Attention, Perception, & Psychophysics (2022) 84:1583–1610
https://doi.org/10.3758/s13414-022-02467-6

# Look at what I can do: Object affordances guide visual attention while speakers describe potential actions

Gwendolyn Rehrig[1] · Madison Barker[1] · Candace E. Peacock[2] · Taylor R. Hayes[3] · John M. Henderson[2] · Fernanda Ferreira[1]

Accepted: 23 February 2022 / Published online: 28 April 2022


## Abstract

As we act on the world around us, our eyes seek out objects we plan to interact with. A growing body of evidence suggests that overt visual attention selects objects in the environment that could be interacted with, even when the task precludes physical interaction. In previous work, objects that afford grasping interactions influenced attention when static scenes depicted reachable spaces, and attention was otherwise better explained by general informativeness. Because grasping is but one of many object interactions, previous work may have downplayed the influence of object affordances on attention. The current study investigated the relationship between overt visual attention and object affordances versus broadly construed semantic information in scenes as speakers describe or memorize scenes. In addition to meaning and grasp maps—which capture informativeness and grasping object affordances in scenes, respectively—we introduce interact maps, which capture affordances more broadly. In a mixed-effects analysis of 5 eyetracking experiments, we found that meaning predicted fixated locations in a general description task and during scene memorization. Grasp maps marginally predicted fixated locations during action description for scenes that depicted reachable spaces only. Interact maps predicted fixated regions in description experiments alone. Our findings suggest observers allocate attention to scene regions that could be readily interacted with when talking about the scene, while general informativeness preferentially guides attention when the task does not encourage careful consideration of objects in the scene. The current study suggests that the influence of object affordances on visual attention in scenes is mediated by task demands.

**Keywords** Eye movements and visual attention · Perception and action · Object-based attention

## Introduction

Gaze behavior can speak volumes about an observer's goals in the present moment (Henderson, 2017; Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013) and how one may act on their environment in the immediate future (David-John et al., 2021; Hayhoe & Ballard, 2005; Hayhoe & Matthis, 2018; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Pelz & Canosa, 2001; Sullivan, Ludwig, Damen, Mayol-Cuevas, & Gilchrist, 2021). When planning physical actions with visual guidance, observers look at objects they intend to interact with (Hayhoe & Ballard, 2005; Hayhoe & Matthis, 2018; Hayhoe et al., 2003), and look ahead to objects involved in later segments of the action sequence (Pelz & Canosa, 2001; Sullivan et al., 2021). Beyond what fixations on objects reveal, gaze dynamics can be used to predict when an observer is about to interact with an object (David-John et al., 2021). This evidence suggests that visual attention is systematically deployed to objects in the environment in the moments leading up to an agent interacting with an object.

The interactions one could perform with an object influence visual attention even when observers are not actively planning to interact with the object. Gomez and Snow (2017) found that object affordances guide overt attention during a visual search task; furthermore,

✉ Gwendolyn Rehrig
glrehrig@ucdavis.edu

[1] Department of Psychology, University of California, Davis, Davis, CA 95616, USA

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

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

the influence of affordances on attention is stronger for physically present objects that are within reach as opposed to 2- or 3-D object representations displayed on a screen (Gomez, Skiba, & Snow, 2018). As observers learned the function or features of novel objects (e.g., as they learned new pulling affordances of a soap container on the ceiling), successfully learning the affordances of novel objects facilitated subsequent search behavior (Castelhano & Witherspoon, 2016). Taken together, the findings suggest a strong influence of object affordances on visual attention in scenes. In the current study, we investigated whether the aforementioned influence of affordances is driven by *specific* object affordances (the ability to be grasped or manipulated) or to affordances broadly defined (an object's ability to be interacted with in any way).

The finding that visual attention orients to objects that afford interaction is consistent with cognitive guidance theory (Henderson, Brockmole, Castelhano, & Mack, 2007). According to cognitive guidance theory, visual attention is not passively pulled to regions of the scene that stand out against their surroundings (as asserted by Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002), but instead cognitive systems push visual attention to information-rich regions of the scene. Visual attention is allocated to informative objects in scenes moreso than to regions that contrast with surrounding areas in luminance, orientation, and other physical properties of the scene, as captured by image-computable saliency maps (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010), even when the information is not task-relevant (Hayes & Henderson, 2019b; Shomstein, Malcolm, & Nah, 2019) and when contrasts in physical salience *are* task-relevant (Peacock, Hayes, & Henderson, 2019b). In recent work, Henderson and Hayes (2017) developed a method to capture the spatial distribution of local semantic information in a scene using meaning maps, which were designed to be comparable to saliency maps. To construct the maps, raters were prompted to rate small patches taken from a real-world scene on the degree to which each patch was informative or recognizable. General informativeness as captured by meaning maps has been shown to account for variance in attention better than Graph-Based Visual Saliency maps (GBVS; Harel, Koch, & Perona, 2006) while observers engaged in aesthetic judgment and memorization tasks (Henderson & Hayes 2017, 2018; Rehrig, Hayes, Henderson, & Ferreira, 2020a), action and scene description tasks (Henderson, Hayes, Rehrig, & Ferreira, 2018; Rehrig, Peacock, Hayes, Henderson, & Ferreira, 2020b), and free-viewing tasks (Peacock, Hayes, & Henderson, 2019a). Furthermore, in a recognition memory task, observers were more likely to resample previously fixated regions in a scene when those regions were informative, as captured by meaning maps (Ramey, Yonelinas, & Henderson, 2020). These findings indicate that semantic information in scenes guides visual attention, consistent with the cognitive guidance theory of overt visual attention.

Meaning maps have shown that the distribution of local semantic information—broadly defined—guides visual attention in a scene (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019a, b; Rehrig et al., 2020a). While meaning maps have proven useful in demonstrating the relationship between scene semantics and visual attention, they were not intended to be a complete representation of semantic information in scenes, but instead were meant to serve as a starting point to quantify semantic information in scenes in a new way. As such, the rating instruction used to construct the original meaning maps was intentionally quite broad—to indicate how informative or recognizable the patch appeared to be, following Antes (1974) and Mackworth and Morandi (1967)—with the understanding that the features queried do not capture all of scene semantics . However, the mapping procedure is flexible in that the instructions can be modified to tap into raters' conceptions of different types of information in the scene. In our previous work (Rehrig et al., 2020b), we altered the rating instruction to investigate whether grasping affordances—the possible grasping interactions that could be performed with objects in the scene—predict visual attention when speakers describe actions that could be carried out in a scene. Our goal was not to develop a computational model that predicts viewer fixations perfectly, but rather to determine what kind of information in scenes is most relevant for the cognitive processes that give rise to overt attention during the action description task. To isolate what information is cognitively relevant, we measured different kinds of information to see what type of information predicted visual behavior best. We constructed physical saliency, meaning, and grasp maps, and correlated each map with an attention map derived from viewer fixations in three action description experiments. For typical real-world scenes, we found that meaning maps explained variance in attention maps best (consistent with Hayes & Henderson, 2019b; Henderson & Hayes, 2017; Henderson et al., 2018; Peacock et al., 2019b), but meaning and grasp maps explained comparable variance in attention when scenes instead depicted reachable spaces. The results suggest that general informativeness guides attention best overall, but grasping object affordances can guide attention as well as general semantic information does when graspable objects are shown within reach of the camera's viewpoint—in other words, when the scene itself is conducive to grasping objects.

In Rehrig et al. (2020b), we found that both grasping object affordances and general informativeness guide visual attention in scenes that depict reachable spaces, contributing to the evidence that object affordances

influence attention. However, it remains puzzling that the influence of object affordances on attention was weak for scenes that were not optimized for grasping, given that object affordances predicted attention well in other studies (Castelhano & Witherspoon, 2016; Gomez et al., 2018; Gomez & Snow, 2017). We suspect object affordances underperformed in Rehrig et al. (2020b) due to the narrow way in which we defined them. Because we operationalized object affordances as grasping affordances specifically—a very narrow type of object affordance—grasp maps were likely unable to capture the influence of object affordances on attention broadly, and therefore our prior work may have underestimated the degree to which object affordances guide visual attention. Rather than mapping scenes to capture another specific variety of object affordance, in the current study we constructed *interact maps* to capture the degree to which any type of object interaction (e.g., grasping, sitting, watching, etc.) was possible in a scene. Once we constructed a broader measure of object affordances, we re-analyzed the fixation data to determine which of the three types of semantic information we quantified was best able to predict attended scene locations using a hierarchical logistic regression model that compared meaning, grasp, and interact map values to determine which features predicted the locations that were attended in the scene.

The current study expands on Rehrig et al. (2020b) in several ways. First, we constructed interact maps for scenes in the Rehrig et al. (2020b) data set to capture broadly-defined object affordances in a scene. Second, we constructed meaning, grasp, and interact maps for the 15 scenes that were not originally included in the Rehrig et al. (2020b) analysis, doubling the number of scenes included in the Experiment 1 data set ($N = 30$ scenes). Third, to explore whether task goals mediate the influence of semantic information in the scene on attention, we analyzed eye movements in two additional data sets for which the task was not to describe the actions possible in a scene: an open-ended scene description task (Henderson et al. 2018; Experiment 4) and a scene memorization task (Rehrig et al. 2020a; Experiment 5). The two additional data sets included the same scenes and the same number of subjects as Experiment 1 in Rehrig et al. (2020b). Finally, we analyzed the data using a new approach inspired by Nuthmann, Einhäuser, and Schütz (2017) and developed by Hayes and Henderson (2021), which enabled us to examine fine-grained differences between regions that were selected for attention over other parts of the scene that were not fixated.

Nuthmann et al. (2017) developed a novel analysis approach that exploits two key assumptions about overt visual attention: 1) the regions of a scene that are prioritized for attention differ from regions that were not attended in ways that are quantifiable (e.g., using saliency maps) and 2) measurable differences between attended regions and unattended regions may explain why those regions were prioritized for attention over others (e.g., the presence of interesting objects). To that aim, Nuthmann et al. (2017) divided scenes into a pre-defined grid and assigned each square in the grid a value of 1 if any fixations fell within the square, or 0 if the square was not fixated, and conducted a logistic mixed-effects regression analysis with the average values for various saliency models and Euclidean distance from the center of the screen for each square in the grid as predictors in the model. Hayes and Henderson (2021) expanded on Nuthmann et al. (2017)'s approach to obviate the need for a grid, instead measuring center proximity (inverted from Euclidean center distance) and feature values (in this case, semantic values for objects computed from ConceptNet) in a 3° diameter window approximating the size of the fovea around each fixation coordinate, as well as from randomly sampled locations that were not fixated.

In the current study, we implemented Hayes and Henderson's (2021) analysis approach on fixation data from 5 eyetracking experiments previously reported in Henderson et al. (2018) and Rehrig et al. (2020a, b). Our goal was to determine whether overt visual attention is guided by semantic information broadly construed, or by object affordances. To that aim, we assessed whether general informativeness, graspability, or interactability predicted visual attention across 3 different types of tasks: 1) action description, in which speakers describe the actions that could be carried out in a scene, 2) scene description, in which speakers describe a scene however they like, and 3) scene memorization, in which observers study a scene in preparation for a later recognition memory task. Because image salience did not predict attention well in our previous work (Henderson et al., 2018; Rehrig et al., 2020b; Rehrig et al., 2020a), we instead focused only on three different operationalizations of semantic information in the new analysis: general informativeness, graspability, and interactability, as captured by meaning, grasp, and interact maps, respectively. Based on our previous work (Henderson et al., 2018; Rehrig et al., 2020a, b), we expected meaning to predict fixated locations well overall across tasks. With respect to the action description Experiments (1–3) specifically, we expected meaning map values to perform better than grasp map values in Experiments 1 and 2, and we expected both meaning and grasp map values to predict fixated locations well in Experiment 3 because the scenes depicted reachable spaces. If general object affordances as captured by interact maps predict attention better than the narrowly-defined grasping affordances, we expected interact map values to predict fixated locations better than grasp map values in

all three action description experiments, and to perhaps rival general informativeness. If object affordances guide attention even when they are less task-relevant, but might still be mentioned in a description, we expect interact map values—and possibly grasp map values—to predict fixated locations when observers described scenes however they liked (Experiment 4). Likewise, if object affordances guide attention generally (as suggested by Gomez et al., 2018; Gomez & Snow, 2017), not just when the task is not explicitly linguistic in nature, then we similarly expect interact map and grasp map values to predict fixated locations well in a scene memorization task (Experiment 5).

## Methods

### Eyetracking data collection

#### Subjects

All subjects were undergraduate students enrolled at the University of California, Davis who participated in exchange for course credit. They spoke English as a first language, were 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. Thirty-two subjects in Experiment 1 participated (2 excluded from analysis); 48 participated in Experiment 2 (8 excluded), 49 participated in Experiment 3 (9 excluded from analysis), 38 participated in Experiment 4 (8 excluded), and 68 participated in Experiment 5 (8 excluded). Across experiments, subjects were excluded from analysis either because their eyes could not be tracked accurately, or due to errors caused by software, hardware, or because of experimenter error. In Experiment 5 only, 30 of the subjects completed a secondary task in addition to memorizing scenes. The secondary task was an articulatory suppression task in which subjects repeated a sequence of digits aloud while viewing the scene, which was intended to prevent subjects from using internal language to facilitate memorization of the scene. The original study showed no effect of articulatory suppression on the relationship between scene informativeness and attention (see Rehrig et al. 2020a for details). For the purpose of the current analysis, we chose to examine the control condition only (the scene memorization task with no secondary task) in order to draw a clean comparison with the description experiments that involved fewer changes in experimental parameters; data from 30 participants in the control condition were analyzed.

#### Stimuli

In all experiments, digitized and luminance-matched photographs of real-world scenes depicting indoor and outdoor environments were presented at 1024 × 768 resolution. There were 30 scenes presented in Experiments 1, 4, and 5, and 20 in Experiment 2 (15 of which were also presented in Experiment 1). In Experiment 3, 20 scenes were presented, 15 of which were photographed by the first and third authors to depict reachable spaces. For those 15 scenes, the authors confirmed that objects in the foreground of the scene were within reach of the scene's viewpoint. The remaining scenes were drawn from other studies: four from Xu, Jiang, Wang, Kankanhalli, and Zhao (2014) and one from Rehrig, Cullimore, Henderson, and Ferreira (2021). Text was removed from each scene presented in Experiment 3 using the clone stamp and patch tools in Adobe Photoshop CS4. One scene in Experiment 2 showed people in the background of the image; faces were not present in the other 54 scenes. See Appendix for all 55 scenes and feature maps.

#### Apparatus

In all experiments, eye movements were recorded with an SR Research EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01) at a sampling rate of 1000 Hz. Head movements were minimized using a chin and forehead rest integrated with the eyetracker's tower mount. Although viewing was binocular, eye movements were recorded from the right eye only. The experiment was controlled using SR Research Experiment Builder software. Audio was recorded digitally at a rate of 48 kHz using a Shure SM86 cardioid condenser microphone.

