when there is no verbal interference.

Experiment 2: Methods

The method for Experiment 2 was the same as Experiment 1, with the following exceptions.

Subjects Sixty-five undergraduates enrolled at the University of California, Davis participated for course credit. All were native speakers of English, at least 18 years old, and had normal or corrected-to-normal vision. They were naive to the purpose of the experiment and provided informed consent as approved by the University of California, Davis Institutional Review Board. Four subjects were excluded from analysis because their eyes could not be accurately tracked, and an additional subject was excluded due to excessive movement; data from the remaining 60 subjects were analyzed.

Shapes and shape sequences We selected the following common shapes for the suppression task: circle, cloud, club, cross, arrow, heart, moon, spade, square, and star. Names for the shapes were either monosyllabic (N=8) or disyllabic (N=2). Shape sequences consisted of three shapes randomly sampled without replacement from the set of 10.

Stimuli Scenes were 60 digitized (1,024 x 768) and luminance-matched photographs of real-world scenes. Thirty were used in Experiment 1, and an additional 30 were drawn from another study. Of the additional scenes, 16 depicted outdoor environments, and 14 depicted indoor environments, and each of the 30 scenes belonged to a unique scene category. People and text were not present in any of the scenes

Another set of 60 digitized images of comparable scenes (similar scene categories from the same time period, no people depicted) served as foils in the memory test. Thirty of these were used in Experiment 1, and an additional 30 were distractor images drawn from a previous study. The Experiment 1 scenes and the additional 30 scenes were equally distributed across blocks.

Apparatus The apparatus was identical to that used in Experiment 1.

Scene memorization procedure Subjects were informed that they would complete two separate experimental blocks, and that in one block each trial would begin with a sequence of three shapes that they would repeat aloud during the scene viewing period.

Following successful calibration, there were four practice trials to familiarize subjects with the task prior to the experimental trials. The first two practice trials were control trials, and the rest were articulatory suppression trials. These consisted of shape sequences (e.g., cloud arrow cloud) that were not repeated in the experimental trials. Before the practice trials, subjects were shown all of the shapes used in the suppression task, alongside the names of each shape (Fig. 5a). Subjects pressed any button on a button box to advance throughout the task.

The trial procedure was identical to Experiment 1, except that the pre-scene articulatory suppression condition displayed the instruction "Study the sequence of shapes shown below. Your task is to repeat these shapes over and over out loud for 12 seconds while viewing an image of the scene", followed by a sequence of three shapes (e.g., square, heart, cross) until the subject pressed a button (Fig. 5b).

Memory test procedure Following the experimental trials in each block, subjects performed a recognition memory in which 30 experimental scenes they saw earlier in the block and 30 foil scenes that they had not seen previously were shown. The remainder of the recognition memory task procedure was identical to that of Experiment 1. The procedure repeated 60 times, after which the block terminated.

Following completion of the first block, subjects started the second with another calibration procedure. In the second block, subjects saw the other 30 scenes (and 30 memory foils) that were not displayed during the first block, and participated in the other condition (suppression if the first block was the control, and vice versa). Each subject completed 60 experimental trials and 120 recognition memory trials total. The scenes shown in each block and the order of conditions were counterbalanced across subjects.

Attention maps Attention maps were generated in the same manner as Experiment 1.

Meaning maps Meaning maps for 30 scenes added in Experiment 2 were generated using the same procedure as the scenes tested in Experiment 1, with the following exceptions.

Raters were 148 UC Davis undergraduate students recruited through the UC Davis online subject pool. All were 18 years or older, had normal or corrected-to-normal vision, and reported no color blindness. Subjects received course credit for participation.

In each survey, catch patches showing solid surfaces (e.g., a wall) served as an attention check. Data from 25 subjects who did not attend to the task (responded incorrectly on fewer than 85% of catch trials), or did not respond to more than 10% of the questions, were excluded. Data from the remaining 123 raters were used to construct meaning maps.

Saliency maps Saliency maps were generated in the same manner as in Experiment 1.

