



Disrupted object-scene semantics boost scene recall but diminish object recall in drawings from memory

Wilma A. Bainbridge^{1,2} · Wan Y. Kwok^{2,3} · Chris I. Baker²

Accepted: 5 April 2021 / Published online: 24 May 2021

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

Abstract

Humans are highly sensitive to the statistical relationships between features and objects within visual scenes. Inconsistent objects within scenes (e.g., a mailbox in a bedroom) instantly jump out to us and are known to catch our attention. However, it is debated whether such semantic inconsistencies result in boosted memory for the scene, impaired memory, or have no influence on memory. Here, we examined the relationship of scene–object consistencies on memory representations measured through drawings made during recall. Participants ($N = 30$) were eye-tracked while studying 12 real-world scene images with an added object that was either semantically consistent or inconsistent. After a 6-minute distractor task, they drew the scenes from memory while pen movements were tracked electronically. Online scorers ($N = 1,725$) rated each drawing for diagnosticity, object detail, spatial detail, and memory errors. Inconsistent scenes were recalled more frequently, but contained less object detail. Further, inconsistent objects elicited more errors reflecting looser memory binding (e.g., migration across images). These results point to a dual effect in memory of boosted global (scene) but diminished local (object) information. Finally, we observed that participants fixate longest on inconsistent objects, but these fixations during study were not correlated with recall performance, time, or drawing order. In sum, these results show a nuanced effect of scene inconsistencies on memory detail during recall.

Keywords Saliency · Binding errors · Global scene processing · Local scene processing

When we view a scene, we automatically parse many aspects of that scene—its overall gist, constituent objects, and their relations to each other and the greater scene layout (Fei-Fei et al., 2007; Oliva & Torralba, 2006). In exploring and understanding that scene, we are guided not only by its visually salient aspects (Zhao & Koch, 2013) but also by its interpretations—or meaning (Henderson & Hayes, 2017). Unsurprisingly, we parse a scene based on our expectations of that scene and its objects, utilizing what can be considered a “scene grammar” (Võ et al., 2019) that guides what types of

objects go in what types of scenes. Thus, when we see scenes containing violations of this grammar—for example, when a beach ball is unexpectedly in a laboratory—they catch our attention. Such inconsistencies in object–scene semantics cause disruptions in our ability to process these images (Greene et al., 2015), and we tend to fixate on these inconsistencies during perceptual and visual search tasks (Henderson et al., 1999; Loftus & Mackworth, 1978; De Graef et al., 1990; Malcolm & Henderson, 2010), even when performing an irrelevant task (Cornelissen & Võ, 2017).

However, even though observers fixate longer on inconsistent objects, it is unclear how scene consistency influences the later memory for that scene and its objects. Some research has reported no memory differences between inconsistent and consistent objects in scenes when encoded incidentally (Cornelissen & Võ, 2017). Other work has reported boosted recognition and recall memory for inconsistent objects across both incidental and intentional memory tasks, along different time scales (Friedman, 1979; Hollingworth et al., 2001; Pezdek et al., 1989). Yet other work studying memory for object–location associations has reported an opposite effect of boosted memory for consistent objects (Draschkow & Võ, 2017). Given the varied results and divergent methods

Wilma A. Bainbridge and Wan Y. Kwok contributed equally to this work.

✉ Wilma A. Bainbridge
wilma@uchicago.edu

¹ Department of Psychology, University of Chicago, 5848 South University Ave, 303 Beecher Hall, Chicago, IL 60637, USA

² Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, MD 20814, USA

³ University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA

across these various studies, there is still a large open question of how object–scene semantics shape the memory representations for a scene. More broadly, beyond asking *whether* memory is impacted by scene semantics, little work has specifically asked *how* that memory is impacted: what specific aspects of a scene’s memory are altered—memory for the entire image, the manipulated object, or the background scene?

In the current study, we compare the underlying visual memory representations for inconsistent and consistent scenes using a visual recall drawing task. Previous work assessing the influences of scene semantics on memory have relied on verbal recall or visual recognition tasks (Cornelissen & Võ, 2017; Friedman, 1979; Hollingworth & Henderson, 1998; Pezdek et al., 1989). However, these types of measures may provide limited information about the nature of a memory—only revealing whether an item is remembered or not, but not what specific visual content of that memory drives its recollection. Recent work has discovered that using drawing-based recall to quantify memory can reveal more fine-grained information than verbal recollection, and requires no assumptions about matched foil images, often necessary for visual recognition tasks (Bainbridge et al., 2019). Drawings have also been known to reveal valuable insight about the memories of children (e.g., Bruck et al., 2000; Otgaar et al., 2016), artists (e.g., Perdreau & Cavanagh, 2015; Vogt & Magnussen, 2016), and patient groups (e.g., Corkin, 2002; Rey, 1941). Thus, a drawing task may reveal subtler differences between memories for consistent and inconsistent scenes than was possible to capture in previous work utilizing verbal-based or recognition-based tasks. Further, drawings can be objectively quantified through online crowd-sourced scoring to reveal a wide range of information, including object detail, spatial accuracy, and inclusion of false additional objects (Bainbridge et al., 2019). With such a task, we can thus examine not only whether an inconsistent object is remembered better than a consistent object, but how inconsistency impacts memory for other objects in the scenes, their spatial relations, and the scene overall.

With these measures, we examined several questions for how scene semantics might influence memory representations. First, we consider how consistency affects recall of the manipulated object in the scene. One possibility is that inconsistent objects are distinctive and easier to remember (Friedman, 1979; Hollingworth et al., 2001; Pezdek et al., 1989). Conversely, we might instead find that consistent objects better fit our scene schemas and thus are easier to remember, as has been found in work analyzing the role of consistency on scene construction (Draschkow & Võ, 2017). A third possibility is that we may observe no memory difference between inconsistent and consistent objects (Cornelissen & Võ, 2017). Second, we consider how consistency affects memory for the overarching image—do participants tend to draw inconsistent or consistent images more frequently from memory, regardless of their memory for the manipulated object?

Finally, beyond memory for the inconsistent/consistent object or its encompassing image, we ask whether there is a difference in memory for the other objects in the scene (what we will hereby refer to as the “background scene”). On one hand, the heightened distinctiveness of an image owing to the presence of an inconsistent object could boost memory for the entire scene, including surrounding objects. On the other hand, these inconsistent objects could create a “spotlight” effect, capturing attention away from surrounding objects (Cornelissen & Võ, 2017) and reducing recognition for objects semantically unrelated to the inconsistent object (Auckland et al., 2007; Davenport, 2007). With this spotlight effect, we might also observe transpositions of objects that are semantically unrelated to their overarching image (Hannigan & Reinitz, 2003). Thus, by analyzing drawings made from memory, we can examine memory performance for the image, the manipulated object, as well as the background scene. Further, using both eye-tracking and computer-vision-based saliency models during image encoding as well as pen-tracking during drawing recall, we can see whether we can replicate previous findings on increased fixations to inconsistent objects (e.g., Cornelissen & Võ, 2017), and whether fixation patterns during perception predict recall performance.

To preview our results, we find an interesting trade-off in memory, in which semantically inconsistent images are recalled more frequently, but with less detail, and with weaker binding between the inconsistent object and the scene, resulting in transpositions of that object across images. Further, while we replicate the observation that individuals fixate inconsistent objects during encoding, we find that recall patterns cannot be explained by fixation patterns or image saliency during encoding.

Methods

Participants

Thirty adults (9 males, 21 females; age $M = 24.8$ years, $SD = 4.5$) were recruited from the local Washington, DC, area for participation in this within-subjects experiment. This sample size was determined *a priori*, to match the same sample size collected in a previous, similar drawing-based experiment that measured high detail from memory drawings with only 15 participants drawing any given image (Bainbridge et al., 2019). The current study also includes far fewer images to hold in memory (12 vs. 30)—thus, we anticipate more drawings will be produced from memory per image, resulting in higher power per image than Bainbridge et al. (2019). However, as this drawing recall methodology is very new, we hope the current study will serve as a basis upon which to conduct power analyses for future drawing studies. Participants were healthy native English speakers with

corrected or normal vision, with the exception of participants with high-prescription glasses, who were not recruited, to avoid calibration issues with the head-mounted eye tracker. No participants or trials were excluded. All participants consented following the guidelines of the National Institutes of Health (NIH) Institutional Review Board (NCT00001360, 93M-0170) and were compensated for their participation.