In Experiments 1, 4, and 5, subjects sat 85 cm away from a 21" monitor such that scenes subtended approximately 27° × 20.5° visual angle, and audio was recorded digitally at a rate of 48 kHz using a Roland Rubix 22 USB audio interface and a Shure SM86 cardioid condenser microphone. In Experiments 2 and 3, subjects sat 83 cm away from a 24.5" monitor such that scenes subtended approximately 27° × 20.5° visual angle at a resolution of 1024 × 768 pixels, presented in 4:3 aspect ratio. For both Experiments 2 and 3, data were collected on two separate systems that were identical except that the operating system for the subject computer in one system was Windows 10, and Windows 7 on the other.

#### Procedure

A calibration procedure was conducted at the beginning of each session 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°. Fixations and saccades were parsed with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s$^2$; SR Research, 2017).

After successful calibration, subjects received task instructions. In Experiments 1 and 2, the instructions were as follows: "In this experiment, you will see a series of scenes. In each scene, think of the average person. Describe what the average person would be inclined to do in the scene. You will have 30 s to respond." In Experiment 3, subjects were instead instructed as follows: "In this experiment, you will see a series of scenes. For each scene, describe what you would do in the scene. You will have 30 s to respond." In Experiment 4, subjects were instructed to describe scenes as follows: "In this experiment, you will see a series of scenes. You will have 30 s to describe the scene out loud.". In Experiment 5, subjects were instructed to study a series of scenes for a later memory test. In each experiment, the instruction was followed by three practice trials that allowed subjects to familiarize themselves with the task and the duration of the response window. Subjects pressed any button on a button box to advance throughout the task.

The task instruction was repeated before subjects began the experimental block (Fig. 1a). Within the block, each subject received a unique pseudo-random trial order that prevented two scenes of the same type (e.g., living room) from occurring consecutively. A trial proceeded as follows. First, a five-point fixation array was displayed to check calibration (Fig. 1b). The subject fixated the center cross and the experimenter pressed a key to begin the trial if the fixation was stable, otherwise the experimenter reran the calibration procedure. The scene was then shown for a period of 30 s (Experiments 1–4) or 12 s (Experiment 5), during which time eye-movements were recorded (Fig. 1c). In Experiments 1–4, audio was also recorded during scene viewing. After the scene viewing period ended, subjects were instructed to press a button to proceed to the next trial (Fig. 1d). The trial procedure repeated until all trials were complete (Experiments 1, 4, & 5 = 30 trials, Experiments 2 and 3 = 20 trials). In Experiment 5 only, subjects completed a recognition memory test comprised of the 30 scenes presented in the experiment and 30 image foils depicting similar scenes.

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 outliers (amplitude >20°).

## Meaning, grasp, and interact map generation

We used the same meaning and grasp maps generated in all three experiments as described in Rehrig et al. (2020b). We additionally mapped 15 scenes for informativeness and graspability, and mapped all 55 scenes for interactability. The mapping procedure was identical between the current study and Rehrig et al. (2020b). In the interest of brevity, we describe details of the mapping procedure only for maps introduced in the current study.

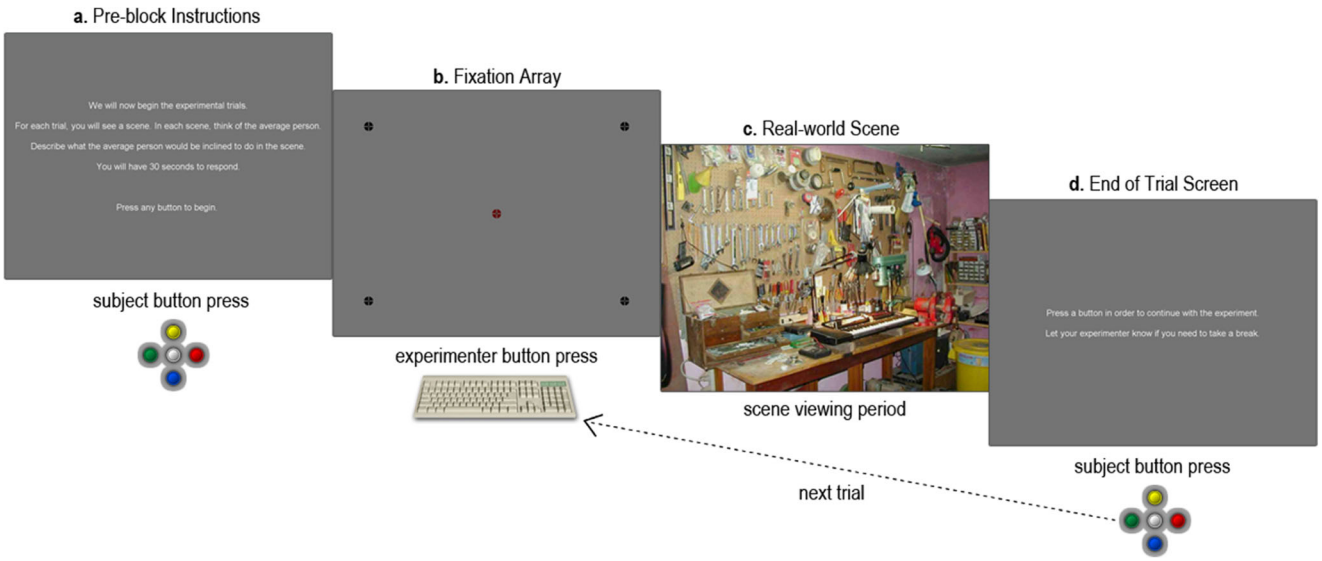



**Fig. 1** Visualization of the trial procedure for each of the 3 eyetracking experiments. First, (a) task instructions were reiterated to subjects following the practice trials. (b) A five-point fixation array was used to gauge calibration quality. (c) A real-world scene was shown for 30 s. Eye-movements were recorded for the duration of the viewing period in all experiments; audio was additionally recorded in Experiments 1–4. (d) Subjects pressed a button to initiate the next trial. After pressing the button, the trial procedure repeated (from b)

### Meaning maps

Meaning maps were generated using a contextualized rating procedure in which subjects viewed small circular patches drawn from the scene alongside a thumbnail image showing the full scene that included a green circle showing what region the patch came from (Peacock et al., 2019a). Each of the 15 scenes (1024 × 768 pixel) was decomposed into a series of partially overlapping circular patches at fine and coarse spatial scales (Fig. 2b&c), resulting in 4,500 unique fine-scale patches (93 pixel diameter) and 1,620 unique coarse-scale patches (217 pixel diameter), 6,120 patches in total.

Raters were 97 undergraduates enrolled at UC Davis who participated through Sona. Students received credit toward a course requirement for participating. Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from the 15 scenes, presented alongside a small (256 × 192 pixel) image of the scene for context. 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 of low-meaning and two examples of high-meaning scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order.

Ten catch trials, which were easy for a human completing the task in good faith to answer correctly, were included in each survey to serve as an attention check. Each catch trial presented a unique catch patch to the subject to rate, which showed a blank surface drawn from the scene (usually a wall or ceiling; see Fig. 3). As in the test trials, catch patches were presented alongside an image showing where in the scene the patch was drawn from so that subjects were not aware the trial was an attention check. If subjects complete the task in accordance with the examples provided in the task instructions, catch patches should be rated as low in meaning (a value of 1 or 2 on the Likert scale). To score catch trial performance, ratings of 2 or lower were considered correct responses, and ratings of 3 or higher were scored as incorrect. Ratings from 34 subjects who scored below 80% on the catch patches were excluded. Each unique patch was rated at least 3 times by 3 independent raters for a total of 18,360 ratings.

Meaning maps were generated from the ratings 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]. This procedure was used for each scene. The final map was blurred using a Gaussian filter via the MATLAB function 'imgaussfilt' with a sigma of 10 (see Fig. 2e for an example meaning map).

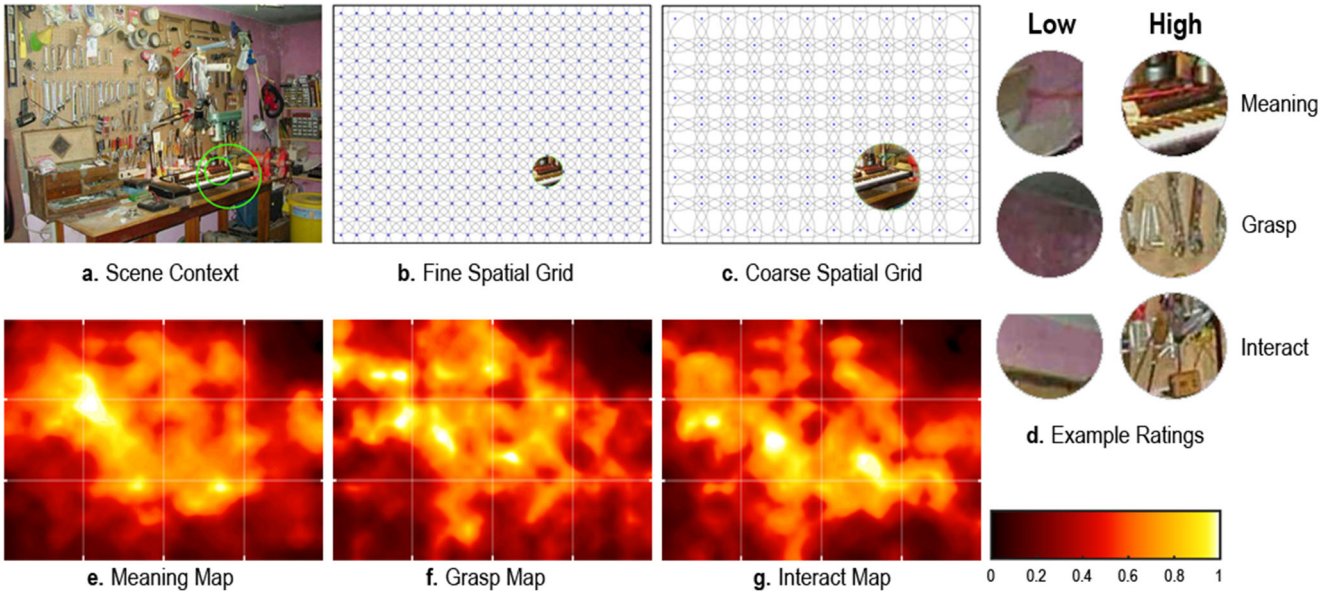



**Fig. 2** (a–d) Feature map generation schematic. (a) Real-world scene. Raters saw the real-world scene and either a fine (inner) or coarse (outer) green circle indicating the origin of the scene patch under consideration. (b–c) Fine-scale (b) and coarse-scale (c) spatial grids used to create scene patches. (d) Examples of scene patches that were rated as low or high with respect to meaning, grasp, and interact. (e–g) Examples of meaning (e), grasp (f), and interact (g) maps for the scene shown in (a)


**Fig. 3** Examples of fine- (a) and coarse-scale (b) catch patches that were included as attention checks

### Grasp maps

Grasp maps were constructed from ratings in the same manner as meaning maps, with the critical exception that subjects rated each patch on how 'graspable' the region of the scene shown in the patch was. In the instructions, we defined 'graspability' as how easily an object depicted in the patch could be picked up or manipulated by hand. If a patch contained more than one object or only part of an object, raters were instructed to use the object or entity that occupied the most space in the patch as the basis for their rating. The remainder of the procedure was identical to the one used to generate meaning maps.

Raters were 83 undergraduates enrolled at UC Davis who participated through Sona. Students received credit toward a course requirement for participating. Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from the 15 scenes, presented alongside a scene thumbnail for context. Subjects were instructed to rate how graspable 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-graspability and high-graspability scene patches in the instructions to ensure that they understood the task. Scene-patch pair presentation order was random. Ratings from 20 subjects that scored below 80% on the catch patches were excluded. Each unique patch was rated at least 3 times by 3 independent raters for a total of 18,360 ratings.

Grasp maps were generated in the same manner as the meaning maps. Ratings were averaged, smoothed, and combined across scales. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. This procedure was used for each scene. An example grasp map can be seen in Fig. 2f.

### Interact maps

Interact maps were constructed in the same manner as meaning and grasp maps, except subjects were asked to rate the region of the scene that was visible in each patch based on how 'interactable' it was. We defined 'interactability' as the extent to which the subject viewed what was shown as an object with which a human might interact. As in the grasp map generation procedure, subjects were again instructed to rate the object that occupied the majority of the patch.

Each of the 55 scenes (1024 × 768 pixel) was decomposed into a series of partially overlapping circular patches at fine and coarse spatial scales (Fig. 2b&c), resulting in 16,500 unique fine-scale patches (93 pixel diameter) and 5,940 unique coarse-scale patches (217 pixel diameter), 22,440 patches in total.

Raters were 328 undergraduates enrolled at UC Davis who participated through Sona.[1] Students received credit toward a course requirement for participating. Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from the 55 scenes, presented alongside a scene thumbnail for context. Subjects were instructed to rate how interactable 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-interactability and high-interactability scene patches in the instructions to ensure that they understood the task. Scene-patch pair presentation order was random. Ratings from 103 subjects that scored below 80% on the catch patches were excluded. Each unique

[1] More raters were used because we had to generate interact maps for 55 scenes, as opposed to only 15 scenes for the other map types.

patch was rated at least 3 times by 3 independent raters (at least 67,320 ratings in total).

Interact maps were generated in the same manner as the meaning and grasp maps. Ratings were averaged, smoothed, and combined across scales. This procedure was used for each scene. An example interact map is shown in Fig. 2g.

Overall, the resulting meaning, grasp, and interact maps were correlated with one another; the correlation was particularly high for grasp and interact maps in Experiment 3 ($M_{R^2} = 0.72$, $SD_{R^2} = 0.10$)(Table 1).

### Analysis

Following Nuthmann et al. (2017), we examined which features influenced visual attention by comparing the feature map values at locations in the scene that were fixated to those for locations that were not, operating on the assumption that differences between regions of the scene that were and were not fixated speak to what information is prioritized for attention. Rather than dividing the scene into a grid (as Nuthmann et al., 2017 did), we elected to use the procedure developed by Hayes and Henderson (2021) to measure meaning, grasp, and interact map values in a window around each location, and compared the values for fixated locations to those of sampled locations in the scene that were not fixated.

Specifically, we conducted a logistic mixed-effects regression analysis in which the dependent variable was whether subjects fixated a location (1) or not (0). The dependent variable was defined as follows. For each subject and each trial, the x,y coordinates corresponding to the subject's fixations were assigned a value of 1 (fixated). A number of locations that were not fixated equal to the number of fixated locations were then randomly sampled from all possible coordinates in the 1024 × 768 image using the 'sample' function from the 'random' module in Python 3. Locations that the subject fixated during that trial, or locations that fell within a 1.5° visual angle (56 pixel) radius around the fixated location, were excluded from the sample space. The randomly sampled coordinates were assigned a value of 0 (not fixated).

We accounted for center bias in our model (Tatler, 2007; Hayes & Henderson, 2019a) using the center proximity measure developed by Hayes and Henderson (2021). We calculated the inverted Euclidean distance between the center of the scene and each other pixel in the image and stored the value for each pixel in a 1024 × 768 matrix. The Euclidean distance was z-scored and inverted for ease of interpretation such that higher values indicate closer proximity to the center of the scene.