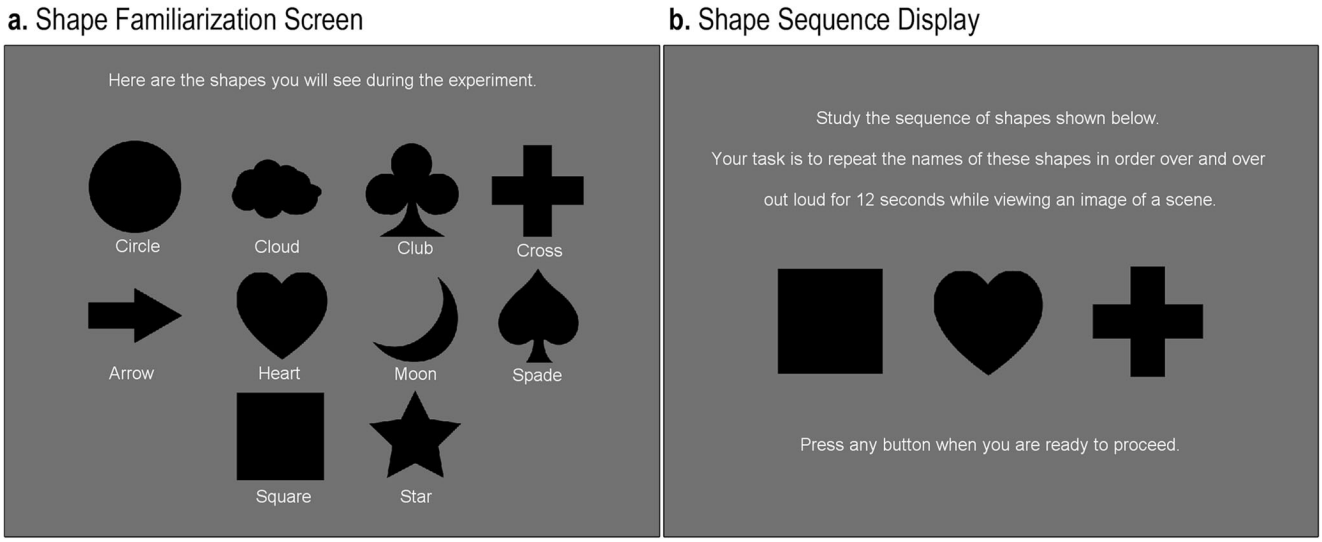

Fig. 5 Experiment 2 suppression task stimuli. (**a**) All ten shapes and shape names shown to subjects prior to the practice trials. (**b**) In the articulatory suppression condition only, the shape repetition task instructions were reiterated to subjects along with a three-shape sequence

Map normalization Maps were normalized in the same manner as in Experiment 1.

Map analyses We determined the degree to which saliency and meaning overlap for the 30 new scenes by computing feature map correlations (R^2) across the maps of 30 scenes, excluding the periphery to control for the peripheral down-weighting associated with center biasing operations. On average, meaning and saliency were correlated (R^2 = 0.51), and this relationship differed from zero (meaning and saliency: $t(29) = 23.52$, $p < 0.001$, 95% CI = [.47 .56]).

Sensitivity analysis We again conducted a sensitivity analysis, which revealed a critical t value of 2.00 and a minimum d_z of 0.47.

Scene-level analysis We correlated meaning and saliency maps with attention maps in the same manner as in Experiment 1. Squared linear and semipartial correlations (R^2) were computed within each condition for each of the scenes. The relationship between meaning and salience with visual attention was analyzed using t-tests. Cohen's d was computed, and effect sizes were interpreted in the same manner as the Experiment 1 results.

Fixation analysis We examined early fixations to replicate the early advantage of meaning over image salience observed in Experiment 1 and previous work (e.g., Henderson & Hayes, 2017). We correlated each feature map (meaning, salience) with attention maps at each fixation (Fig. 6b). Map-level correlations and t-tests were conducted in the same manner as Experiment 1.

Experiment 2: Results

We sought to replicate the results of Experiment 1 using a more robust experimental design. If verbal encoding is not required to extract meaning from scenes, we expected an advantage of meaning over salience in explaining variance in attention for both conditions. We again conducted paired t-tests within task conditions.

Scene-level analysis

Linear correlations Meaning accounted for 36% of the average variance in attention in the control condition (M = 0.36, SD = 0.16) and salience accounted for 25% (M = 0.25, SD = 0.14; Fig. 6). The advantage of meaning over salience was significant and the effect size was large ($t(59) = 6.74$, p < .001, 95% CI = [0.08 0.15], d = 0.80, d 95% CI = [0.53 1.07], d_z = 0.79). Meaning accounted for 45% of the variance in attention in the suppression condition (M = 0.45, SD = 0.15) and salience accounted for 27% (M = 0.27, SD = 0.13). Consistent with Experiment 1, the advantage of meaning over salience was significant even with verbal interference, and the effect size was large ($t(59) = 9.83$, p < .001, 95% CI = [0.14 0.22], d = 1.24, d 95% CI = [0.91 1.58], d_z = 1.30).