A total of 1,725 online scorers were recruited from online crowd-sourcing task platform Amazon Mechanical Turk (AMT), acknowledging their participation following the guidelines of the NIH Office of Human Subjects Research Protections (OHSRP), and were also compensated for their participation. The number of online scorers per task was selected to be identical to prior online scoring studies of drawings (Bainbridge et al., 2019; Bainbridge et al., 2021).

Stimuli

Stimulus images were created from 12 distinctive scene images from different scene categories, half indoor (bathroom, bedroom, classroom, kitchen, laboratory, laundry room) and half outdoor (campsite, construction site, neighborhood street, playground, swimming pool, backyard). The original object and scene images came from publicly available photographs on Google Images, found by searches of the scene category and object names. Adobe Photoshop was used to naturally add an object to each image (referred to throughout as the “manipulated object”) that was either consistent or inconsistent with the scene semantics (see Fig. 1). The scene images were paired, and these object manipulations were conducted within the pairs, so that the consistent object in a given image was also used as the inconsistent object in its paired image, and vice versa. For example, in the consistent condition, a lab scene contained a microscope and a pool scene contained a beach ball (see Fig. 1). In the inconsistent condition, the lab scene had a beach ball and the pool scene had a microscope. The consistent and inconsistent object were placed at the same size and in the same location within a given scene, and

shadowing and lighting were added to each object to integrate it naturally with the surrounding scene. This resulted in a set of 24 stimuli, comprising of a consistent and inconsistent version of each of the 12 scenes (and, similarly, each of 12 objects had a consistent image and an inconsistent image). To confirm that we successfully manipulated scene consistency, all images were rated online by Amazon Mechanical Turk (AMT) workers ($N = 15$ per image; $N = 67$ total) on a 5-point Likert Scale on how typical (“normal”) it was for that object to be in the scene ($1 = \textit{very abnormal}$, $5 = \textit{very normal}$). As expected, consistent objects were rated to be significantly more normal than inconsistent objects (Consistent: $M = 4.4$, $SD = 1.1$; Inconsistent: $M = 1.6$, $SD = 1.1$; Wilcoxon signed rank test: $Z = 2.20$, $p = .028$, effect size $r = .64$). During the main experiment, each participant saw 12 images (one of each background scene), with half consistent images and half inconsistent images. Which images were consistent or inconsistent was counterbalanced across participants, so each of the 24 images was seen by 15 participants, akin to Bainbridge et al. (2019).

All 24 stimuli were annotated with outlines for every object by the authors in advance of the experiments, using online tool LabelMe (Russell et al., 2008). These annotations allow us to create object-based online scoring experiments, and compare drawings to ground-truth information of object size and location. Objects were defined as nameable, separable, visually distinct items, larger than a 50-pixel diameter. Visually uniform object parts were not labeled (e.g., the leg of a chair), but detachable components were (e.g., windows on a house). While the manipulated object was intentionally inserted to be a key object in the foreground, each scene contained multiple other foreground and background objects ($M = 39.3$ objects, $SD = 20.5$, $\text{Min} = 13$, $\text{Max} = 77$).

Experimental procedures

At the beginning of the experiment, participants were told to carefully examine each image as they would be later tested on



Fig. 1 Two example sets of consistent and inconsistent scenes. (Left) A toy car or mop bucket in a bathroom or playground. (Right) A beach ball or a microscope on a swimming pool deck or in a laboratory

their memory. Participants were informed that this was a memory task, so that they were motivated to actively encode the image details for long-term memory. However, participants were unaware of the nature of the memory task, and were unlikely to expect a drawing task, so they could not employ strategies specifically honed for drawing as a task. The experiment was split into four phases (see Fig. 2).

The first phase was a study phase, in which participants viewed each of the images for 10 seconds while their eye movements were tracked with a head-mounted EyeLink 1000 Plus eye-tracking device. Participants studied 12 images in total, determined as the average number of scenes recalled by participants in a prior memory drawing study ($M = 12.1$ images; Bainbridge et al., 2019). We anticipated that twelve images per participant would maximize the power of our study; more images would likely result in a low recall rate for any given image, while fewer images could be too easy and reduce the experimental power. Between the presentation of each stimulus image, a fixation cross was displayed to the right of the image on the screen, in order to avoid biasing eye movements to the center. After the participant fixated on the cross, the next stimulus was then displayed. Each image was displayed at $1,200 \times 800$ pixels on a 24-inch screen at 28 degrees of visual angle.

The second phase was a digit span distractor task intended to disrupt verbal working memory strategies. Participants saw a consecutive series of digits varying by 3–9 digits in length, and then had to repeat back the series of digits from memory when prompted. This repeated for 21 trials, and introduced an approximately 6-minute delay between the study and test phases.

The third phase was the drawing recall test phase. Participants were given sheets of paper with a rectangular outline with dimensions matching those of the original images, and were asked to draw as many images as they could remember in as much detail as possible. Participants drew on a Wacom Paper Pro tablet, which allowed participants to draw with an inked pen

on paper while it simultaneously recorded pen strokes digitally in real time. Participants were told to draw the images in any order. They were also given colored pencils if they wanted to include color detail in their drawings, but were asked to include color only if they specifically recalled it. They were instructed to add color after completing the pen drawing of all the objects they recalled. Participants were also told they could label objects if they wished to clarify what they were. Participants were given as much time as they needed and took 27 minutes on average for the recall phase ($SD = 8$).

In the fourth and final phase, participants completed a recognition phase, in which they made a series of recognition-based judgments of the images. They were shown the 12 scene images they studied randomly interspersed with 12 closely matched foil scenes of the same scene categories. All scene images had a gray occluding ellipse covering the manipulated (consistent or inconsistent) object. Foil images had a gray occluding ellipse placed in a plausible similar location. First, for a given image, participants were asked if they had seen the scene during the study phase (scene recognition). If they said yes, they were presented with four object images and had to indicate which was the object they saw in that scene. The four choices of object images were: (1) the inconsistent object, (2) the consistent object, (3) a different exemplar from the inconsistent object category, and (4) a different exemplar from the consistent object category. This question tested both object category recognition (e.g., if you studied an inconsistent scene, did you falsely remember seeing a consistent object?), as well as specificity to the exemplar within the same category (e.g., did you remember that you saw that specific microscope in the scene, or a microscope in general?).

Online scoring procedures

The resulting 275 drawings were scanned and uploaded to AMT, to crowd-source worker ratings on several properties of the memory drawings. Specifically, four different

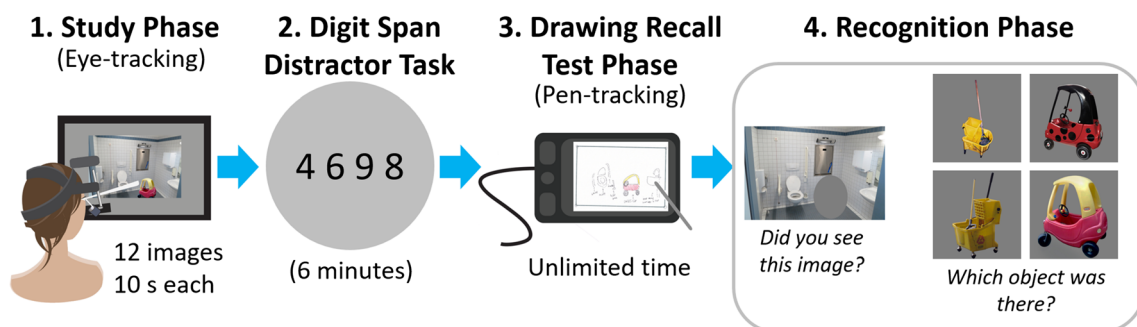


Fig. 2 The experimental procedures. The experiment consisted of four phases: (1) A study phase in which participants studied 12 randomly ordered images (6 consistent, 6 inconsistent) for 10 s each while their fixations were tracked; (2) a digit span distractor task in which participants had to verbally recall digit series; (3) a drawing recall test phase in

which participants drew the studied scenes from memory while their pen movements were tracked; and (4) a recognition phase in which participants had to separately recognize the scene and the manipulated object

sets of measures were collected for each drawing. For all rating tasks, each AMT worker could participate in as many trials as they wanted. Depictions of the four online tasks can be seen in Fig. 3, with the precise instructions given to AMT workers. AMT workers did not know the origins of these drawings, the different image conditions (i.e., consistent vs. inconsistent scenes), nor the nature of the main drawing experiment. Drawings were also randomly mixed, so that if an AMT worker participated on multiple trials, they would not know if they were scoring drawings from the same (or different) person or conditions. Thus, AMT workers scored these drawings blind to the conditions.