For each x,y coordinate pair, we then computed the mean feature and center proximity map values corresponding to a 3° visual angle (113 pixel) diameter window around the coordinate. We defined a mask for the region around the fixation using a 56 pixel radius. The mask was then used to extract an array of map values for the meaning, grasp, interact, and center proximity maps, and the mean of each array was stored as the average feature map values corresponding to the x,y coordinate under consideration (Fig. 4).

A logistic mixed-effects model was constructed for each experiment's data using the 'glmer' function of the 'lme4' package in R (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2021). Each model was maximally specified to include fixed effects of center proximity, meaning, grasp, and interact, as well as interactions between each. Random intercepts and random slopes corresponding to fixed effects and their interactions were included in both random effect structures. To facilitate model convergence, all predictors were centered and scaled using the 'scale' function in base R prior to analysis, and random slopes and intercepts were uncorrelated. All models used the default optimizer (bobyqa). Random effects were included for subjects and items (scenes). Because the data sets in use are large, the maximum number of model iterations was increased to 100,000.

**Table 1** Correlations ($R^2$) between feature maps

| Experiment | Feature maps | Correlation ($R^2$) | |
|---|---|---|---|
| | | *M* | *SD* |
| 1,4,&5 | Meaning × Grasp | 0.538 | 0.185 |
| 1,4,&5 | Meaning × Interact | 0.451 | 0.209 |
| 1,4,&5 | Grasp × Interact | 0.499 | 0.22 |
| 2 | Meaning × Grasp | 0.526 | 0.224 |
| 2 | Meaning × Interact | 0.496 | 0.222 |
| 2 | Grasp × Interact | 0.562 | 0.195 |
| 3 | Meaning × Grasp | 0.471 | 0.175 |
| 3 | Meaning × Interact | 0.469 | 0.182 |
| 3 | Grasp × Interact | 0.715 | 0.097 |

## Results

### Experiment 1

In Experiment 1, 30 subjects were asked to describe actions the average person could carry out in each of 30 real-world scenes. We predicted that object affordances as captured by interact and grasp maps would predict regions in the scene that were selected for attention as subjects described possible actions, because objects that can be interacted with are task-relevant.

Locations in the scene that were fixated were more informative on average ($M = 3.05$, $SD = 0.76$) than

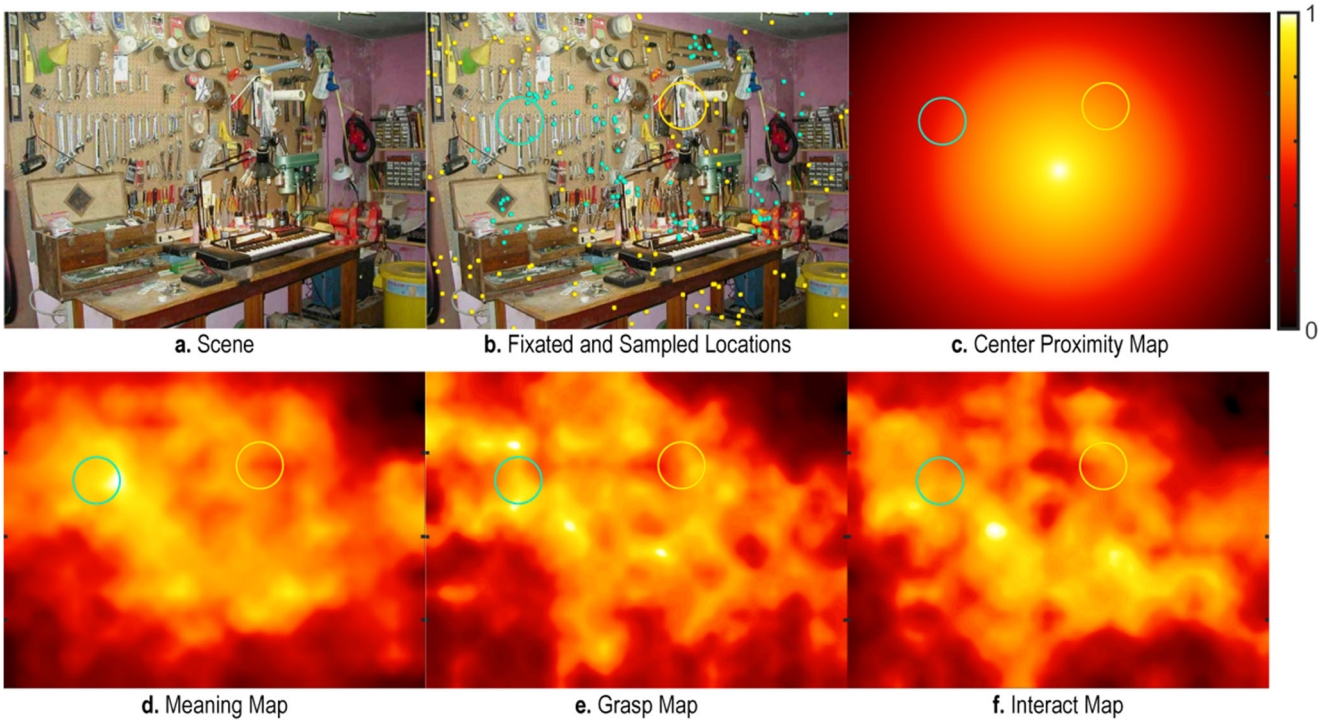



**Fig. 4** Visualization of analysis approach. (a) Real-world scene. (b) Scene overlaid with fixated (yellow) and randomly sampled (cyan) location coordinates. Circles illustrate the mask radius used to compute average feature map values around each fixated (cyan) or sampled (yellow) coordinate. (c) Center proximity map. (d–f) Meaning (d), grasp (e), and interact (f) maps for the scene shown in (a)

randomly sampled locations ($M = 2.62$, $SD = 0.77$) (Fig. 5a). Locations that were fixated also had higher grasp map values on average ($M = 3.26$, $SD = 0.86$) than randomly sampled locations that were not ($M = 2.91$, $SD = 0.88$). Interact map values were also higher on average for locations that were fixated ($M = 3.16$, $SD = 0.89$) than those that were not fixated ($M = 2.74$, $SD = 0.90$). Consistent with center bias, fixated locations had higher center proximity on average ($M = 0.60$, $SD = 0.95$) than randomly sampled locations that were not fixated ($M = -0.28$, $SD = 0.89$).

Consistent with the hypothesis that object affordances broadly influence visual attention, there was a simple main effect of interact such that subjects were more likely to fixate locations that had higher interact map values ($\beta = 0.48$, $z = 4.40$, $p < .0001$) (Table 2). Counter to our predictions, there was no simple main effect of meaning ($\beta = 0.10$, $z = 1.14$, $p = 0.26$). There was a reliable interaction between grasp and interact such that locations in the scene that had low interact map values were more likely to be fixated if they had high grasp map values ($\beta = -0.17$, $z = -2.29$, $p = .02$). The model revealed a simple main effect of center proximity such that subjects were more likely to fixate locations near the center of the image ($\beta = 0.82$, $z = 13.67$, $p < .0001$), consistent with center bias (Tatler, 2007). There was a reliable interaction between center proximity and meaning such that locations further from the screen center were more likely to be fixated if they had higher meaning map values ($\beta = -0.15$, $z = -2.88$, $p = .004$), and an opposite reliable interaction between center proximity and grasp such that locations further from the center of the scene were less likely to be fixated if they had high grasp map values ($\beta = 0.17$, $z = 2.69$, $p = .007$) (Fig. 6). Finally, there was a marginal interaction between interact and center proximity such that regions of the scene in the periphery were marginally more likely to be fixated if they had high interact map values ($\beta = -0.13$, $z = -1.91$, $p = .06$). No other predictors were significant.

In sum, interact map values predicted fixated locations in the scene better than meaning or grasp, which were only influential in interactions which revealed that observers deviated from the center of the image to pursue locations that were more informative or interactable, but not highly graspable. As expected, fixated locations were closer to the center of the image, reflecting center bias.

## Experiment 2

In Experiment 2, we again asked 40 subjects to describe actions the average person could carry out in each of 20 real-world scenes, and once again we anticipated object

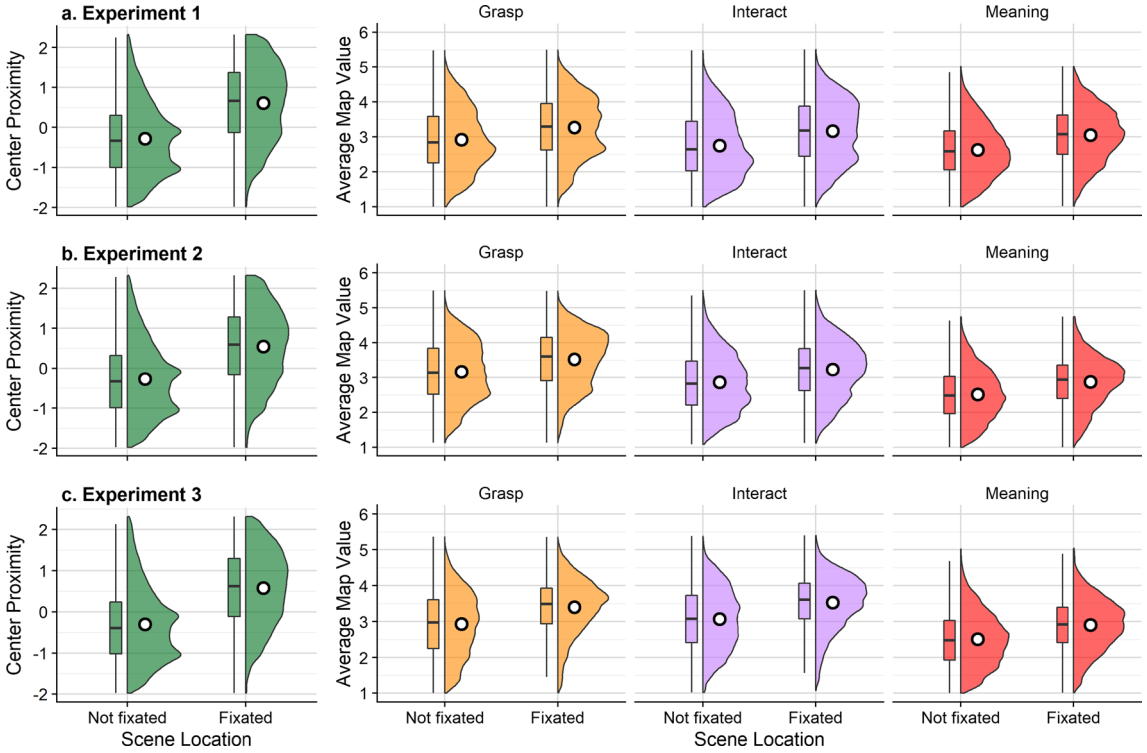



**Fig. 5** Hybrid violin and box plots. Data for each of the three experiments is shown on separate rows. In each row, the left panel shows center proximity (green) both for sampled coordinates that were not fixated and for fixated locations (x-axis). The right panel shows the average grasp map values around the image coordinate (yellow-orange), average interact map values (violet), and average meaning map values (red), shown separately for locations that were randomly sampled and fixated locations (x-axis). Center proximity values and map values reflect $z$-values and Likert ratings (1–6), respectively. White points superimposed over the violins indicate the grand mean. On the box plots to the left of each violin, black horizontal lines correspond to the median, colored boxes indicate the 25% and 75% quartile boundaries, and black vertical lines show ± 1.5 IQR (the interquartile range)

**Table 2** Experiment 1 logistic mixed-effects model output

| Predictors | Fixed effects | | | | Random effects ($SD$) | |
|---|---|---|---|---|---|---|
| | $\beta$ | $SE$ | $z$ | $p$ | Subject | Scene |
| Intercept | 0.004 | 0.09 | 0.05 | 0.96 | 0.10 | 0.44 |
| Meaning | 0.10 | 0.09 | 1.14 | 0.26 | 0.16 | 0.44 |
| Grasp | −0.07 | 0.09 | −0.76 | 0.45 | 0.17 | 0.43 |
| Interact | 0.48 | 0.11 | 4.40 | <.0001⋆ | 0.29 | 0.51 |
| Meaning:Grasp | −0.03 | 0.09 | −0.37 | 0.71 | 0.15 | 0.43 |
| Meaning:Interact | 0.08 | 0.09 | 0.86 | 0.39 | 0.19 | 0.45 |
| Grasp:Interact | −0.17 | 0.08 | −2.29 | 0.02⋆ | 0.13 | 0.37 |
| Meaning:Grasp:Interact | −0.05 | 0.05 | −0.98 | 0.32 | 0.06 | 0.22 |
| Center Proximity | 0.82 | 0.06 | 13.67 | <.0001⋆ | 0.22 | 0.21 |
| Center Proximity:Meaning | −0.15 | 0.05 | −2.88 | 0.004⋆ | 0.09 | 0.24 |
| Center Proximity:Grasp | 0.17 | 0.06 | 2.69 | 0.007⋆ | 0.11 | 0.28 |
| Center Proximity:Interact | −0.13 | 0.07 | −1.91 | 0.06† | 0.12 | 0.31 |
| Center Proximity:Meaning:Interact | −0.06 | 0.05 | −1.20 | 0.23 | 0.06 | 0.21 |
| Center Proximity:Grasp:Interact | −0.02 | 0.07 | −0.24 | 0.81 | 0.11 | 0.30 |
| Center Proximity:Meaning:Grasp | 0.09 | 0.06 | 1.57 | 0.12 | 0.07 | 0.28 |
| Center Proximity:Meaning:Grasp:Interact | 0.008 | 0.03 | 0.23 | 0.82 | 0.05 | 0.15 |

⋆Denotes a significant predictor or interaction

†Denotes a marginal predictor or interaction

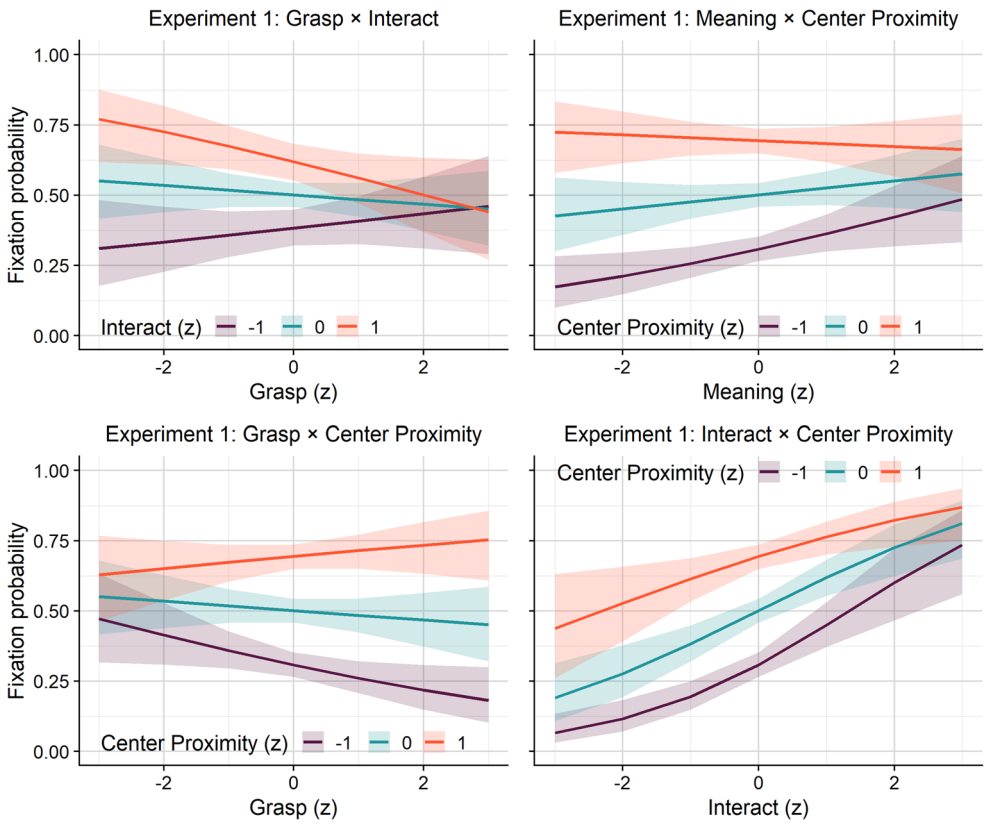



**Fig. 6** Estimated fixation probability (y-axis) for marginal and significant interactions for $z$-scored predictors in Experiment 1. Shaded regions indicate 95% confidence intervals. The top row shows the interactions of grasp (x-axis) with interact (lines) and meaning (x-axis) with center proximity (lines). The bottom row shows interactions of grasp (x-axis) and center proximity (lines) and interact (x-axis) and center proximity (lines)

affordances (as captured by grasp and interact maps) would predict the regions that were fixated in the scene.