Semipartial correlations To account for the relationship between meaning and salience, we partialled out the shared variance explained by both. When the shared variance explained by salience was accounted for in the control condition, meaning explained 15% of the average variance in attention (M = 0.15, SD = 0.10), while salience explained 3% of the average variance after accounting for the variance explained by meaning (M = 0.03, SD = 0.05). The advantage

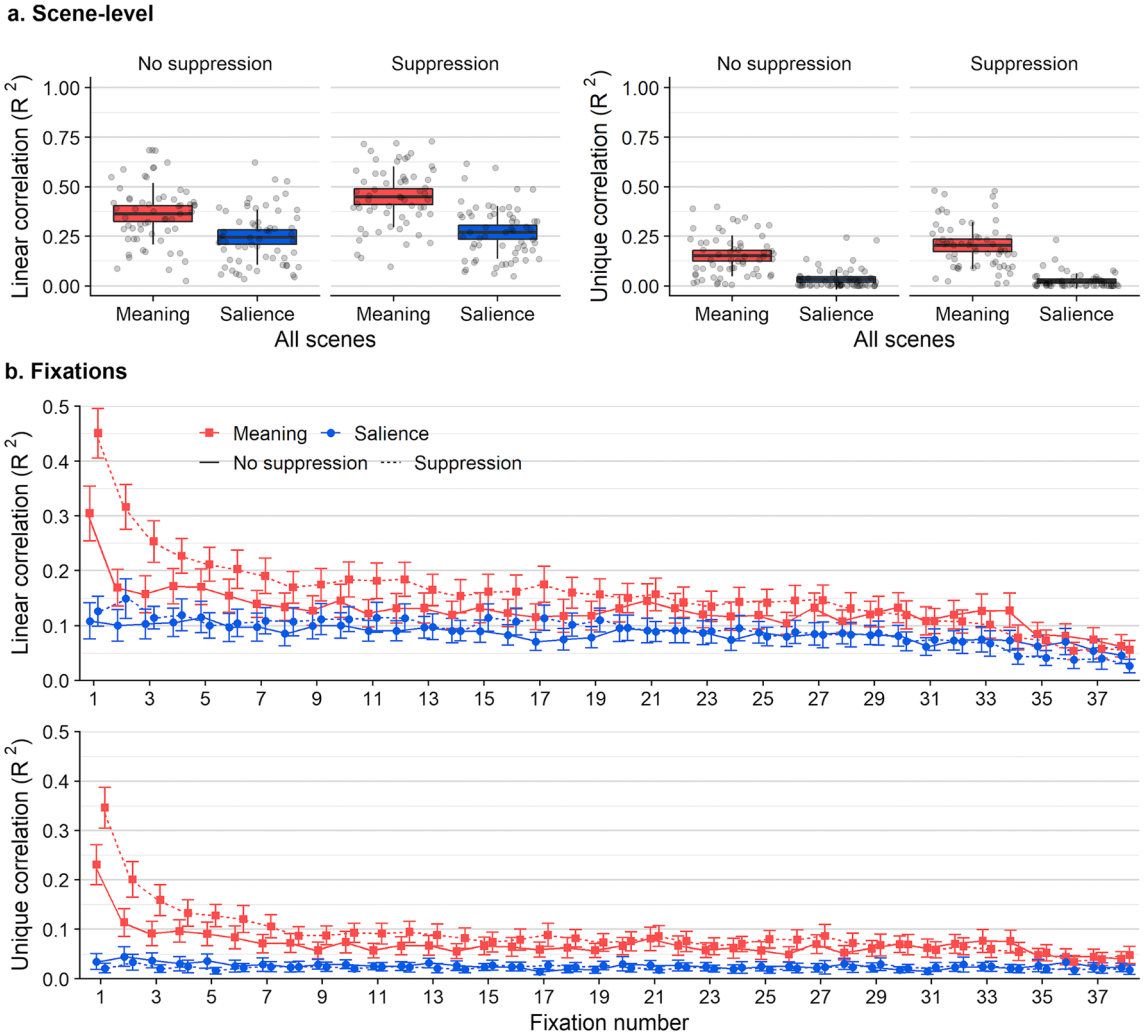

Fig. 6 (**a**) Box plots showing linear correlations (left) and semipartial correlations (right) between feature maps (meaning, saliency) and attention maps. The scatter box plots show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning (red box) and salience (blue box) across 30 scenes. (**b**) Line graphs showing linear correlations (top) and semipartial correlations (bottom) between feature maps and attention maps for each fixation (1–38) when subjects engaged in a memorization task only (solid lines) or additionally an articulatory suppression task (dashed lines). Error bars indicate 95% confidence intervals