Drawing match scoring AMT workers rated how well each drawing matched one of the images participants studied, providing a measure of diagnosticity of that drawing (see Fig. 3a, top left). For a given trial, they were presented with a drawing with an image next to it and rated on a 5-point Likert scale the likelihood that it was a drawing of that image (1= *definitely not*, 5 = *definitely*). Across trials, drawings were tested against each of the 12 images seen by the participant who made the drawings. Twelve ratings were collected for each drawing–image pair (with a total of 144 ratings per drawing across the 12 pairs), and 611 AMT workers participated in total. The image with the highest match rating with a drawing (averaged across the 12 AMT workers for that drawing–image

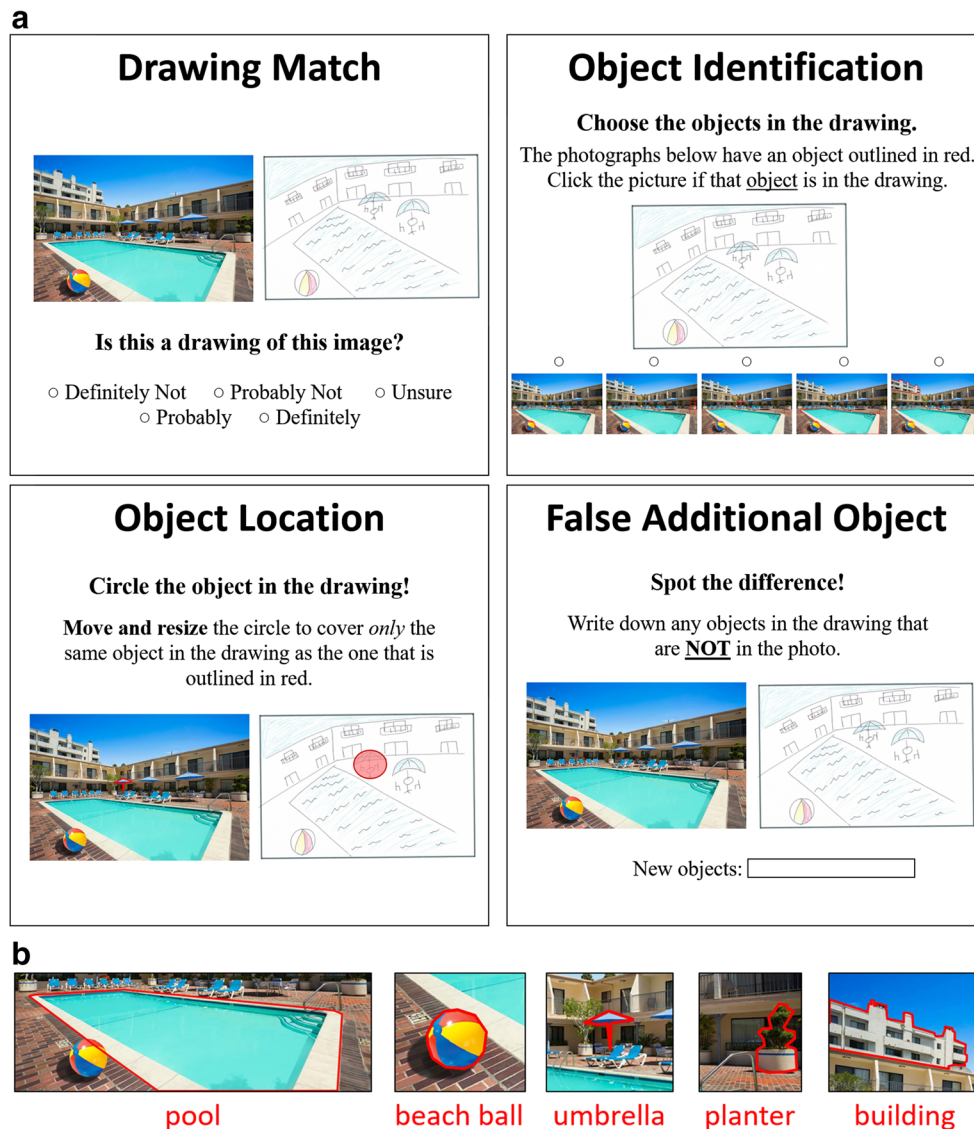


Fig. 3 a Example trials from the online experiments. Shown are depictions of example trials from the four online experiments: (1) drawing match scoring, (2) object identification scoring, (3) object location scoring, and (4) false additional object scoring. **b** Up-close view of highlighted objects. The object identification scoring and object location scoring

tasks required online participants to compare a highlighted object with the drawings. Shown here are the close-up examples of objects highlighted in red from “object identification” (Panel a). Outlines were created *a priori* using LabelMe (see Methods). Online participants saw the full image when making responses (as in Panel a)

pair) was selected as the corresponding image (“original image”) for that drawing. To score overall image recall performance, each of the 12 images each participant saw was given a binary score (1 = was drawn, 0 = was not drawn) as determined by if a drawing was matched to it by the AMT scorers. Number of recalled images was calculated as the sum of those 12 binary scores. If a participant made multiple drawings of the same image, that image was only given a single score for being remembered. However, these duplicate drawings were scored for other measures (below) and their objects present were taken as the union across duplicates within participants (rather than us selecting a single drawing to be scored for a given image). Seventeen participants drew multiple drawings of the same image.

Object identification scoring For each drawing-image pair resulting as the highest match from the drawing match scoring, AMT workers determined which objects from that image were included in each drawing (see Fig. 3a, top right). For a given trial, they were presented with a drawing and five copies of the matched image with a different object highlighted on it. Objects were highlighted to AMT workers with a red outline, determined by the LabelMe outline created *a priori* (see Fig. 3b, see Stimuli). AMT workers had to click on which of those five objects (if any) were present in the drawing. Five AMT workers rated each object, and 679 AMT workers participated in total. Using these object outlines rather than object names of the objects allows AMT workers to decide on the presence of an object using object identity, detail, size, and spatial information—so that presence can be determined when there are multiple exemplars for a given object type (e.g., the multiple umbrellas in Fig. 3a). The five objects shown to any given AMT worker were randomly selected and counterbalanced, so that across AMT workers, five ratings were collected for every single object from each image. Objects were determined to be in the drawing if at least three out of five workers said it was in the drawing. In analyses comparing object recall for consistent versus inconsistent scenes, one participant was excluded because they did not draw any consistent scenes. Participant recall performance for an image was measured as the number of objects they drew for a given image, divided the total number of objects present in that image.

Object location and size scoring AMT workers determined the locations and sizes of each object present in the drawings (see Fig. 3a, bottom left). For a given trial, they were presented with a drawing and its matched image with an object highlighted on it. On the drawing, they had to place and resize an ellipse to encircle that specified object. Five AMT workers made ellipses for each object and 453 AMT workers participated in total. The final ellipse was determined by the median centroid and radii in the x and y directions. This scoring was

conducted for all objects determined to exist in a given drawing, based on ratings in the object identification scoring. One participant did not draw any consistent scenes, and so they were not included in analyses comparing locations of objects in consistent versus inconsistent scenes. One participant also did not draw any inconsistent manipulated objects, and so an analysis comparing the location and size of manipulated objects only included 28 of the 30 total participants.