As in Experiment 1, fixated locations had higher average meaning map values ($M = 2.87$, $SD = 0.70$) than randomly sampled locations that were not fixated ($M = 2.51$, $SD = 0.73$), and higher grasp map values ($M = 3.52$, $SD = 0.79$) than randomly sampled locations ($M = 3.17$, $SD = 0.84$). Consistent with our hypothesis and the results of Experiment 1, fixated locations in the scene had higher interact map values ($M = 3.23$, $SD = 0.81$), than those that were not fixated ($M = 2.86$, $SD = 0.83$). Finally, fixated locations were closer to the center of the image on average ($M = 0.54$, $SD = 0.94$) than locations that were sampled from parts of the scene that were not fixated ($M = -0.27$, $SD = 0.90$).

Consistent with Experiment 1, there was a simple main effect of interact such that subjects were more likely to fixate locations that had higher interact map values ($\beta = 0.38$, $z = 2.67$, $p = 0.008$) (Table 3). There was a reliable interaction between meaning and grasp such that locations with high meaning values were more likely to be fixated if they also had high grasp map values ($\beta = 0.23$, $z = 2.63$, $p = 0.009$) (Fig. 7). The model revealed a simple main effect of center proximity reflecting center bias ($\beta = 1.15$, $z = 8.47$, $p < .0001$). There was a marginal interaction between meaning and center proximity such that locations further from the center of the image were marginally more likely to be fixated if they were informative ($\beta = -0.19$, $z = -1.80$, $p = 0.07$), and a marginal three-way interaction between grasp, interact, and center proximity such that regions close to the center of the image were marginally more likely to be fixated when they had low interact map values if they had high grasp map values ($\beta = -0.21$, $z = -1.88$, $p = 0.06$). No other predictors were significant.

Consistent with Experiment 1, interact map values predicted fixated locations in the scene, and meaning and grasp did not predict fixated locations well independently, though each had some influence in interactions. There was a significant effect of center bias such that fixated locations were closer to the center of the image, and there was a marginal interaction between center proximity, interact, and grasp such that regions in the center of the image that were low in interactability were more likely to be fixated if were high in graspability.

**Table 3** Experiment 2 logistic mixed-effects model output

| Predictors | Fixed effects | | | | Random effects ($SD$) | |
|---|---|---|---|---|---|---|
| | $\beta$ | $SE$ | $z$ | $p$ | Subject | Scene |
| Intercept | 0.11 | 0.10 | 1.02 | 0.31 | 0.52 | 0.28 |
| Meaning | −0.007 | 0.14 | −0.05 | 0.96 | 0.24 | 0.60 |
| Grasp | 0.10 | 0.10 | 0.98 | 0.33 | 0.17 | 0.42 |
| Interact | 0.38 | 0.14 | 2.67 | 0.008* | 0.28 | 0.59 |
| Meaning:Grasp | 0.23 | 0.09 | 2.63 | 0.009* | 0.20 | 0.41 |
| Meaning:Interact | 0.11 | 0.13 | 0.84 | 0.40 | 0.27 | 0.52 |
| Grasp:Interact | −0.18 | 0.11 | −1.58 | 0.11 | 0.19 | 0.47 |
| Meaning:Grasp:Interact | −0.07 | 0.05 | −1.23 | 0.22 | 0.10 | 0.22 |
| Center Proximity | 1.15 | 0.14 | 8.47 | <.0001* | 0.73 | 0.31 |
| Center Proximity:Meaning | −0.19 | 0.11 | −1.80 | 0.07† | 0.17 | 0.45 |
| Center Proximity:Grasp | 0.0009 | 0.11 | 0.008 | 0.99 | 0.18 | 0.43 |
| Center Proximity:Interact | −0.003 | 0.11 | −0.03 | 0.97 | 0.16 | 0.45 |
| Center Proximity:Meaning:Grasp | 0.09 | 0.11 | 0.83 | 0.41 | 0.18 | 0.46 |
| Center Proximity:Meaning:Interact | 0.02 | 0.08 | 0.28 | 0.78 | 0.21 | 0.28 |
| Center Proximity:Grasp:Interact | −0.21 | 0.11 | −1.88 | 0.06† | 0.24 | 0.45 |
| Center Proximity:Meaning:Grasp:Interact | 0.002 | 0.04 | 0.04 | 0.97 | 0.14 | 0.16 |

*Denotes a significant predictor or interaction

†Denotes a marginal predictor or interaction

## Experiment 3

In Experiment 3, we asked 40 subjects to describe actions that they *personally* would carry out in each of 20 real-world scenes, which depicted reachable spaces (Josephs & Konkle, 2020). We anticipated object affordances (as captured by grasp and interact maps) might predict the regions that were fixated in the scene more strongly than in the first two experiments because the task instruction was personalized and the scenes depicted spaces that afford object interactions particularly well.

Fixated locations again had higher average meaning map values ($M = 2.90$, $SD = 0.71$) than sampled locations did ($M = 2.51$, $SD = 0.75$). Grasp map values were also higher

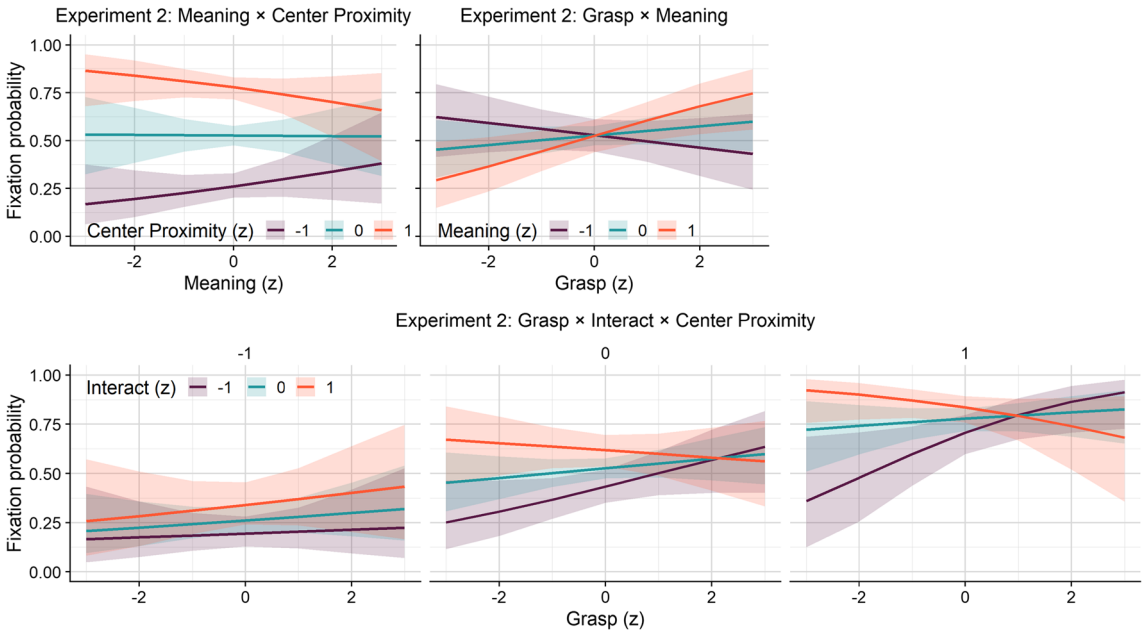



**Fig. 7** Estimated fixation probability (y-axis) for marginal and significant interactions for $z$-scored predictors in Experiment 2. Shaded regions indicate 95% confidence intervals. The top row shows 2-way interactions: meaning (x-axis) with center proximity (lines) and grasp (x-axis) with meaning (lines). The bottom row shows the 3-way interaction between grasp (x-axis), interact (lines), and center proximity (facets)

on average for fixated locations ($M = 3.40$, $SD = 0.77$) than sampled locations ($M = 2.93$, $SD = 0.88$). Finally, fixated locations in the scene again had higher interact map values ($M = 3.53$, $SD = 0.74$), than randomly sampled locations that were not fixated did ($M = 3.07$, $SD = 0.85$). Once again, fixated locations were closer to the center of the image ($M = 0.57$, $SD = 0.91$) on average than randomly sampled locations were ($M = -0.31$, $SD = 0.90$).

Consistent with the previous two experiments, there was a simple main effect of interact such that subjects were more likely to fixate locations that had higher interact map values ($\beta = 0.22$, $z = 2.53$, $p = 0.01$) (Table 4). There was a marginal effect of grasp such that regions were marginally more likely to be fixated when they had higher grasp map values ($\beta = 0.15$, $z = 1.67$, $p = 0.095$). The model revealed a simple main effect of center proximity reflecting center bias ($\beta = 1.02$, $z = 14.52$, $p < .0001$). There was a reliable interaction between center proximity, meaning, and interact such that locations further from the center of the image were more likely to be fixated when both interact and meaning map values were high ($\beta = -0.15$, $z = -2.36$, $p = 0.02$), and a reliable interaction between center proximity, grasp, and interact such that locations in the periphery with low interact map values were more likely to be fixated if they had high grasp map values (Fig. 8). No other predictors were significant.

In Experiment 3, interact map values again predicted fixated locations in the scene better than meaning or grasp, though grasp was a marginal independent predictor. There was again a reliable center bias on fixated locations, and there were reliable interactions between center proximity, meaning, and interact map values and center proximity, grasp, and interact map values.

## Experiment 4

To determine whether the finding that interactability predicts fixated locations generalizes to a description task for which object interactions are less task-relevant, we applied the analysis performed on the action description tasks (Experiments 1–3) to fixation data from an open-ended description task (Henderson et al., 2018) that used the same 30 scenes presented in Experiment 1.

If object interactions guide attention in scenes even when actions are less task-relevant, we anticipate that the analysis will show a strong predictive relationship between interact map values and fixated locations; however, if interact map values predicted well in Experiments 1–3 because object interactions were highly task relevant—but not generally more important than general informativeness for visual attention—we expect meaning map values to predict fixated locations better than interact map values.

**Table 4** Experiment 3 logistic mixed-effects model output

| Predictors | Fixed effects | | | | Random effects ($SD$) | |
|---|---|---|---|---|---|---|
| | $\beta$ | $SE$ | $z$ | $p$ | Subject | Scene |
| Intercept | 0.04 | 0.04 | 1.00 | 0.32 | 0.11 | 0.13 |
| Meaning | 0.01 | 0.07 | 0.16 | 0.88 | 0.25 | 0.27 |
| Grasp | 0.15 | 0.09 | 1.67 | 0.095† | 0.22 | 0.35 |
| Interact | 0.22 | 0.09 | 2.53 | 0.01⋆ | 0.19 | 0.35 |
| Meaning:Grasp | 0.12 | 0.11 | 1.16 | 0.25 | 0.12 | 0.45 |
| Meaning:Interact | 0.04 | 0.08 | 0.44 | 0.66 | 0.11 | 0.33 |
| Grasp:Interact | −0.09 | 0.10 | −0.89 | 0.37 | 0.12 | 0.44 |
| Meaning:Grasp:Interact | −0.03 | 0.03 | −1.11 | 0.27 | 0.08 | 0.10 |
| Center Proximity | 1.02 | 0.07 | 14.52 | <.0001⋆ | 0.34 | 0.19 |
| Center Proximity:Meaning | −0.05 | 0.06 | −0.93 | 0.35 | 0.11 | 0.22 |
| Center Proximity:Grasp | −0.11 | 0.08 | −1.36 | 0.18 | 0.10 | 0.33 |
| Center Proximity:Interact | 0.002 | 0.08 | 0.03 | 0.98 | 0.11 | 0.35 |
| Center Proximity:Meaning:Grasp | 0.04 | 0.07 | 0.58 | 0.56 | 0.16 | 0.23 |
| Center Proximity:Meaning:Interact | −0.15 | 0.06 | −2.36 | 0.02⋆ | 0.16 | 0.22 |
| Center Proximity:Grasp:Interact | 0.12 | 0.06 | 1.96 | 0.049⋆ | 0.12 | 0.23 |
| Center Proximity:Meaning:Grasp:Interact | 0.02 | 0.03 | 0.64 | 0.52 | 0.06 | 0.11 |