of meaning over salience was significant, and the effect size was large ($t(59) = 6.75$, $p < .001$, 95% CI = [0.08 0.15], $d = 1.52$, d 95% CI = [0.86 2.17], $d_z = 0.90$). Meaning explained 20% of the unique variance on average after shared variance was partialled out in the articulatory suppression condition ($M = 0.20$, $SD = 0.12$), and salience explained 2% of the average variance after shared variance with meaning was accounted for ($M = 0.02$, $SD = 0.04$), and the advantage was significant with a large effect size ($t(59) = 9.83$, $p < .001$, 95% CI = [0.14 0.22], $d = 2.19$, d 95% CI = [1.38 3.00], $d_z = 1.25$).

Consistent with Experiment 1, we found a large advantage of meaning over salience in accounting for variance in attention in both conditions, for both linear and semipartial correlations, and the value of the t statistic and d_z exceeded the thresholds obtained in the sensitivity analysis.

Fixation analysis

Linear correlations In the control condition, meaning accounted for 30% of the average variance in attention during the first fixation ($M = 0.30$, $SD = 0.19$), 17% during the second ($M = .17$, $SD = .13$), and 16% during the third ($M = .16$, $SD = .13$). Salience accounted for 11% of the variance at the first fixation ($M = .11$, $SD = .13$) and 10% of the variance during the second and third fixations (2: $M = 0.10$, $SD = 0.11$; 3: $M = 0.10$, $SD = 0.11$). The advantage of meaning was significant for all three fixations, and effect sizes were medium or large (1: $t(59) = 8.17$, $p < .001$, 95% CI = [0.15 0.24], $d = 1.17$, d 95% CI = [0.80 1.54]; 2: $t(59) = 3.62$, $p = .001$, 95% CI = [0.03 0.11], $d = 0.57$, d 95% CI = [0.23 0.90]; 3: $t(59) = 3.36$, $p < .001$, 95% CI = [0.02 0.09], $d = 0.46$, d 95% CI = [0.17 0.74]). In the suppression condition, meaning accounted for

45% of the average variance during the first fixation (M = 0.45, SD = 0.17), 32% during the second (M = 0.32, SD = 0.16), and 25% during the third (M = 0.25, SD = 0.15). Salience accounted for 13% of the average variance during the first fixation (M = 0.13, SD = 0.10),15% during the second (M = 0.15, SD = 0.14), and 11% during the third (M = 0.11, SD = 0.08). The advantage of meaning over salience was significant for all of the three fixations (1: $t(59) = 14.01$, $p < .001$, 95% CI = [0.28 0.37], d = 2.21, d 95% CI = [1.63 2.79; 2: $t(59) = 7.65$, $p < .001$, 95% CI = [0.12 0.21], d = 1.13, d 95% CI = [0.75 1.50]; 3: $t(59) = 8.20$, $p < .001$, 95% CI = [0.10 0.17], d = 1.10, d 95% CI = [0.76 1.44]).