False additional object scoring AMT workers determined the presence of additional objects in the drawings that were not in the original images (see Fig. 3a, bottom right). For a given trial, they were presented with a drawing and its corresponding image and had to write down all objects that existed in the drawing, but not the image. Fifteen AMT workers rated each image, and 200 AMT workers participated in total. Any objects listed by at least five workers were counted as false alarms.

Fixation, pen-tracking, and saliency analyses

From the EyeLink 1000 Plus, we extracted eye-movement patterns for each participant to each image, as a list of locations on the image and their fixation times. To obtain a metric of fixation time per object per participant, we computed the total fixation time across all pixels within a given object. We also looked at fixation order by object, by comparing the order in which the manipulated object had its first fixation in relation to the first fixation on all other objects (e.g., of all objects, was the manipulated object fixated first, second, etc?). A participant’s fixation order was then normalized by total number of objects in the drawing.

Using the tablet recordings of the pen movements, we also calculated amount of time spent drawing each manipulated object per participant. An in-lab scorer watched the video of pen strokes created by the drawing tablet for each drawing. The start and end time of pen strokes for the manipulated objects were noted for each image, for the first span of time in which the object was drawn. Time spent coloring objects was not included, as participants were instructed to add color after completing their drawing (and the tablet could not track colored pens/pencils). Object drawing time also did not include any time spent returning to add details to an object later. Total amount of time spent drawing the object was calculated as the difference between the end time and the start time, normalized by total amount of time spent on the drawing. Similarly, we calculated sequential drawing order per participant by assigning an order to each object based on first pen stroke on that object. Drawing order was then normalized by total number of objects in the drawing. One participant was removed from the drawing time and drawing order analyses due to a technical glitch with the pen tablet software (resulting in $N = 29$ for these analyses).

To compute image saliency scores, we used two state-of-the-art computer vision algorithms designed to predict human fixation time: DeepGaze II (Kümmerer et al., 2016) and Graph-Based Visual Saliency (Harel et al., 2007). Both models aim to predict human fixations of an image, but DeepGaze II is a more recent approach that utilizes a wide range of feature types (i.e., both low-level and high-level visual information) and is trained on human fixation data, while Graph-Based Visual Saliency (GBVS) is a more established method commonly tested by attention researchers, that relies solely on image-computable low-level visual features. Specifically, DeepGaze II predicts fixation time based on features from the VGG-19 deep neural network for object identification combined with a readout network trained for saliency prediction based on human fixations (Kümmerer et al., 2016). In contrast, GBVS is a model that identifies visually dissimilar regions of an image (Harel et al., 2007). We were curious to see whether these two models would perform differently in their predictions of recall and fixation behavior, given current debates on the success of these models in predicting scene semantics (Hayes & Henderson, 2019; Henderson et al., 2021; Pedziwiatr et al., 2021). For both metrics, we obtained saliency heatmaps for each of the stimulus images (see Fig. 8). Object-based saliency was then calculated as the average saliency across the pixels of that object, normalized by the average saliency of the entire image.

Finally, to generate heatmaps of recall for each image, we calculated a recall score for each object, calculated as the number of participants who drew that object, divided by the number of participants who drew the image containing that object. This allows us to create a heatmap of how well different objects in an image were remembered, that can be directly compared with heatmaps formed from fixation patterns, pen movements, or saliency measures.

We thus have multiple values for each object in a given image: (1) fixation time on that object (averaged or by participant); (2) average time spent drawing that object (averaged or by participant); (3) fixation order on that object (by participant); (4) drawing order of that object (by participant); (5) an average GBVS saliency score of the object; (6) an average DeepGaze II saliency score of the object; and (7) proportion of participants recalling that object. Analyses were conducted at the levels of these different object scores, not on the heatmaps themselves.

Data analyses

For most analyses, we conducted paired samples t-tests within subjects to compare the above metrics between participants' drawings of consistent scenes versus inconsistent scenes. We first tested these metrics for normality using a Kolmogorov–Smirnov goodness-of-fit test, and found none of these were significantly different from a normal distribution (all $ps > .05$).

For metrics with limited ranges (e.g., Likert scales, number of drawings made), we instead conducted nonparametric paired samples Wilcoxon signed rank tests. Effect sizes are included with all significant statistical tests.

Results

Drawings are highly diagnostic of their images

The first step is identifying the images that correspond to each drawing. Further, given the range of people's drawing abilities and memory, can a separate group of participants tell what image a drawing represents? AMT workers saw individual drawings matched with each of the 12 images studied by participants in the main experiment, and judged the likelihood that the drawing was of that image on a scale of 1 (*definitely not*) to 5 (*definitely*). Overall, it was clear to AMT workers what images matched the drawings, with only a single image getting a score above 3 on average (see Fig. 4). For all further analyses, the highest rated image was taken as the corresponding image for each drawing. Importantly, there was no significant difference in ratings between consistent and inconsistent images (Consistent: $M = 3.9$, $SD = 0.5$; Inconsistent: $M = 3.7$, $SD = 0.6$; Wilcoxon signed rank test: $Z = 1.70$, $p = .090$). This means that both semantically consistent and inconsistent

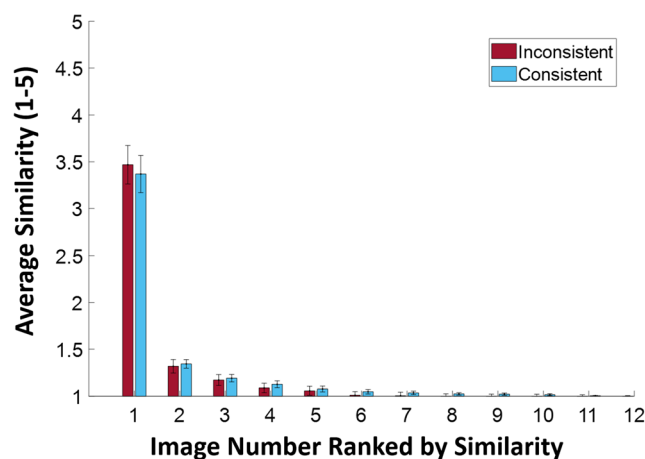


Fig. 4 Diagnosticity of the drawings across images for the two conditions. The average ratings made by online scorers of the similarity of participants' drawings to each of the 12 images they saw, ranked from highest to lowest. Similarity here was assessed on a scale of 1 (low) to 5 (high), with a question asking how likely it was that a drawing was of a given image (the "drawing match scoring" experiment). Red bars indicate drawings made of inconsistent images, while blue bars indicate drawings made of consistent images. The high spike for image #1 and quick drop-off for images #2–12 (all averaging below a rating of 1.5) for both conditions indicates that it was clear to AMT scorers that a given drawing was highly similar to only one image and dissimilar from all others. In other words, drawings were highly diagnostic of their images. There was no significant difference in diagnosticity between consistent and inconsistent images. Error bars indicate standard error of the mean

drawings could be matched with their original images, and were equally diagnostic of their original image. However, this rating of diagnosticity serves as a relatively coarse metric, as several different features could contribute to being able to successfully match a drawing to an image (i.e., the manipulated object, properties of the background scene). A closer look at the content in these drawings may reveal key differences between consistent and inconsistent images.

More inconsistent scenes are recalled than consistent scenes

The memory drawing experiment resulted in 275 total drawings, with 126 drawings of consistent images, and 149 drawings of inconsistent images (see Fig. 5). This reflects a general tendency to recall inconsistent images over consistent images (chi-squared test for proportions: $\chi^2 = 3.85, p = .050$, effect size $\varphi = 0.12$). Each participant on average drew 9.2 drawings from memory out of the 12 that they studied ($SD = 2.16$, Min = 5, Max = 12). Of those drawings, participants drew more inconsistent images than consistent images from memory (Wilcoxon signed rank test: $Z = 2.10, p = .036, r = .38$), drawing on average 5.0 inconsistent images ($SD = 1.5$) and 4.2 consistent images ($SD = 1.5$). Thus, memory for inconsistent images overall was better than that for consistent images.