⋆Denotes a significant predictor or interaction

†Denotes a marginal predictor or interaction

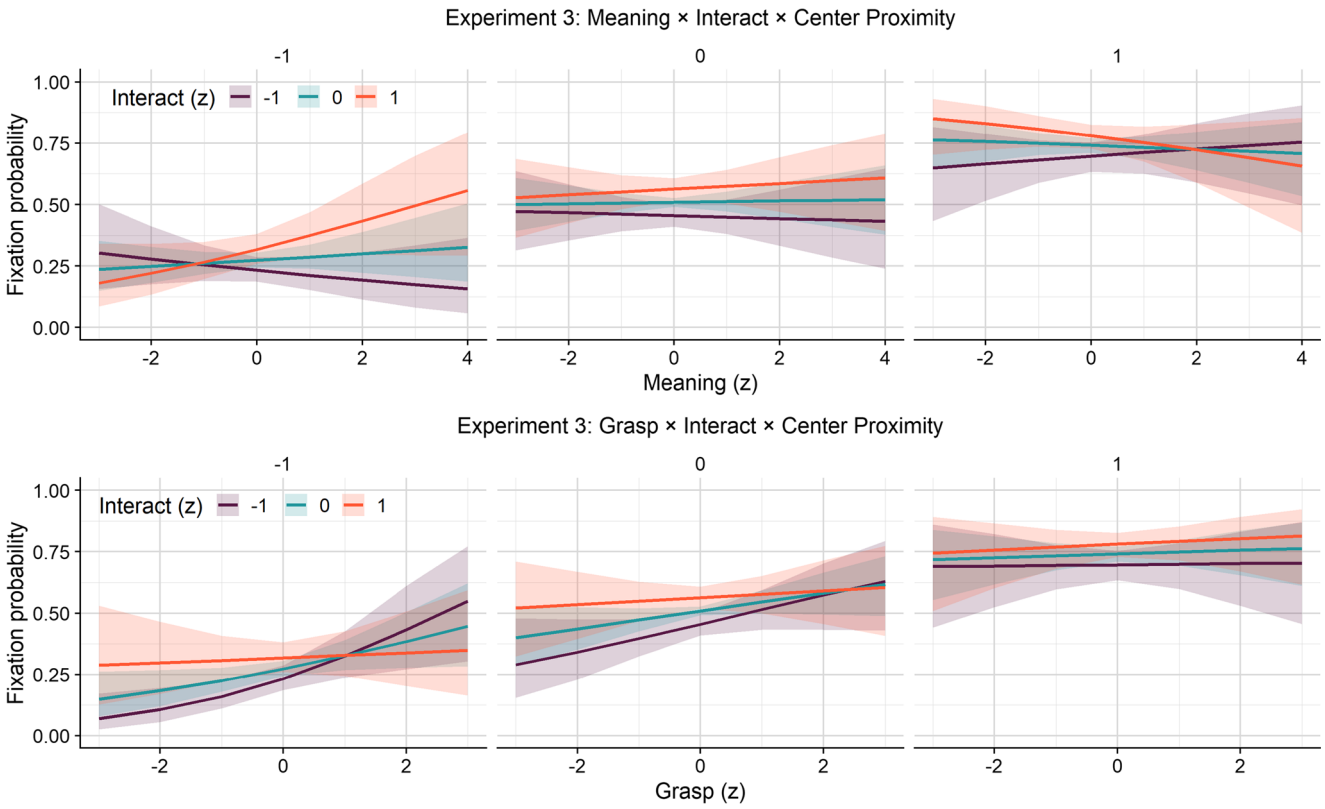



**Fig. 8** Estimated fixation probability (y-axis) for significant 3-way interaction between $z$-scored predictors in Experiment 3: in the top row, grasp (x-axis), interact (lines), and center proximity (facets); in the bottom row, meaning (x-axis), interact (lines), and center proximity (facets). Shaded regions indicate 95% confidence intervals

The analysis was identical to that of Experiments 1–3, with the following exception: the maximal model produced singular fit, therefore the random slope that accounted for negligible variance (an interaction between center proximity, meaning, and grasp in the subject random effect) was pruned from the model (following Barr, Levy, Scheepers, & Tily 2013). The resulting model converged without error.

When subjects described scenes however they liked, the average meaning map values were higher for fixated locations ($M = 3.30$, $SD = 0.68$) than sampled locations ($M = 2.51$, $SD = 0.73$) (Fig. 9). Grasp map values were also higher on average for fixated ($M = 3.48$, $SD = 0.83$) as opposed to sampled locations ($M = 2.81$, $SD = 0.85$). Consistent with the action description experiments, fixated locations in the scene also had higher interact map values

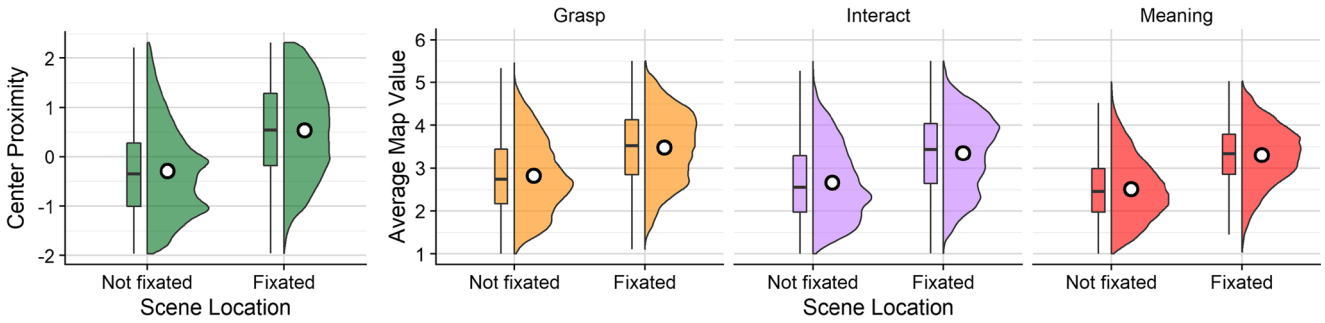



**Fig. 9** Hybrid violin and box plots for predictors Experiment 4. The left panel shows center proximity (green) both for sampled coordinates that were not fixated and for fixated locations (x-axis). The right panel shows the average grasp map values around the image coordinate (yellow-orange), average interact map values (violet), and average meaning map values (red), shown separately for locations that were randomly sampled and fixated locations (x-axis). Center proximity values and map values reflect $z$-values and Likert ratings (1–6), respectively. White points superimposed over the violins indicate the grand mean. On the box plots to the left of each violin, black horizontal lines correspond to the median, colored boxes indicate the 25% and 75% quartile boundaries, and black vertical lines show ± 1.5 IQR (the interquartile range)

($M = 3.35$, $SD = 0.88$) than randomly sampled locations that were not fixated ($M = 2.66$, $SD = 0.86$). Finally, fixated locations were, on average, closer to the center of the image ($M = 0.53$, $SD = 0.93$) than randomly sampled locations were ($M = -0.29$, $SD = 0.90$).

As in the action description experiments, in the open-ended scene description task there was a simple main effect of interact such that subjects were more likely to fixate locations that had higher interact map values ($\beta = 0.41$, $z = 2.66$, $p = 0.008$) (Table 5). Counter to the action description tasks, there was a simple main effect of meaning such that subjects were more likely to fixate locations with higher meaning map values ($\beta = 0.90$, $z = 7.70$, $p < .0001$). As expected, the model revealed a simple main effect of center proximity reflecting center bias ($\beta = 0.44$, $z = 4.08$, $p < .0001$). There was a marginal interaction between center proximity, meaning, and grasp such that locations near the center of the scene were marginally more likely to be fixated if they had high grasp and meaning map values ($\beta = .21$, $z = 1.93$, $p = 0.05$), and there was a reliable 4-way interaction between center proximity, meaning, grasp, and interact such that locations in the periphery of the scene that had high meaning map values, but lower grasp and interact map values, were more likely to be fixated ($\beta = -0.20$, $z = -2.63$, $p = 0.009$) (Fig. 10). No other predictors were significant.

In Experiment 4, both interact and meaning map values predicted fixated locations in the scene, whereas grasp map values did not, and there was again a reliable center bias on fixated locations.

## Experiment 5

To determine whether interactability predicts fixated locations well in a task that does not encourage the viewer to think about objects in the scenes and how they would interact with those objects, we applied the analysis performed in Experiments 1–4 to fixation data from a scene memorization task (Rehrig et al., 2020a) that used the same 30 scenes presented in Experiments 1 and 4. In Experiment 5, 30 subjects memorized 30 real-world scenes for a period of 12 s each in preparation for a later recognition memory task. Following Rehrig et al. (2020a), we expect general informativeness to predict fixated locations well. If the strong predictive relationship between object interactability and attention observed in Experiments 1–4 generalizes beyond language tasks, we additionally expect interact map values to predict fixated locations.

When subjects studied scenes for a later memorization task, the average meaning map values were higher for fixated locations ($M = 3.34$, $SD = 0.69$) than for randomly sampled locations that had not been fixated ($M = 2.60$,

**Table 5** Experiment 4 logistic mixed-effects model output

| Predictors | Fixed effects | | | | Random effects ($SD$) | |
|---|---|---|---|---|---|---|
| | $\beta$ | $SE$ | $z$ | $p$ | Subject | Scene |
| Intercept | 0.05 | 0.14 | 0.33 | 0.74 | 0.08 | 0.78 |
| Meaning | 0.90 | 0.12 | 7.70 | < .0001* | 0.19 | 0.59 |
| Grasp | 0.25 | 0.15 | 1.64 | 0.10 | 0.17 | 0.80 |
| Interact | 0.41 | 0.16 | 2.66 | 0.008* | 0.39 | 0.74 |
| Meaning:Grasp | −0.07 | 0.12 | −0.58 | 0.56 | 0.14 | 0.61 |
| Meaning:Interact | 0.07 | 0.13 | 0.54 | 0.59 | 0.17 | 0.65 |
| Grasp:Interact | −0.01 | 0.12 | −0.09 | 0.93 | 0.11 | 0.60 |
| Meaning:Grasp:Interact | −0.05 | 0.09 | −0.50 | 0.62 | 0.08 | 0.46 |
| Center Proximity | 0.44 | 0.11 | 4.08 | < .0001* | 0.28 | 0.50 |
| Center Proximity:Meaning | −0.18 | 0.12 | −1.52 | 0.13 | 0.14 | 0.58 |
| Center Proximity:Grasp | 0.19 | 0.13 | 1.44 | 0.15 | 0.07 | 0.68 |
| Center Proximity:Interact | −0.04 | 0.12 | −0.31 | 0.75 | 0.08 | 0.62 |
| Center Proximity:Meaning:Grasp | 0.21 | 0.11 | 1.93 | 0.05† | – | 0.54 |
| Center Proximity:Meaning:Interact | −0.006 | 0.12 | −0.05 | 0.96 | 0.04 | 0.64 |
| Center Proximity:Grasp:Interact | 0.07 | 0.09 | 0.76 | 0.45 | 0.07 | 0.44 |
| Center Proximity:Meaning:Grasp:Interact | −0.20 | 0.08 | −1.63 | 0.009* | 0.07 | 0.37 |

*Denotes a significant predictor or interaction

†Denotes a marginal predictor or interaction

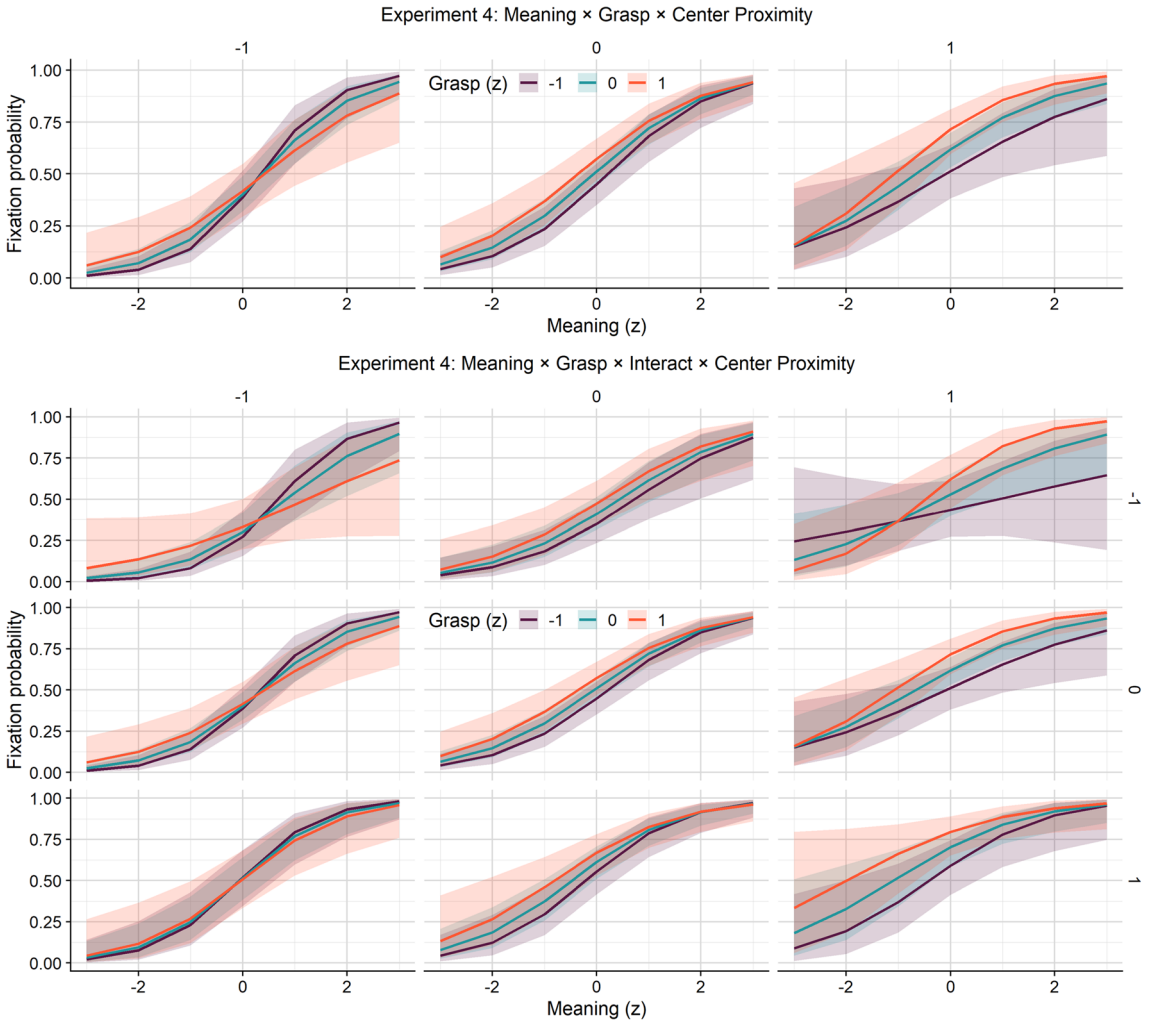



**Fig. 10** Estimated fixation probability (y-axis) for interactions between $z$-scored predictors in Experiment 4. The top figure illustrates a marginal three-way interaction between meaning (x-axis), grasp (lines) and center proximity (columns). The bottom figure visualizes a reliable four-way interaction between meaning (x-axis), grasp (lines), interact (rows) and center proximity (columns). Shaded regions indicate 95% confidence intervals

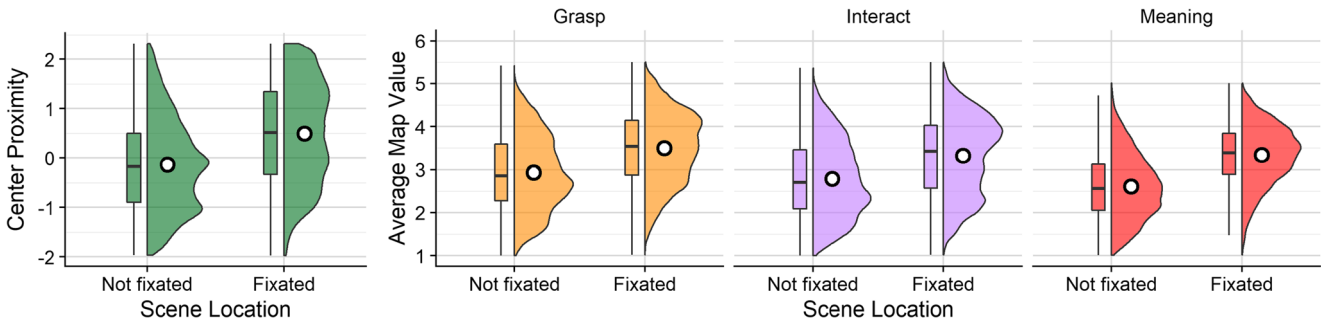



**Fig. 11** Hybrid violin and box plots for predictors Experiment 5. The left panel shows center proximity (green) both for sampled coordinates that were not fixated and for fixated locations (x-axis). The right panel shows the average grasp map values around the image coordinate (yellow-orange), average interact map values (violet), and average meaning map values (red), shown separately for locations that were randomly sampled and fixated locations (x-axis). Center proximity values and map values reflect $z$-values and Likert ratings (1–6), respectively. White points superimposed over the violins indicate the grand mean. On the box plots to the left of each violin, black horizontal lines correspond to the median, colored boxes indicate the 25% and 75% quartile boundaries, and black vertical lines show ± 1.5 IQR (the interquartile range)