Semipartial correlations Because meaning and salience were correlated, we partialled out shared variance explained by both and analyzed semipartial correlations computed for each of the initial three fixations. In the control condition, after the shared variance explained by both meaning and salience was partialled out, meaning accounted for 23% of the average variance at the first fixation (M = 0.23, SD = 0.16), 11% of the variance during the second (M = 0.11, SD = 0.11), and 9% during the third (M = 0.09, SD = 0.10). After shared variance with meaning was partialled out, salience accounted for 3% of the average unique variance at the first fixation (M = 0.03, SD = 0.06) and 4% at the second and third (2: M = 0.04, SD = 0.08; 3: M = 0.04, SD = 0.06). The advantage of meaning was significant for all three fixations (1: $t(59) = 8.17$, $p < .001$, 95% CI = [0.15 0.24], d = 1.71, d 95% CI = [1.06 2.36]; 2: $t(59) = 3.62$, $p < .001$, 95% CI = [0.03 0.11], d = 0.74, d 95% CI = [0.28 1.20]; 3: $t(59) = 3.37$, $p < .001$, 95% CI = [0.02 0.09], d = 0.69, d 95% CI = [0.24 1.15]). In the suppression condition, after the shared variance with salience was partialled out, meaning accounted for 35% of the variance on average during the first fixation (M = 0.35, SD = 0.16), 20% of the variance at the second (M = 0.20, SD = 0.14), and 16% during the third (M = 0.16, SD = 0.12). After the shared variance with meaning was partialled out, on average salience accounted for 2% of the variance at the first and third fixations (1: M = 0.02, SD = 0.04; 3: M = 0.02, SD = 0.03) and 3% of the variance at the second (M = 0.03, SD = 0.06). The advantage of meaning was significant for all three fixations, with large effect sizes (1: $t(59) = 14.01$, $p < .001$, 95% CI = [0.28 0.37], d = 3.06, d 95% CI = [2.03 4.08]; 2: $t(59) = 7.65$, $p < .001$, 95% CI = [0.12 0.21], d = 1.61, d 95% CI = [0.98 2.25]; 3: $t(59) = 8.20$, $p < .001$, 95% CI = [0.10 0.17], d = 1.66, d 95% CI = [1.04 2.28]).

The results of Experiment 2 replicated those of Experiment 1: meaning held a significant advantage over salience when the entire viewing period was considered and when we limited our analysis to early viewing, both for linear and semipartial correlations.

Memory: Recognition As an attention check, we totaled the number of hits, correct rejections, misses, and false alarms on the recognition task for each subject (Fig. 7a). The totals for each response category ranged from 0 to 30. Recognition performance was high in both conditions. In the control condition, subjects correctly recognized scenes shown in the memorization task 97% of the time on average (M_{hits} = 0.97, SD_{hits} = 0.18), while subjects correctly recognized scenes 91% of the time after they had engaged in the suppression task during memorization (M_{hits} = 0.91, SD_{hits} = 0.29). In the control condition, subjects falsely reported that a foil had been present in the memorization scene set 1% of the time on average ($M_{false\ alarms}$ = 0.01, $SD_{false\ alarms}$ = 0.11), and in the suppression condition, the average false-alarm rate was 2% ($M_{false\ alarms}$

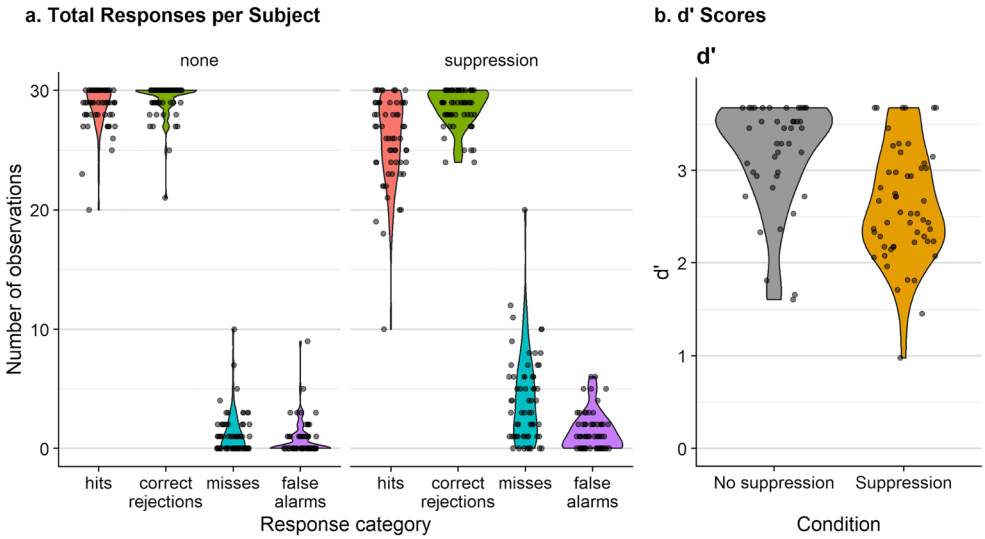

Fig. 7 (**a**) Violin plot showing the total number of recognition task responses for each subject (individual points), broken into hits, correct rejections, misses, and false alarms. (**b**) Violin plot showing d' values for each subject

= 0.02, $SD_{\text{false alarms}}$ = 0.15). Overall, recognition accuracy was higher in the control condition than the suppression condition, though the difference was small.