More objects are recalled in consistent scenes than in inconsistent scenes

Next, we looked at the amount of detail available in each drawing by having AMT workers judge which objects from the original image were included in each drawing. Each image contained on average 39.3 objects (with the same number of objects across the consistent and inconsistent versions of a given scene), and participants drew on average 9.0 objects per image ($SD = 2.9$), or 77.6 objects on average across the experiment. Participants drew a significantly higher proportion of the objects in consistent drawings versus inconsistent drawings (Consistent: $M = 23.4\%, SD = 6.9\%$; Inconsistent: $M = 19.8\%, SD = 8.1\%$), paired t test, excluding the manipulated object: $t(28) = 2.56, p = .016, d = 0.52$. We then looked to see whether there were differences in the tendency to draw the manipulated object (the consistent or inconsistent object). We found no significant difference between consistent or inconsistent drawings in the proportion containing the manipulated object (Consistent: $M = 53.8\%, SD = 25.8\%$; Inconsistent: $M = 65.0\%, SD = 28.1\%$), $t(28) = 1.37, p = .181$. This indicates that while semantically inconsistent objects were recalled just as frequently as their consistent counterparts, there was reduced memory for objects in the inconsistent background scenes than in consistent ones.

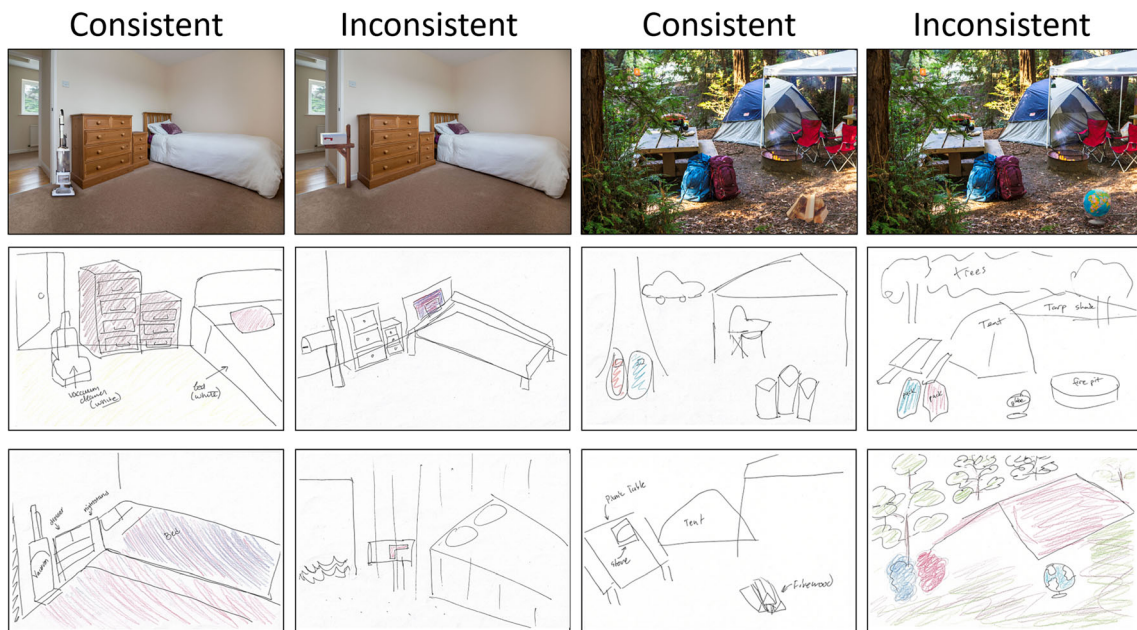


Fig. 5 Example drawings for four of the stimulus images. Shown here are two example drawings each for the consistent and inconsistent bedroom scene, and the consistent and inconsistent camp scene. Each drawing was taken from a different participant, showcasing the

impressive level of both object and spatial detail in the memory drawings for a range of people. The key question in this study is whether there are differences in memory detail between drawings for the consistent and the consistent scenes

The nature of object errors in consistent versus inconsistent scenes

We then investigated whether there were differences between conditions in the types of object errors that were made in the drawings. Overall, participants included relatively few false additional objects in their drawings, only drawing 25 objects that did not exist in the consistent images (across the 126 drawings from all participants), and 21 objects that did not exist in the inconsistent images (across 149 drawings). Within participants, there was no significant difference in the number of false additional objects they drew for inconsistent scenes versus consistent scenes, $t(29) = 0.64$, $p = .526$. Thus, differences in scene semantics did not appear to induce differences in false memories in these drawings.

However, participants made intriguing errors with the manipulated object in their drawings (see Fig. 6). In 18 drawings (13 inconsistent, 5 consistent), participants made drawings of only the manipulated object, unable to recall the surrounding background scene. In six drawings (6 inconsistent, 0 consistent), participants drew a detailed scene and included a circle with an unspecified object; they remembered that the manipulated object was there, but not what it was. Participants were not explicitly instructed to draw such “fuzzy” objects, so these occurred spontaneously by the participant. Finally, in 16 drawings (13 inconsistent, 3 consistent), participants transposed the manipulated object so that it was in a different scene they had viewed. All of these errors occurred significantly more frequently for inconsistent than consistent scenes (chi-squared test of proportions, Isolated Object: $\chi^2 = 7.11$, $p = .008$, $\varphi = 0.63$; Unspecified Object: $\chi^2 = 12.00$, $p = 5.32 \times 10^{-4}$, $\varphi = 1.41$; Transposed Object: $\chi^2 = 12.50$, $p = 4.07 \times 10^{-4}$,

$\varphi = 0.88$). These results imply that a disruption of scene semantics may result in a looser binding in memory of that inconsistent object with its encompassing scene.

Equally high spatial accuracy (location and size) in consistent and inconsistent images

While there are differences in object memory based on scene semantics, are there also differences in spatial accuracy for the objects in the drawings? AMT workers indicated the size and location of each object in the drawing by placing an ellipse on the drawn object. With that ellipse, we calculated mean location error (centroid x and y error) and size error (radius x and y error) for each object. In both conditions, low amounts of spatial error were found, although errors were larger in magnitude in the Y -direction than the X -direction. Errors of object location were transpositions of less than 11% of the size of the entire image (X -direction: Consistent $M = 2.2\%$, Inconsistent $M = 0.4\%$, Y -direction: Consistent $M = 9.3\%$, Inconsistent $M = 10.8\%$). Errors in size were on average less than 4% of an image’s pixels (Width: Consistent $M = 2.2\%$, Inconsistent $M = 2.0\%$; Height: Consistent $M = 2.9\%$, Inconsistent $M = 3.7\%$). For the manipulated object, there was no significant difference between consistent and inconsistent drawings in spatial accuracy, either in terms of location accuracy (X -direction), $t(27) = 0.94$, $p = .357$; (Y -direction), $t(27) = 1.24$, $p = .226$, or object size (Width), $t(27) = 0.51$, $p = .613$; (Height), $t(27) = 1.34$, $p = .191$. There were also no differences between conditions in accuracy for location or size of the other objects in the scene (X -location), $t(28) = 1.18$, $p = .249$; (Y -location), $t(28) = 0.93$, $p = .361$; (Width), $t(28) = 1.24$, $p = .227$; (Height), $t(28) = 1.43$, $p = .163$. These results indicate that