**Table 6** Experiment 5 logistic mixed-effects model output

| Predictors | Fixed effects | | | | Random effects ($SD$) | |
|---|---|---|---|---|---|---|
| | $\beta$ | $SE$ | $z$ | $p$ | Subject | Scene |
| Intercept | −0.24 | 0.12 | −1.97 | 0.05 | 0.13 | 0.62 |
| Meaning | 1.49 | 0.13 | 11.70 | <.0001* | 0.09 | 0.64 |
| Grasp | 0.12 | 0.16 | 0.75 | 0.46 | 0.05 | 0.80 |
| Interact | −0.15 | 0.17 | −0.92 | 0.36 | 0.12 | 0.87 |
| Meaning:Grasp | −0.23 | 0.11 | −2.05 | 0.04* | 0.06 | 0.53 |
| Meaning:Interact | 0.12 | 0.14 | 0.88 | 0.38 | 0.05 | 0.71 |
| Grasp:Interact | 0.14 | 0.15 | 0.94 | 0.35 | 0.03 | 0.75 |
| Meaning:Grasp:Interact | 0.04 | 0.08 | 0.54 | 0.59 | 0.05 | 0.37 |
| Center Proximity | 0.25 | 0.11 | 2.23 | 0.03* | 0.36 | 0.46 |
| Center Proximity:Meaning | −0.06 | 0.13 | −0.46 | 0.65 | 0.09 | 0.67 |
| Center Proximity:Grasp | 0.01 | 0.13 | 0.08 | 0.94 | 0.03 | 0.64 |
| Center Proximity:Interact | 0.07 | 0.13 | 0.56 | 0.57 | 0.08 | 0.64 |
| Center Proximity:Meaning:Grasp | −0.16 | 0.11 | −1.50 | 0.13 | 0.05 | 0.52 |
| Center Proximity:Meaning:Interact | 0.11 | 0.13 | 0.82 | 0.41 | 0.08 | 0.67 |
| Center Proximity:Grasp:Interact | 0.02 | 0.09 | 0.22 | 0.82 | 0.05 | 0.42 |
| Center Proximity:Meaning:Grasp:Interact | −0.06 | 0.08 | −0.72 | 0.47 | 0.04 | 0.40 |

*Denotes a significant predictor or interaction

$SD = 0.75$) (Fig. 11). Grasp map values were also higher on average for fixated ($M = 3.50$, $SD = 0.83$) as opposed to sampled locations ($M = 2.92$, $SD = 0.87$). Consistent with the scene description experiments, fixated locations in the scene also had higher interact map values ($M = 3.31$, $SD = 0.91$) than randomly sampled locations that were not fixated ($M = 2.78$, $SD = 0.89$). Finally, fixated locations were, on average, closer to the center of the image ($M = 0.49$, $SD = 1.02$) than randomly sampled locations were ($M = -0.14$, $SD = 0.94$).

Unlike Experiments 1–3, but consistent with Experiment 4, in the scene memorization task there was a simple main effect of meaning: Subjects were more likely to fixate locations that had higher meaning map values ($\beta = 1.49$, $z = 11.70$, $p < .0001$) (Table 6). There was a reliable interaction between meaning and grasp such that locations that had low meaning map values were more likely to be fixated when the corresponding grasp map values were high ($\beta = -0.23$, $z = -2.05$, $p = 0.04$) (Fig. 12). Consistent with all other experiments, the model revealed a simple main effect of center proximity reflecting center bias ($\beta = 0.25$, $z = 2.23$, $p < 0.03$). No other predictors were significant.

In stark contrast to the previous experiments, of the three feature maps used, meaning map values were the only reliable independent predictor of fixated locations in Experiment 5, though there was a reliable interaction between meaning and grasp map values. Consistent with all of the previous experiments, there was a reliable effect of center bias on fixated locations.

## General discussion

In four data sets used in the current analysis, fixated locations were predicted by interact map values such that locations that were highly interactable were more likely to be fixated, consistent with the prediction that interactability could rival general informativeness in predicting overt visual attention, which follows from the hypothesis that object affordances influence attention in scenes. However, interact map values predicted fixated locations only for description tasks (Experiments 1–4), but failed to predict fixated locations when the task did not have an explicit

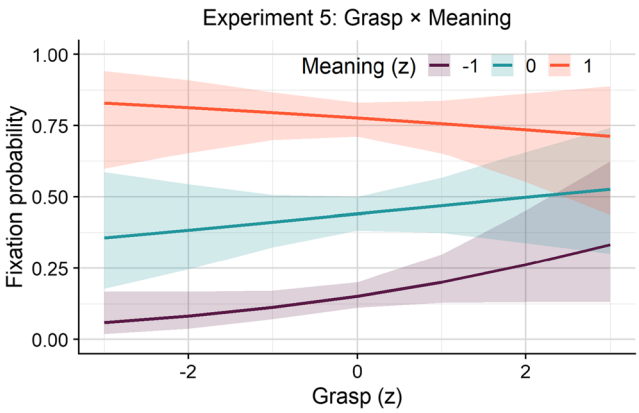



**Fig. 12** Estimated fixation probability (y-axis) for significant interaction between $z$-scored predictors in Experiment 5: grasp (x-axis) and meaning (lines). Shaded regions indicate 95% confidence intervals

language component (Experiment 5). When the task was not to describe the scene (scene memorization), only meaning map values predicted what locations in the scene were fixated. Partially consistent with Rehrig et al. (2020b) and with our predictions, higher grasp map values marginally predicted fixated locations only when scenes depicted reachable spaces (Experiment 3), and otherwise grasp contributed to reliable interactions in all experiments. Counter to our predictions, meaning map values were not a significant predictor as a simple main effect in any of the action description experiments; however, general informativeness was influential in tasks for which object interactions were less task-relevant (Experiments 4 & 5).

Our findings for Experiments 1–4 suggest that object affordances broadly defined (as captured by interact maps) predict locations prioritized for visual attention in scenes during description tasks; however, in a task that did not encourage the viewer to think about objects in the scene or their interactions (Experiment 5), affordances as operationalized in the current study did not predict fixated locations. The aforementioned findings are difficult to reconcile with those in the literature that show an influence of object affordances on attention in visual search tasks (Castelhano & Witherspoon, 2016; Gomez et al., 2018; Gomez & Snow, 2017). One possible explanation put forth by Rehrig et al. (2020b) is that prior work demonstrating a role of object affordances on attention in more traditional visual attention experiments (such as visual search; Castelhano and Witherspoon, 2016; Gomez et al., 2018; Gomez & Snow, 2017) may have been driven by other object-related information (such as recognizabilty or informativeness) that was better captured by informativeness than general affordances in the current study. However, it might also be the case that the 2-dimensional nature of the task used in the current study was unable to speak to the role of object affordances in guiding attention to physically present objects as demonstrated by Gomez et al. (2018). We leave the challenging task of investigating whether attentional guidance is better explained by general informativeness or object affordances for 3-dimensional or physically present objects to future work.

The influence of affordances on attention in the description tasks (Experiments 1–4) is consistent with literature implicating object affordances in language processing broadly (Borghi, 2012; Borghi & Riggio, 2009; Feven-Parsons & Goslin, 2018; Glenberg et al., 2009; Glenberg & Kaschak, 2002; Grafton, Fadiga, Arbib, & Rizzolatti, 1997; Harpaintner, Sim, Trumpp, Ulrich, & Kiefer, 2020; Kaschak & Glenberg, 2000; Martin, 2007). Neuroimaging studies have revealed motor activation associated with object-related cognitive processes (Martin, 2007), and specifically with language processes such as silently naming an object (Grafton et al., 1997), or making lexical decisions about action words (Harpaintner et al., 2020). A priming study showed an object's semantics are not prioritized over its affordances when processing object names (Feven-Parsons & Goslin, 2018). Evidence from language comprehension suggests that we interpret sentences through human action (Glenberg et al., 2009; Glenberg & Kaschak, 2002; Kaschak & Glenberg, 2000): for example, an object's affordances can facilitate detection of the object's name in sentences (Borghi, 2012; Borghi & Riggio, 2009). Studies of language-mediated visual attention suggest that, while listening to speech presented concurrent with scene viewing, observers attend to objects in a scene with affordances that are compatible with those of the events or objects mentioned (Altmann & Kamide, 1999; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Chambers, Tanenhaus, & Magnuson, 2004; Kako & Trueswell, 2000; Kamide, Altmann, & Haywood, 2003), particularly when object affordances are task-relevant (Salverda, Brown, & Tanenhaus, 2011). Altmann and Kamide (2007) argued that processes of language comprehension activate conceptual representations associated with a referent (such as an object or event), and, in turn, visual attention seeks out objects in the scene that have compatible object affordances. The results of the current study are compatible with the idea that the mediating effects of language on visual attention described by Altmann and Kamide (2007) may extend to language production tasks, and to the allocation of attention in real-world scenes. However, the current study cannot differentiate between the possibility that object affordances influenced attention in Experiments 1–4 because description tasks engage the language system, or because description tasks encourage the observer to think about objects in the scene and the interactions they afford more carefully than other tasks would. Future work will be needed to determine which of the two possibilities best explains the observed relationship between object affordances and visual attention.

It is worth noting that Experiments 1, 4, & 5 used the same stimuli and maps, and tested the same number of subjects, yet the influence of the types of semantic information we quantified in the current study (informativeness, graspability, and interactability) on attention differed in each. We attribute the difference to the observer's task, which changed across the three experiments. When object affordances were most task relevant—as observers described the potential actions available to them in a scene—interactability predicted attended locations better than meaning or graspability (Experiment 1), but when observers simply described what they saw,

informativeness and interactability both guided attention (Experiment 4). Finally, when object affordances were least task-relevant, informativeness influenced visual attention and interactability did not (Experiment 5). The difference in findings dependent on the task instruction supports the idea that object affordances exert a greater influence on cognition when they are task-relevant (Ostarek & Huettig, 2019).

In the previous analysis using much of the same data, grasping affordances only explained variation in fixation density effectively when the scenes depicted reachable spaces, which led us to conclude that affordances guide attention for stimuli that are especially conducive to acting on the environment (Rehrig et al., 2020b). In contrast, the present analysis revealed that object affordances broadly, as captured by interact maps, predicted fixated locations during the same action description experiments even in scenes that were less clearly conducive to interaction, including the scenes for which graspability did not explain attention well in the prior study. Consistent with Rehrig et al. (2020b), grasping affordances—as captured by grasp maps—marginally predicted fixated locations when scenes depicted reachable spaces. Through comparing the results of the current analysis with those reported in Rehrig et al. (2020b), we conclude that grasping affordances influence attention only when objects would be within reach (conducive to grasping), despite the fact that possible grasping interactions are task-relevant in all scenes, but object affordances more broadly exert a strong influence on attention when the possible actions in the scene are relevant to the speaker's goals. These findings suggest that the possible grasping actions in an environment are only relevant to observers when the object is within reach and thus would be readily acted upon; however, an alternative explanation is simply that any highly constrained, specific affordance—be it grasping, lifting, sitting, etc.—would underperform in our model relative to a representation that captures a wide range of possible interactions with the environment. It is further possible that graspability would perform as well as, or perhaps better, than interactability in a task for which grasping interactions specifically were highly task-relevant—for example, if observers were asked to study a scene for the purpose of planning to sanitize objects in the scene, or to pack the items in the room for a move. We leave investigation of the latter possibility to future work.

Our findings further illustrate the flexibility and utility of the mapping procedure, originally developed to construct meaning maps, in capturing different types of semantic information in scenes (Henderson, Hayes, Peacock, & Rehrig, 2021). The meaning, grasp, and interact maps used in the current study are all primarily derived from stored semantic representations of objects and scene categories that comprise semantic knowledge for scenes. Although each map taps semantic representations in a similar way, and the maps are correlated with one another, the fact that each differed in their ability to predict fixated locations across tasks indicates that the different maps tapped dissociable forms of semantic information, which suggests raters in the crowd-sourcing task were sensitive to variation in the instructions and followed them diligently.