We then computed d' in the same manner as Experiment 1 (Fig. 7b). In the control condition, d' scores were higher on average (M = 3.43, SD = 0.60) than in the suppression condition (M = 2.76, SD = 0.71). To determine whether the difference in means was significant, we conducted a paired t-test, which revealed a significant difference with a large effect size ($t(59)$ = 6.62, $p < 0.001$, 95% CI = [0.47 0.88], d = 1.01, d 95% CI = [0.64 1.39]).

For Experiment 2, while recognition accuracy was high overall, recognition was significantly better in the control condition, when subjects memorized scenes and did not engage in the suppression task.

Experiment 2: Discussion

The attention results of Experiment 2 replicated those of Experiment 1, providing further evidence that incidental verbalization does not modulate the relationship between scene meaning and visual attention during scene viewing. Recognition performance was significantly worse in the suppression condition than in the control condition, which we cannot attribute to individual differences given that the interference manipulation was implemented within-subject. One possibility is that the shape name interference imposed greater cognitive load than the digit sequence interference; however, we cannot determine whether that was the case based on the current experiment.

General discussion

The current study tested two competing hypotheses concerning the relationship (or lack thereof) between incidental verbal encoding during scene viewing and attentional guidance in scenes. First, the relationship between scene meaning and visual attention could be mediated by verbal encoding, even when it occurs incidentally. Second, scene meaning guides attention regardless of whether incidental verbalization is available, and verbal encoding does not mediate use of scene meaning. We tested these hypotheses in two experiments using an articulatory suppression paradigm in which subjects studied scenes for a later memorization task and either engaged in a secondary task (digit or shape sequence repetition) to suppress incidental verbalization, or had no secondary task. In both experiments, we found an advantage of meaning over salience in explaining the variance in attention maps whether or not incidental verbalization was suppressed. Our results did not support the hypothesis that verbal encoding mediates attentional guidance by meaning in scenes. To the extent that observers use incidental verbalization during scene viewing, it does not appear to mediate the influence of meaning on visual attention, suggesting that meaning in scenes is not necessarily interpreted through the lens of language.

Our attentional findings do not support saliency-based theories of attentional guidance in scenes (e.g., Parkhurst et al., 2002). Instead, they are consistent with prior work showing that regions with higher image salience are not fixated more (Tatler et al., 2005) and that top-down information, including task demands, plays a greater role than image salience in guiding attention from as early as the first fixation (Einhäuser, Rutishauser, & Koch, 2008). Consistent with cognitive guidance theory, scene meaning – which captures the distribution of information across the scene – predicted visual attention better in both conditions than image salience did. Because our chosen suppression manipulation interfered with verbalization strategies without imposing undue executive load (Allen et al., 2017), our findings demonstrate that the advantage of meaning over salience was not modulated by the use of verbal encoding during scene viewing. Instead, we suggest that domain-general cognitive mechanisms (e.g., a central executive) may push attention to meaningful scene regions, although additional work is required to test this idea.

Many of the previous studies that showed an effect of internal verbalization strategies (via interference paradigms) tested simpler displays, such as arrays of objects (Meyer et al., 2007), color patches (Winawer et al., 2007), or cartoon images (Trueswell & Papafragou, 2010), while our stimuli were real-world photographs. Unlike real-world scenes, observers cannot extract scene gist from simple arrays, and may process cartoons less efficiently than natural scenes (Henderson & Ferreira, 2004). It is possible that verbal encoding exerts a greater influence on visual processing for simpler stimuli: the impoverished images may put visual cognition at a disadvantage because gist and other visual information that we use to efficiently process scenes are not available.

Limitations and future directions

We cannot know with certainty whether observers in our suppression task were unable to use internal verbal encoding. However, we would expect the secondary verbal task to have at least impeded verbalization strategies (e.g., Frank et al., 2012; Hermer-Vazquez et al., 1999; Trueswell & Papafragou, 2010; Winawer et al., 2007), and that should have impacted the relationship between meaning and attention if verbal encoding is involved in processing scene meaning. Furthermore, the suppression tasks we used (three-digit or three-shape sequences) were comparable to tasks that eliminated verbalization effects in related work (e.g., Lupyan, 2009), and so should have suppressed inner speech. We suspect that a more demanding verbal task would have imposed greater cognitive load, which could confound our results because we would not be able to separate effects of verbal interference from those of cognitive load.