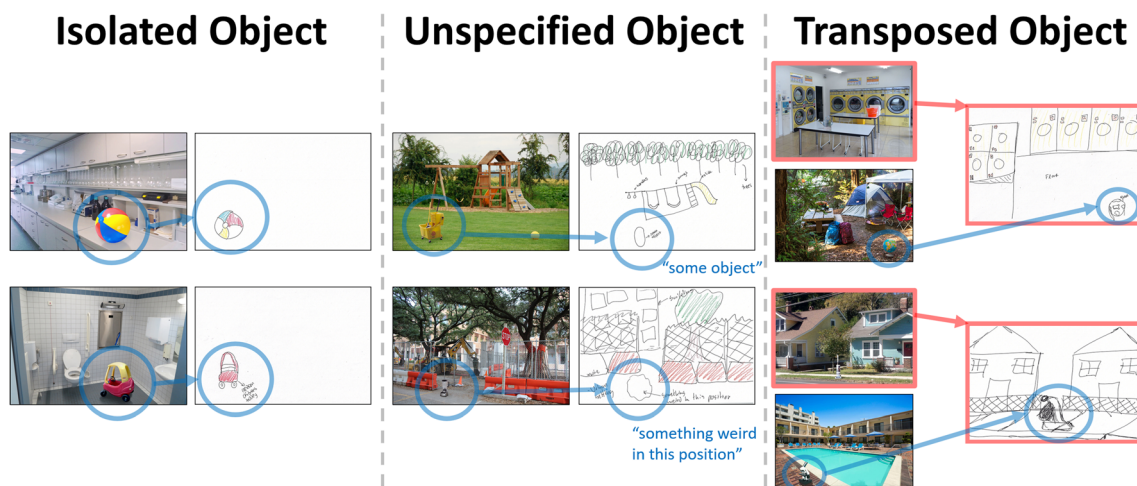


Fig. 6 Examples of memory errors made by participants for the manipulated object. Each example is taken from a different participant. We identified three types of errors: (1) drawing the object in isolation (left); (2) drawing a detailed scene with a circle noting recollection of an unspecified object at that location, but not its identity; and (3) transposing

the object to a different scene. For the unspecified object errors, the text labeling the circle is included in larger font. These errors occurred overwhelmingly more often when objects were inconsistent with their background scenes ($p < .01$ for all error types)

manipulations of object semantics do not appear to affect spatial accuracy in memory.

Comparing eye fixations, visual saliency, and recall

Prior studies have observed an influence of scene semantics on fixation time across the scene (De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978; Malcolm & Henderson, 2010). We examined whether there was a tendency for participants to fixate longer on the inconsistent objects in our paradigm. Further, we investigated whether fixation time and order (see Fig. 7) could be predicted by visual saliency of the image. Finally, we tested the degree to which these metrics related to recall of the information in the images.

We looked at fixation time for each participant for each object during the study phase, where they viewed each image for 10 s. On average, participants fixated for significantly longer on the inconsistent manipulated object than the consistent manipulated object (Inconsistent: $M = 1,271.7$ ms, $SD = 553.3$ ms; Consistent: 725.4 ms, $SD = 365.4$ ms), $t(26) = 5.44$, $p = 1.05 \times 10^{-5}$, $d = 1.17$; three participants were not measured as fixating on the manipulated object. For the other objects in the scenes, participants spent numerically more time looking at them in the consistent condition than the inconsistent condition, but this difference was not statistically significant (Inconsistent: $M = 6,439.8$ ms total across all other objects, $SD = 1,126.8$ ms; Consistent: $M = 6,956.5$ ms, $SD = 1,101.5$

ms), $t(29) = 1.97$, $p = 0.059$. There was also no difference in time spent fixating nonobject regions of the image (Inconsistent: $M = 1,592.2$ ms, $SD = 1,823.8$ ms; Consistent: $M = 1,592.2$ ms, $SD = 1,682.1$ ms), $t(29) = 0.16$, $p = .871$. Thus, there is no clear evidence that increased fixations on the inconsistent object detracted from fixations on other objects, preventing their encoding into memory.

We then looked to see whether current state-of-the-art visual saliency algorithms DeepGaze II and GBVS could predict these fixation times (see Methods). Collapsing across conditions, we found a significant correlation between DeepGaze-predicted saliency and fixation time on the manipulated objects (Spearman’s rank correlation: $\rho = 0.634$, $p = .001$) as well as the same correlation for GBVS-predicted saliency and fixation time ($\rho = 0.634$, $p = .001$). However, there was no significant difference in either saliency score measure between inconsistent and consistent objects in the same scene, DeepGaze: $t(11) = 0.48$, $p = .643$; GBVS: $t(11) = 1.42$, $p = .182$. These results indicate that visual saliency may be able to partially account for fixation durations, but it does not show a clear relationship to semantic consistency; we discuss these implications later in the Discussion.

Next, we investigated whether these metrics could predict the proportion of people who recalled the manipulated object (see Fig. 8). We observed no significant correlation between mean fixation time across participants and recall proportion for each manipulated object ($\rho = 0.22$, $p = .308$).

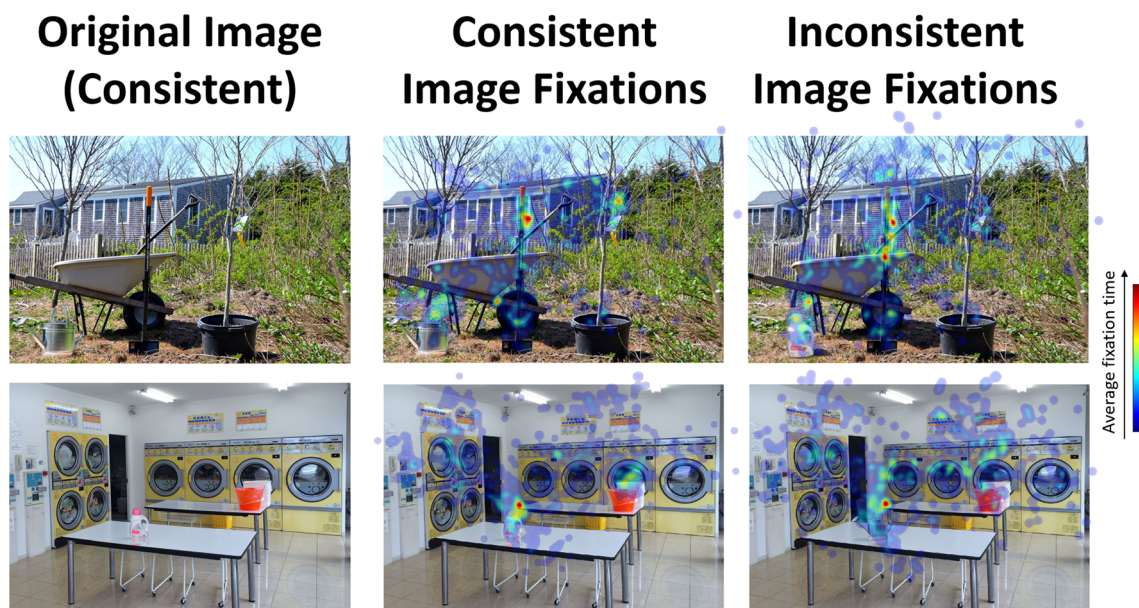


Fig. 7 Example average fixation heatmaps. Shown here are example heatmaps of fixation time (averaged across participants) for the consistent and inconsistent versions of two paired scenes where their objects were swapped (i.e., a watering can or laundry detergent in a backyard scene or a laundry scene). Red indicates higher total fixation time on average, and blue indicates lower total fixation time. For the

backyard scene, the inconsistent detergent bottle causes more fixations than the consistent watering can. For the laundry scene, both the watering can and detergent bottle elicit fixations. Fixation heatmaps are generated here for visualization purposes and were created using EyeLink’s Data Viewer. However, analyses were conducted at the level of individual fixations, without smoothing