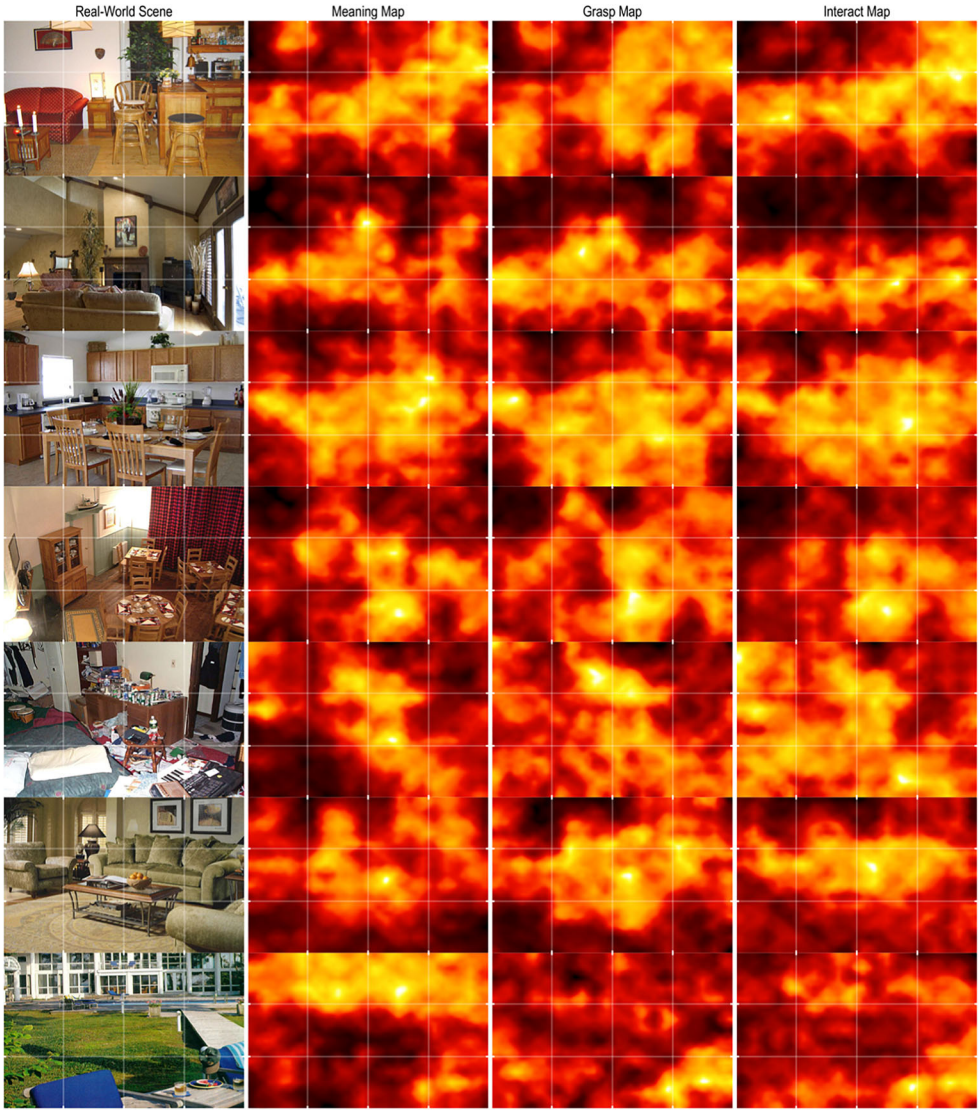
## Conclusion

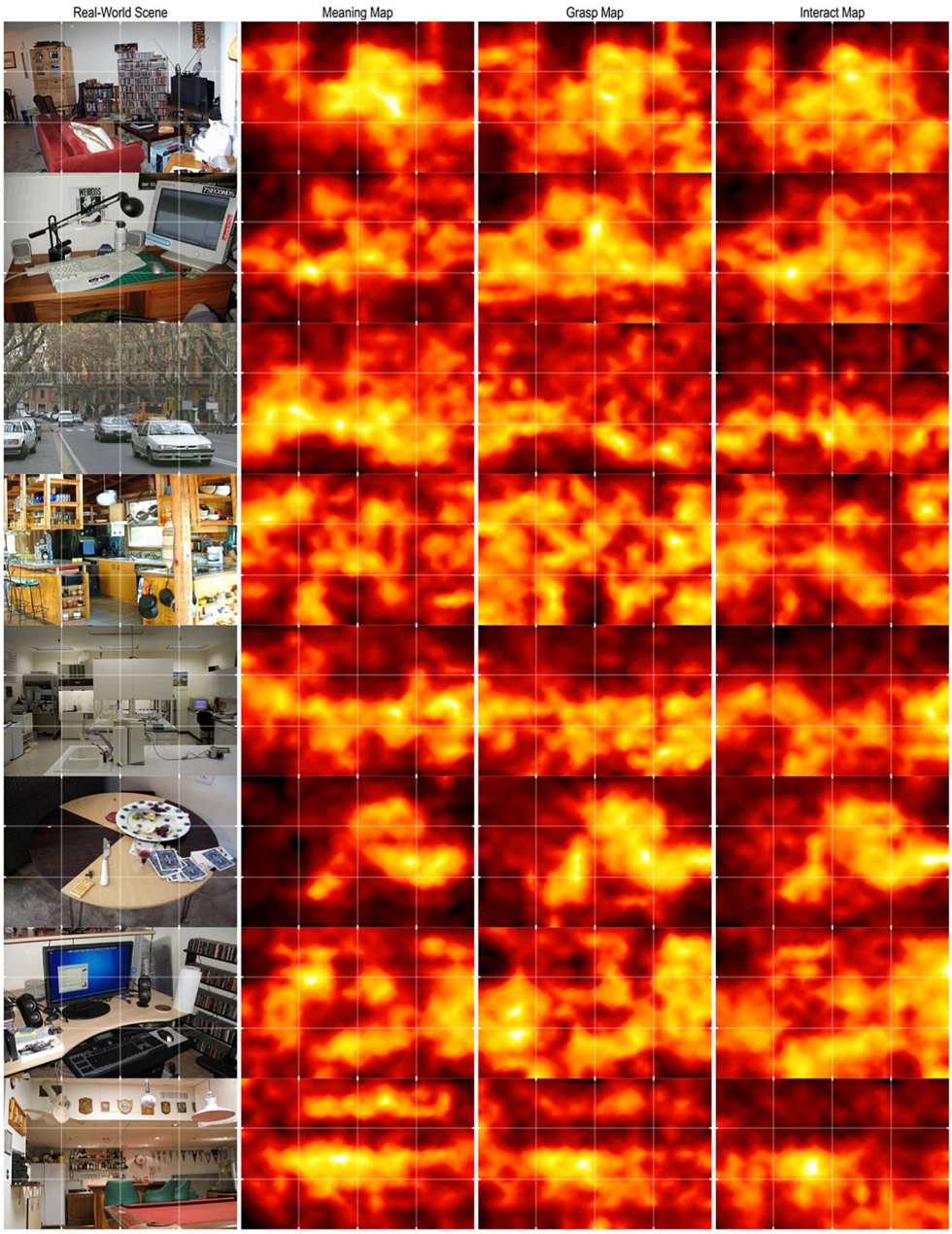
The current analysis investigated what type of semantic information guides attention in a scene. We conducted a novel analysis on existing data sets (Henderson et al., 2018; Rehrig et al., 2020a, b) and determined which of three forms of semantic information best accounted for overt visual attention: (1) general informativeness, the informativeness or recognizability of scene regions, (2) graspability, the degree to which what is shown in a region can be grasped, and (3) interactability, the degree to which a scene region depicts objects that can be interacted with in any way. Of the forms of semantic information tested, interactability was the strongest predictor of locations speakers fixated across three action description experiments, suggesting that the actions objects in a scene afford exert a strong influence on attention during action descriptions, moreso than what the results originally reported in Rehrig et al. (2020b) suggested. When speakers described scenes however they liked (Henderson et al., 2018), both interactability and informativeness predicted fixated locations; however, only informativeness predicted fixated locations when the task had no explicit language component (scene memorization; Rehrig et al., 2020a). Consistent with Altmann and Kamide (2007), the results suggest that object affordances guide attention when the language system is engaged—to a greater degree than informativeness does, at least when affordances are especially task-relevant (Experiments 1–3; consistent with Salverda et al., 2011)—while informativeness guides attention when the task does not encourage observers to carefully consider the objects in the scene, and, by extension, the interactions those objects afford. More generally, the finding that different semantic aspects of a scene influence the allocation of visual attention differently depending on the viewer's task offers additional, compelling evidence for the cognitive guidance theory of eye movement control (Henderson et al., 2007).

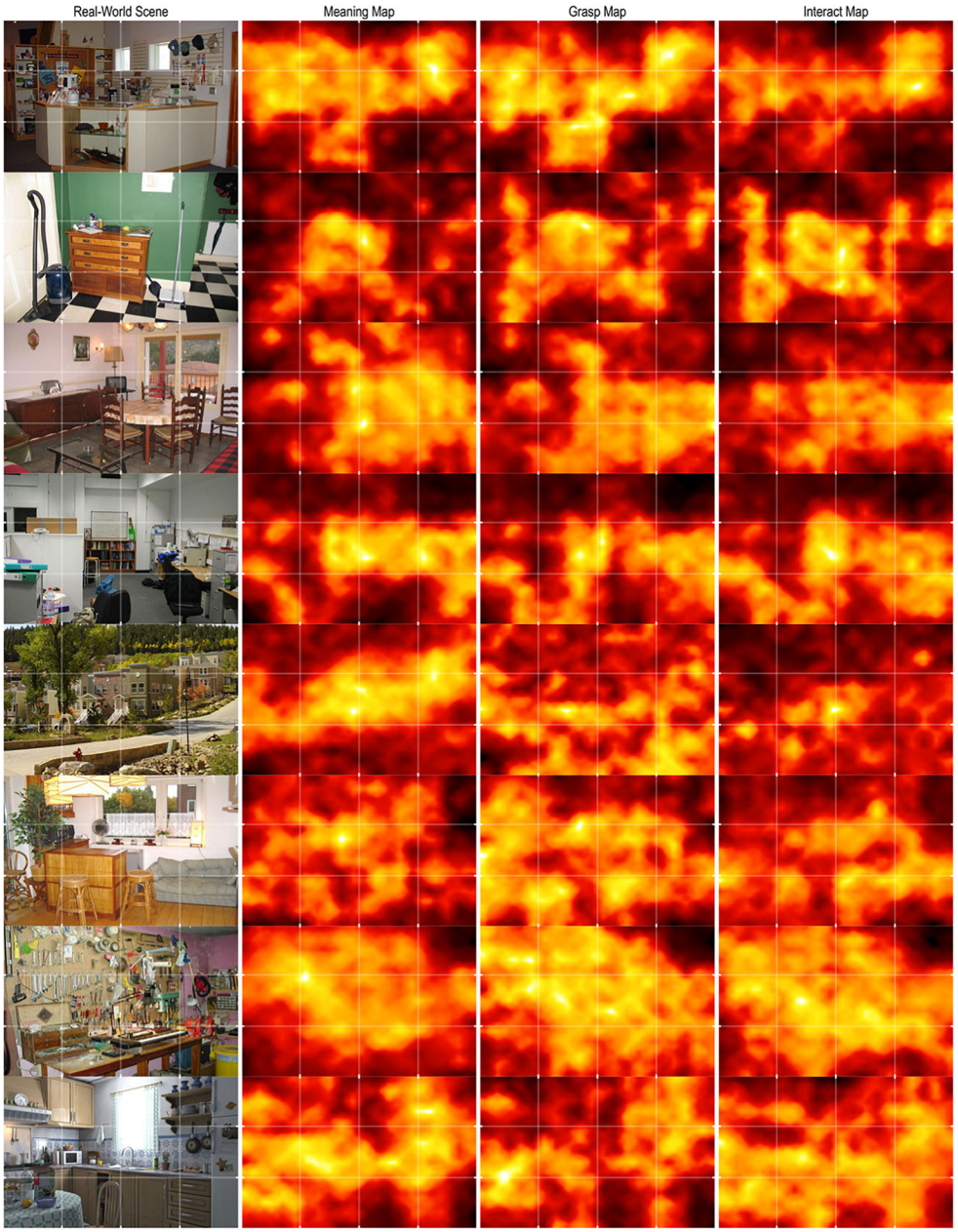
## Appendix

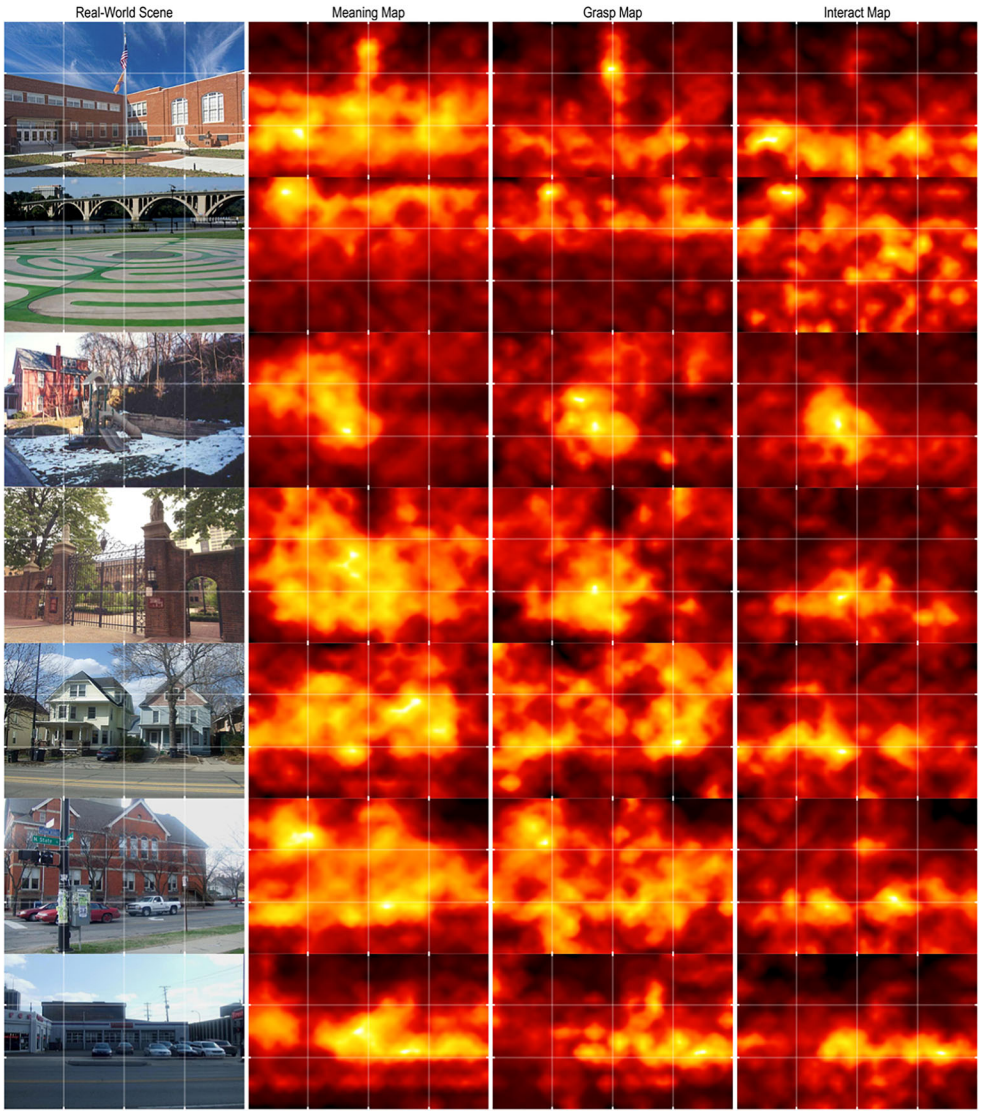
Scenes and feature maps unique to each experiment.