Subjects in the control condition did not perform a secondary non-verbal task (e.g., a visual working-memory task). Given that our findings did not differ across conditions, we suspect controlling for the secondary task's cognitive load would not have affected the outcome. Recall that prior work has shown digit repetition tasks do not pose excessive cognitive load (Allen et al., 2017), and we would have expected lower recognition accuracy in the suppression condition if the demands of the suppression task were too great. However, we cannot be certain the verbal task did not impose burdensome cognitive load in our paradigm, and therefore this remains an issue for further investigation.

Our results are limited to attentional guidance when memorizing scenes. It is possible that verbal encoding exerts a greater influence on other aspects of visual processing, or that the extent to which verbal encoding plays a role depends on the task (Lupyan, 2012). Verbal interference may be more disruptive in a scene categorization task, for example, than in scene memorization, given that categorization often involves verbal labels.

Conclusion

The current study investigated whether internal verbal encoding processes (e.g., thought in the form of language) modulate the influence of scene meaning on visual attention. We employed a verbal interference paradigm to control for incidental verbalization during a scene memorization task, which did not diminish the relationship between scene meaning and attention. Our findings suggest that verbal encoding does not mediate scene processing, and contribute to a large body of empirical support for cognitive guidance theory.

Open Practices Statement The experiment and analyses reported here were not pre-registered. Supplemental material are available at osf.io/8mbyv/. Data are available on request.

References

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2017). Executive and perceptual distraction in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 43*(9), 1677-1693.

Cohen, J. (1988). The effect size index: d. Statistical power analysis for the behavioral sciences, 2, 284-288.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1-19.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160.

Ferreira, F., & Rehrig, G. (2019). Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, 1-11.

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive Psychology,* 64(1-2), 74-92.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Proceedings of Neural Information Processing Systems (NIPS), 19*, 545-552.

Hayes, T. R., & Henderson, J. M. (2019a). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review*, 26(5), 1683-1689.

Hayes, T. R., & Henderson, J. M. (2019b). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, 1–10. https://doi.org/10.3758/s13414-019-01849-7.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16,* 219-222.

Henderson, J. M., & Ferreira, F. (2004). Scene Perception for Psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1-58). New York, NY, US: Psychology Press.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behavior, 1*(10), 743.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision, 18*(6), 10. https://doi.org/10.1167/18.6.10

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*(5), 850-856.

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8,* 13504.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach. *Vision, 3*(2), 19.

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology, 39*(1), 3-36.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40(10-12), 1489-1506.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*(1), 161-170.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563.

Kummerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4789-4798).

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review, 16*(4), 711-718.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3*, 54.

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, B., de Sousa, H., Cansler, B. L., Shayan, S., de Vos, C., Senft, G., Enfield, N. J., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., Brown,

P., Hill, C., Le Guen, O., Hirtzel, V., van Gijn, R., Sicoli, M. A., & Levison, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences, 115*(45), 11369-11376.

Maloney, E. A., Barr, N., Risko, E. F., & Fugelsang, J. A. (2019). Verbal working memory load dissociates common indices of the numerical distance effect: Implications for the study of numerical cognition. *Journal of Numerical Cognition, 5*(3), 337-357.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific Reports,* 8:1079.

Meyer, A. S., & Damian, M. F. (2007). Activation of distractor names in the picture-picture interference paradigm. *Memory & Cognition, 35*(3), 494-503.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*(4), 710-716.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107-123.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica, 198*, 102889.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*, 20-34.

Perry, L. K., & Lupyan, G. (2013). What the online manipulation of linguistic activity can tell us about language and thought. *Frontiers in Behavioral Neuroscience, 7*, 122.

SR Research (2017). *EyeLink 1000 Plus User Manual, Version 1.0.2*. Mississauga, ON: SR Research Ltd.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1-17. doi: https://doi.org/10.1167/7.14.4.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 25,* 643-659.

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language, 63*(1), 64-82.

Ünal, E., & Papafragou, A. (2016). Interactions between language and mental representations. *Language Learning, 66*(3), 554-580.

Van Dijck, J-P. & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition, 119,* 114-119.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*(19), 7780-7785.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*(3), 0058.