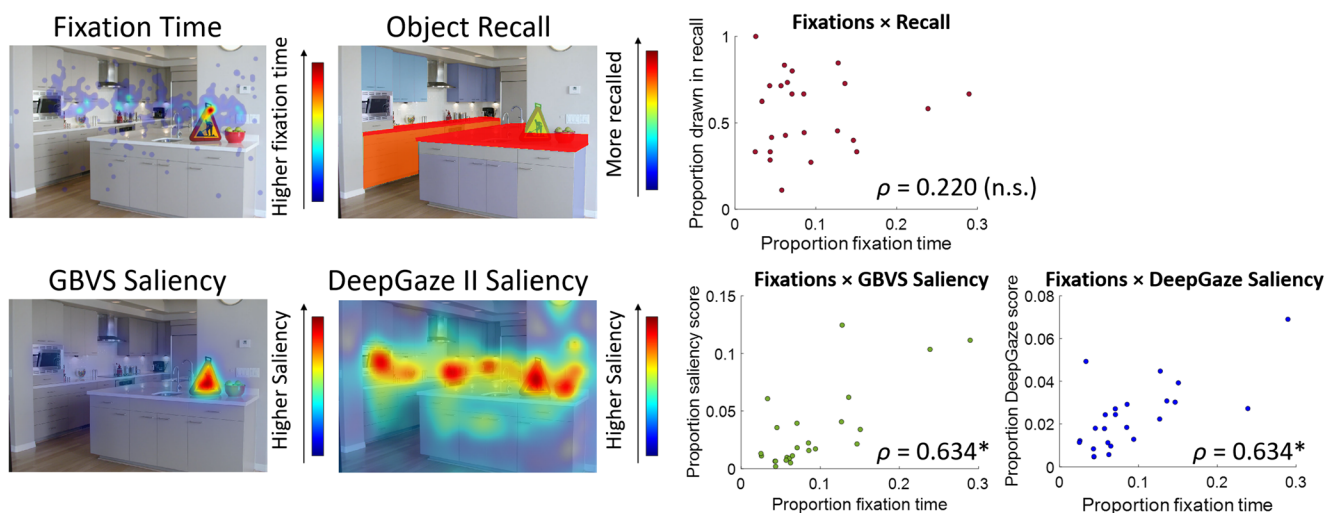


Fig. 8 Comparison of fixation time during study, visual saliency, and recall success. (Left) For each stimulus image, we looked at four types of information: (1) eye fixation times across the pixels of the image, (2) proportion of participants recalling each object in the image, (3) visual saliency of the image calculated using Graph-Based Visual Saliency (Harel et al., 2007), and (4) visual saliency of the image calculated using DeepGaze II (Kümmerer et al., 2016). (Right) Scatterplots of average

fixation time with the three other metrics (recall proportion, GBVS visual saliency, and DeepGaze II visual saliency). Each point represents one of the 24 stimulus images, and indicates the average score for the manipulated object. While saliency metrics were significantly correlated with fixation time, no measure was significantly correlated with recall success. Correlations reported here are Spearman's ρ , with * indicating significant correlations

As a secondary analysis, we conducted an analysis of variance (ANOVA) across all manipulated objects, to see whether fixation time differed based on two factors: (1) whether the object was in a consistent or inconsistent scene, and (2) whether that object was recalled or not. For fixation time, we replicated our significant effect of consistency, where inconsistent objects were fixated longer, $F(1, 229) = 14.30$, $p = 1.99 \times 10^{-4}$, $\eta^2 = 0.06$. We also observed significantly higher fixations for objects that were recalled than for those that were forgotten, $F(1, 229) = 4.25$, $p = 0.001$, $\eta^2 = 0.02$; Recalled: $M = 1,308.5$ ms, $SD = 1,153.3$ ms; Forgotten: $M = 923.5$ ms, $SD = 862.9$ ms, although we observed no significant interaction, $F(1, 229) = 0.39$, $p = .53$. Thus, fixation times do show some relationship to recall success, although this does not appear to be modulated by the consistency of an object with its scene. We also investigated the relationship of computational visual saliency to recall performance. We observed no significant correlation between recall proportion and DeepGaze-predicted saliency ($\rho = 0.137$, $p = .524$), nor GBVS-predicted saliency ($\rho = 0.201$, $p = .346$). For DeepGaze saliency, an ANOVA showed no significant difference between consistent and inconsistent objects ($p = .477$), nor recalled or forgotten objects ($p = .525$), nor a statistical interaction ($p = .284$). Similarly, an ANOVA for GBVS saliency showed no significant difference between consistent and inconsistent scenes ($p = .244$), nor recalled or forgotten objects ($p = .714$), nor a statistical interaction ($p = .455$). Thus, it does not appear that inconsistent objects were much more visually salient than consistent objects, and importantly, image-

based saliency cannot account for differences in memory performance between the consistent and inconsistent objects.

Comparing the temporal order of recall for consistent and inconsistent scenes

In conjunction with recording eye movements during study of the images, we also recorded real-time pen movements during recall of the images (see Fig. 9). Participants did not spend a significantly different amount of time drawing the inconsistent versus consistent manipulated objects (Inconsistent: $M = 16.08$ s, $SD = 7.75$; Consistent: $M = 15.85$ s, $SD = 11.26$), $t(27) = 0.07$, $p = .942$, nor a significantly different amount of time drawing inconsistent versus consistent images (Inconsistent: $M = 2.03$ min, $SD = 0.62$; Consistent: $M = 2.28$ min, $SD = 0.89$), $t(27) = 1.53$, $p = .139$. There was also no significant difference in the order in which inconsistent versus consistent objects were drawn. $t(25) = 1.01$, $p = .323$. It was thus not the case that inconsistent objects were drawn for longer or drawn earlier. There was also no significant correlation at the level of the participant between amount of time spent fixating the manipulated object and amount of time drawing the manipulated object (Spearman's rank correlation: $\rho = 0.21$, $p = .130$). Similarly, there was no significant correlation between fixation order and drawing order for the manipulated object ($\rho = 0.09$, $p = .449$). Thus, time spent during the drawing recall phase does not reveal clear differences between the inconsistent and consistent images, nor a clear relationship to fixations during study.

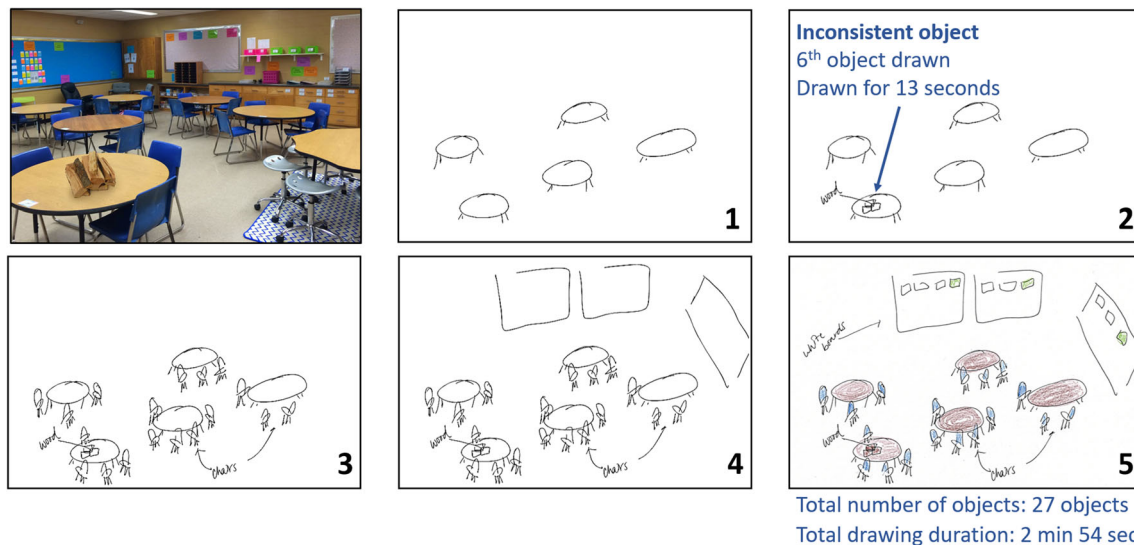


Fig. 9 Example of pen tracking output and recall order analyses. For each drawing, the pen tablet recorded a video of the pen strokes in order. This figure shows 5 ordered example frames from one drawing video of an inconsistent classroom scene (with a pile of logs). For the manipulated

object, a scorer noted the order in which the object was drawn (normalized by total number of objects), and the length of time it was drawn for (normalized by total time spent on the drawing)