**Experiment 1, 4, & 5 Scenes**

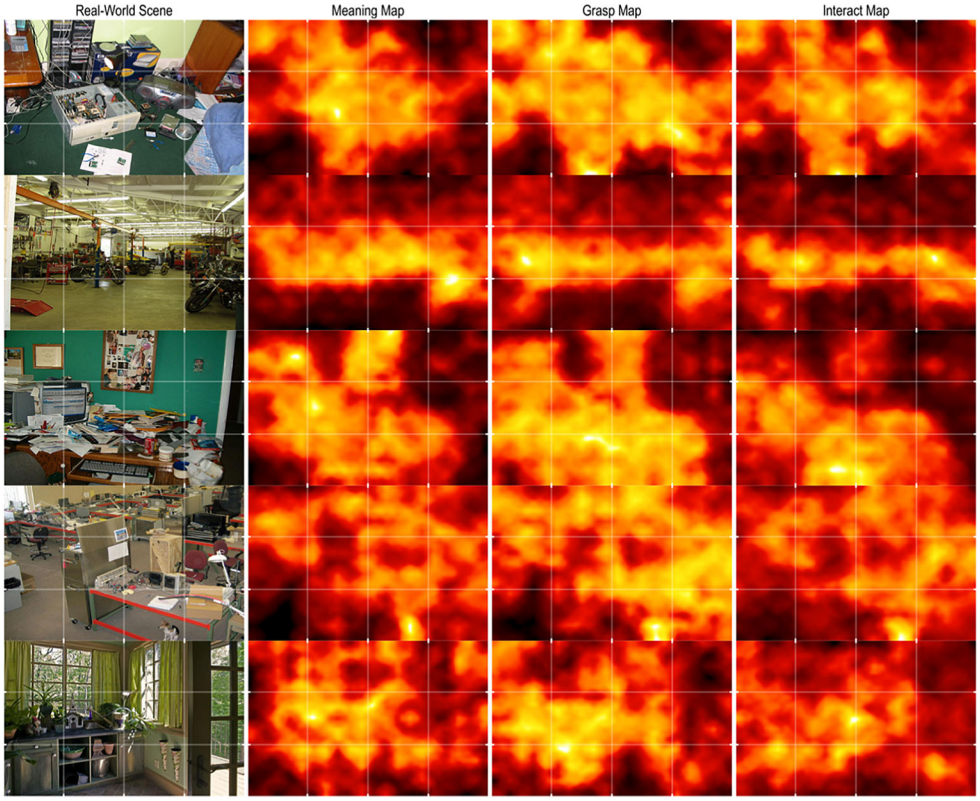



Real-World Scene
Meaning Map
Grasp Map
Interact Map

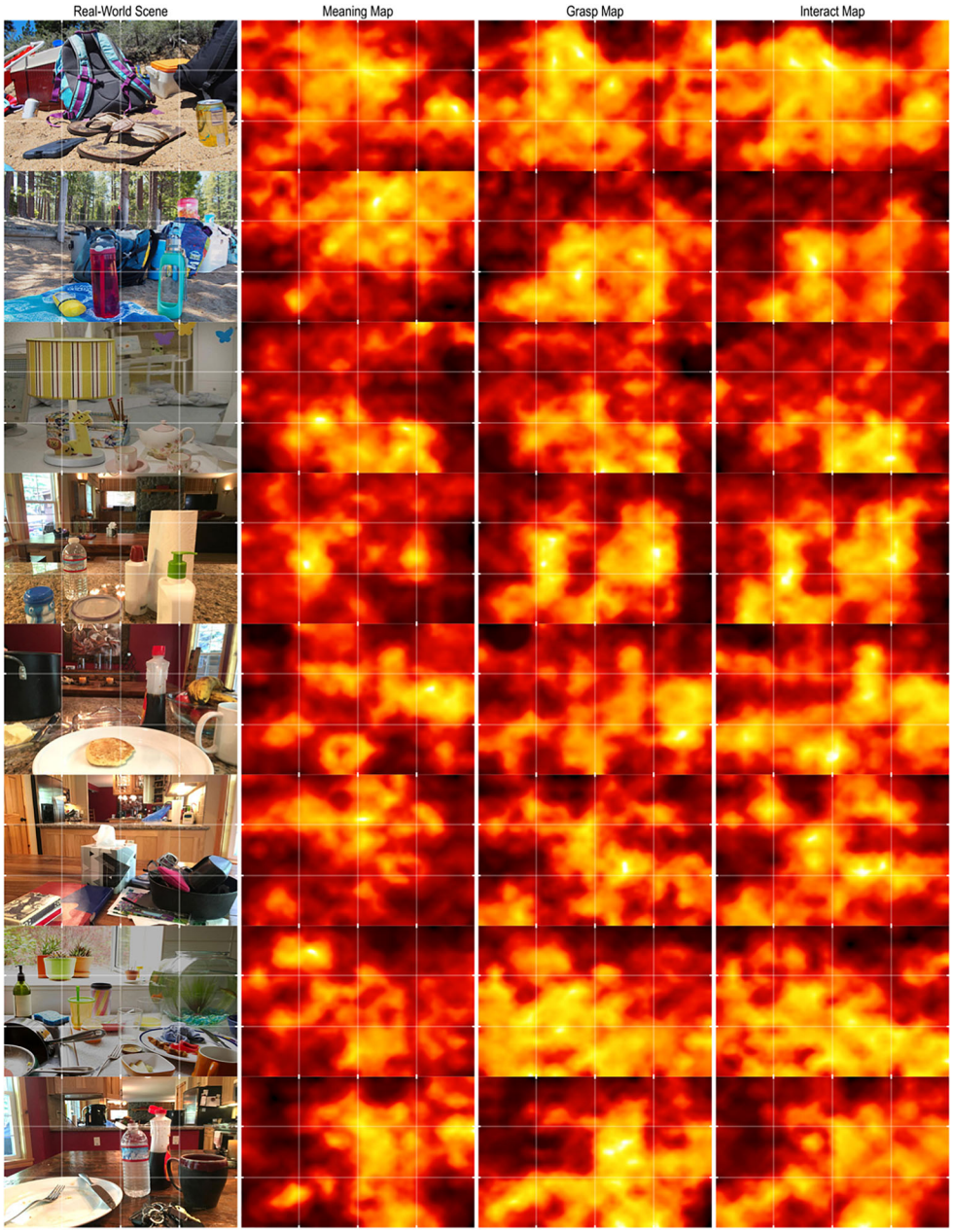

Real-World Scene
Meaning Map
Grasp Map
Interact Map

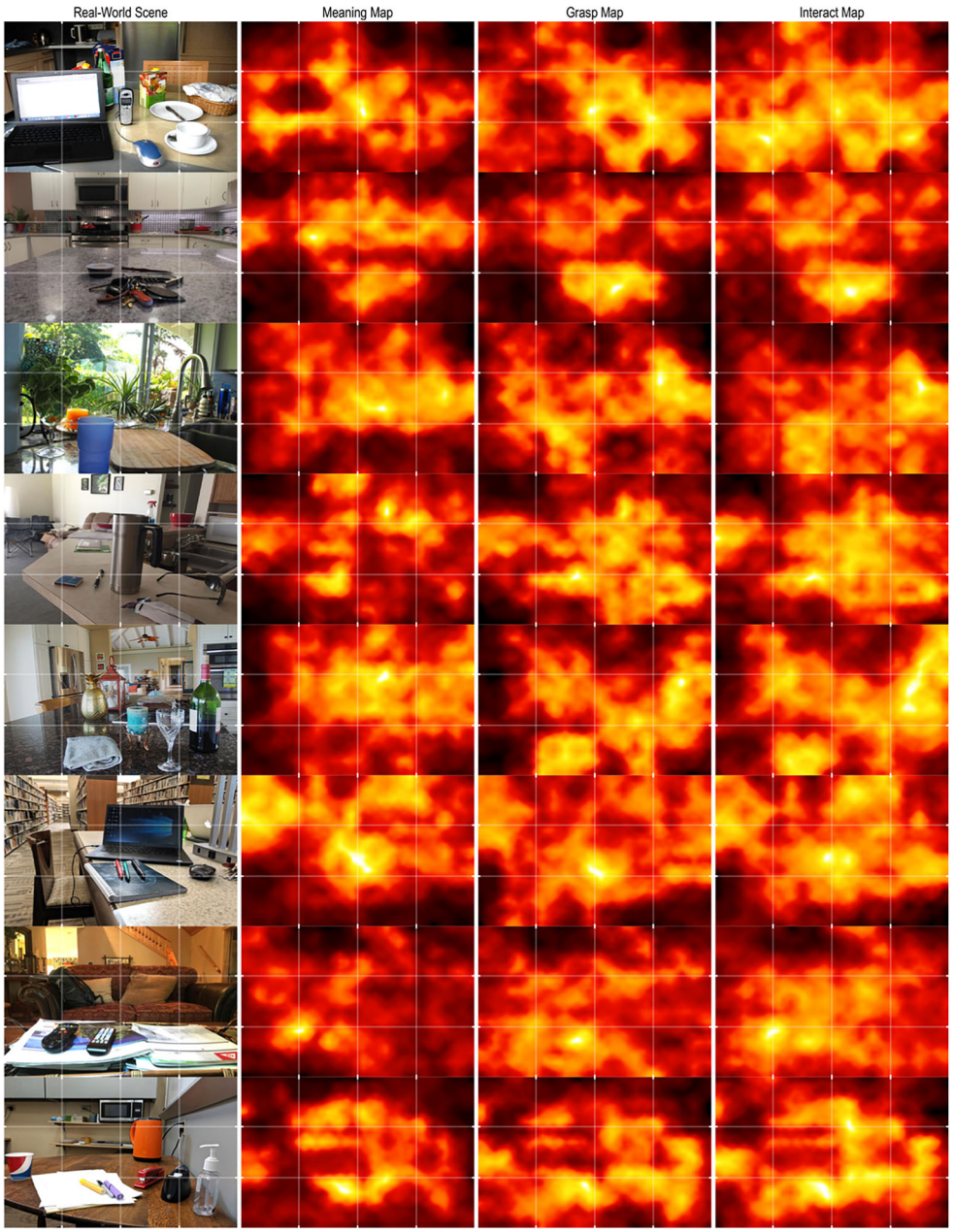

Real-World Scene
Meaning Map
Grasp Map
Interact Map

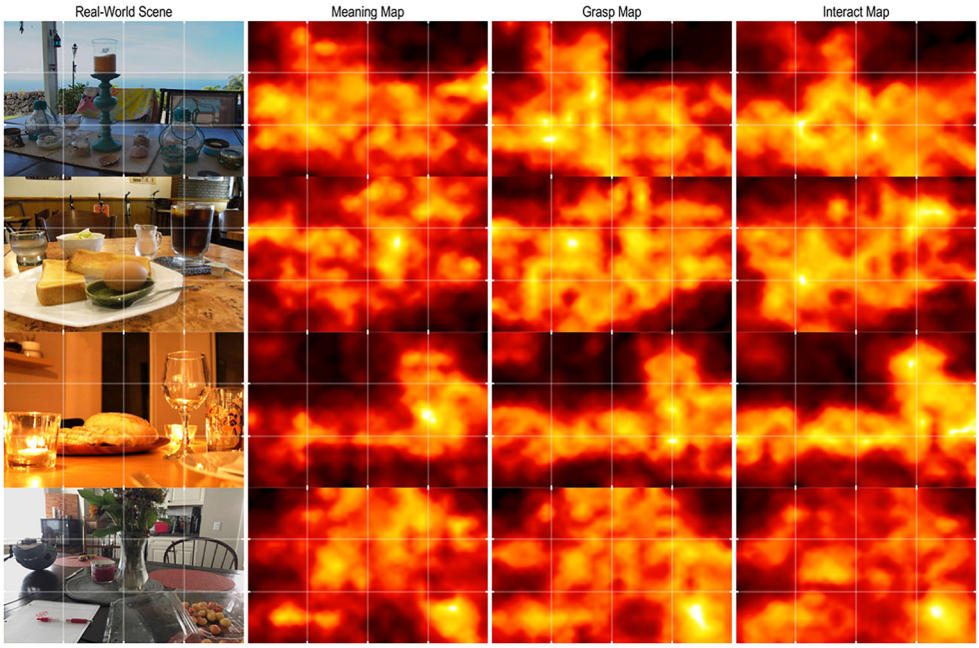
## Experiment 2 Scenes



## Experiment 3 Scenes




Real-World Scene
Meaning Map
Grasp Map
Interact Map




**Acknowledgements** This work was funded by the National Institute of Health grants R01 HD100516 awarded to Fernanda Ferreira, and R01 EY027792 awarded to John Henderson, and by National Science Foundation grants BCS1650888 awarded to Fernanda Ferreira and BCS2019445 awarded to John Henderson. We thank our research assistants who collected the data for their fantastic work.

**Open Practices Statement** The experiment and analyses reported here were not pre-registered. Data available on request.

## References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Borghi, A. M. (2012). *Language and action in cognitive neuroscience* (Chap. Language comprehension: Action, affordances and goals, pp. 143–162). Psychology Press.

Borghi, A. M., & Riggio, L. (2009). Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research*, *1253*, 117–128.

Castelhano, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science*, *27*(5), 606–621.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 687.

David-John, B., Peacock, C. E., Zhang, T., Murdison, T. S., Benko, H., & Jonker, T. R. (2021). Towards gaze-based prediction of the intent to interact in virtual reality. *Virtual Reality*, *7*.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18–18.

Feven-Parsons, I. M., & Goslin, J. (2018). Electrophysiological study of action-affordance priming between object names. *Brain and Language*, *184*, 20–31.

Glenberg, A. M., Becker, R., Klötzer, S., Kolanko, L., Müller, S., & Rinck, M. (2009). Episodic affordances contribute to language comprehension. *Language and Cognition*, *1*(1), 113–135.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565.

Gomez, M. A., Skiba, R. M., & Snow, J. C. (2018). Graspable objects grab attention more than images do. *Psychological Science*, *29*(2), 206–218.

Gomez, M. A., & Snow, J. C. (2017). Action properties of object images facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1115.

Grafton, S. T., Fadiga, L., Arbib, M. A., & Rizzolatti, G. (1997). Premotor cortex activation during observation and naming of familiar tools. *Neuroimage*, *6*(4), 231–236.

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

Harpaintner, M., Sim, E.-J., Trumpp, N. M., Ulrich, M., & Kiefer, M. (2020). The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex*, *124*, 1–22.

Hayes, T. R., & Henderson, J. M. (2019a). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. Attention, Perception, & Psychophysics, pp. 1–10.

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

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, 1–7. https://doi.org/10.1177/0956797621994768.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009.

Hayhoe, M., & Matthis, J. S. (2018). Control of gaze in natural environments: Effects of rewards and costs, uncertainty and memory in target selection. *Interface Focus*, *8*(4), 20180009.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences*, *21*(1), 15–23.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye Movements*, (pp. 537–III): Elsevier.

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

Henderson, J. M., Hayes, T., Peacock, C., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kümmerer, Wallis, Bethge & Teufel (2021). *Cognition*, 104742.

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., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PLOS ONE*, *8*(5), 1–6. https://doi.org/10.1371/journal.pone.0064937.

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.

Josephs, E. L., & Konkle, T. (2020). Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, *117*(47), 29354–29362. https://doi.org/10.1073/pnas.1912333117.

Kako, E., & Trueswell, J. C. (2000). Verb meanings, object affordances, and the incremental restrictions of reference. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 22).

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language*, *43*(3), 508–529.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*, 547–552.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20–20.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, *11*, 491.

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science*, *28*(6), 593–599.

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*(1), 20–34.

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, *41*(25–26), 3587–3596.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramey, M. M., Yonelinas, A. P., & Henderson, J. M. (2020). Why do we retrace our visual steps? Semantic and episodic memory in gaze reinstatement. *Learning & Memory*, *27*(7), 275–283.

Rehrig, G., Cullimore, R. A., Henderson, J. M., & Ferreira, F. (2021). When more is more: Redundant modifiers can facilitate visual search. *Cognitive Research: Principles and Implications*, *6*, 10.

Rehrig, G., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020a). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. *Memory & Cognition*, *48*, 1181–1195.

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020b). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *46*(9), 1659–1681.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, *137*(2), 172–180.

Shomstein, S., Malcolm, G. L., & Nah, J. C. (2019). Intrusive effects of task-irrelevant information on visual selective attention: Semantics and size. *Current Opinion in Psychology*, *29*, 153–159. https://doi.org/10.1016/j.copsyc.2019.02.008.

Sullivan, B., Ludwig, C. J. H., Damen, D., Mayol-Cuevas, W., & Gilchrist, I. D. (2021). Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *Journal of Vision*, *21*(3), 13. https://doi.org/10.1167/jov.21.3.13.

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.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), 28.