Postdrawing recognition task performance

Finally, we also tested participants' memory for the items using a visual recognition task following the drawing recall task. When tested for their recognition of each background scene (with the manipulated object concealed by a gray circle), participants had very high recognition accuracy regardless of whether the scene was originally consistent or inconsistent (Inconsistent: Mean hit rate = 92.2%, $SD = 11.4\%$; Consistent: $M = 88.9\%$, $SD = 13.4\%$), with no significant difference between the groups, $t(29) = 1.00$, $p = .326$. There was also no significant difference in false recognitions of matched foil images from the same scene category (Inconsistent: Mean false alarm rate = 11.2%, $SD = 14.8\%$; Consistent: $M = 10.0\%$, $SD = 11.2\%$), $t(29) = 0.45$, $p = .656$. Participants then were presented with four possible objects to fill in the obscured part of the image: (1) the inconsistent exemplar, (2) the consistent exemplar, (3) a different exemplar image from the inconsistent object category, (4) a different exemplar image from the consistent object category. There was no significant difference between inconsistent and consistent scenes in participants being able to choose the correct item out of the four options (Inconsistent: Mean hit rate = 28.2%, $SD = 20.9\%$; Consistent: $M = 26.5\%$, $SD = 14.6\%$), $t(28) = 0.39$, $p = .697$. There were also no significant differences in the types of errors made by participants. Participants across groups were equally likely to choose an object that appeared in another image (Inconsistent: $M = 27.1\%$, $SD = 19.1\%$; Consistent: $M = 23.5\%$, $SD = 18.5\%$), $t(28) = 0.79$, $p = .434$, or an object of the correct category but an incorrect exemplar (Inconsistent: $M = 18.3\%$, $SD = 16.6\%$; Consistent: $M = 24.8\%$, $SD = 18.3\%$), $t(28) = 1.34$, $p =$

.192. Thus, while we measured differences in recall performance in the drawing task, clear differences in recognition performance did not appear. That being said, participants reported this object recognition task was difficult (occurring after the relatively effortful drawing recall task and with very closely matched foil objects), and performance was relatively low.

Discussion

In this study, we tested how memory representations may differ based on consistent or inconsistent object–scene semantics using a visual recall drawing task. We found that scenes containing inconsistent objects were recalled more often, but with less detail. Further, object–scene inconsistencies resulted in a weaker binding between the object and its scene, with the inconsistent object sometimes drawn in isolation, with an unspecified object identity, or transposed into an entirely different scene. In contrast, while semantically consistent scenes were recalled less frequently, their successful recollections contained more object details, and fewer errors.

These results provide important evidence on the impact of object–scene semantics on memory (Cornelissen & Vö, 2017; Draschkow & Vö, 2017; Friedman, 1979; Hollingworth et al., 2001; Pezdek et al., 1989). Using drawing as a memory output allows for a fine-grained look at how object–scene semantics influence memory representations, and we observe a nuanced trade-off for semantically inconsistent scenes in which memory for the overall image is better, but memory for the objects within it is worse. This dual result could account for the fact that some work had previously observed diminished memory

for inconsistent images (Draschkow & Vö, 2017) while others observed improved memory (Friedman, 1979; Hollingworth et al., 2001; Pezdek et al., 1989). In fact, both effects may be occurring simultaneously at different levels of stimulus information (i.e., the image, the objects, and the background scene). This simultaneous effect may be akin to the trade-off of capacity and precision observed in visual working memory (Roggeman et al., 2014), in which inconsistent scene semantics may result in higher capacity with less precision. Our findings also provide evidence for the contextual guidance model suggesting two parallel pathways for scene processing: one for gist-based global information and one for object-based local information (Torralba et al., 2006; Vö & Wolfe, 2015), as well as fuzzy-trace theory, which posits parallel storage and dissociable retrieval of verbatim versus gist information, which has been shown to parsimoniously account for several false memory findings (Brainerd & Reyna, 2002). While scene–object inconsistencies may result in a distinctive scene with boosted memory for the gist of the scene, they may prevent the ability to use a scene template to fill in local, precise object details (Hollingworth, 2009; Malcolm & Henderson, 2009). A disruption of the scene–object semantics may also result in looser binding of an object to its scene, resulting in a “spotlighting” on the inconsistent object (Cornelissen & Vö, 2017), and a tendency to migrate objects across memory episodes (Hannigan & Reinitz, 2003). Within memory, semantically inconsistent objects may impair abstraction of the scene from a schema template (Hock & Schmelzkopf, 1980; Intraub, 1997), resulting in a loss of schema-coherent details. While the current study focused on the consistency of a single object with its greater scene, investigations of recall for more complex semantic manipulations (e.g., manipulating the semantic relationships of the objects to each other) may provide further insight on how semantics during perception influence the memory representation for a scene.

Considering the role of scene–object consistencies on memory has important real-world implications for how we design scenes, and how we test memory. Some of the first seminal work looking at scene consistency and memory tested memory for real graduate student offices (Pezdek et al., 1989), and recent work has brought questions about scene memory into virtual reality (Helbing et al., 2020). It will be exciting to see whether our findings can help guide the design of real-world scenes, based on what aspects we wish to be memorable (Bainbridge, 2019): a key object, all objects, or the gist of the scene. In some cases, one may want to enhance a specific object even at the cost of surrounding objects being forgotten, while in other cases the goal will be to make an entire landscape memorable. Drawing is also a task that has been historically used as a clinical tool to measure patient groups (Corkin, 2002; Rey, 1941), and recent work has applied these same drawing quantification techniques to aphantasia, a condition of absent visual imagery (Bainbridge et al., 2021). The

current task manipulating scene grammar could potentially reveal insight into groups with differing abilities at visual, semantic, or mnemonic processing, such as individuals across the life span.

While we observed differences in recall for consistent and inconsistent scenes, we also observed several similarities between memory representations for consistent and inconsistent scenes. Between these two conditions, recalled drawings tended to be equally diagnostic, have equally high spatial accuracy (in terms of both object location and size), and have equally rare numbers of additional objects inserted into the drawings. We also did not observe differences between the two conditions in visual recognition performance (although this could be due to the difficulty of the recognition task). Thus, while scene semantics may influence some aspects of a memory (e.g., memory for other objects in an image), it may have less of a sway on other aspects of that memory (e.g., spatial accuracy). Indeed, various work has suggested differences in how object and spatial information may be coded in memory (Bainbridge et al., 2021; Farah & Hammond, 1988; Staresina et al., 2011). While the current work investigates scene semantics, other work has suggested that scene *syntax*—the spatial arrangement of semantically consistent objects within a scene—as a similarly meaningful organizational principle for scenes (Vö et al., 2019). An experiment manipulating scene syntax rather than semantics (e.g., moving a consistent object to an inconsistent location) may result in higher spatial error but preserved object accuracy in memory.

While the current study serves as important evidence towards a dissociation of scene versus object memory for inconsistent scenes, some caveats of this work motivate future studies. We utilized stimulus images that were manipulated to appear natural, regardless of the object consistency. However, future studies could explore similar methods using stimuli that prioritize systematic manipulation of the images (e.g., keeping object size, location, lighting, and shadowing consistent across all images), rather than naturalness of the stimuli. Further, while we decided *a priori* on a sample size validated from prior work (Bainbridge et al., 2019), some results showed small to medium effect sizes or null effects that would be valuable to replicate in follow-up research. A future study could also increase the number of images per participant in order to look at influences of saliency and fixation patterns on within-participant recall performance; however, we do note that participants may not be able to recall many more images. Also, as this task required detailed memorization of the scene, all visual information was highly task relevant. However, prior research has shown that some tasks such as visual search can result in higher recall performance than explicit memorization (Draschkow et al., 2014). It would be interesting to see if an incidental study task would drive even stronger differences between object and scene recollection. Relatedly, it would be interesting to see whether different

explicit instructions (such as telling participants to indicate vaguely remembered objects) would influence the information present in memory drawings. Finally, there are still many open questions about how drawing as a recall task itself may influence remembered information. Some work in children has shown that drawing of a memory can increase accurately recalled information, but it also has a tendency to increase false memories (Bruck et al., 2000; Otgaar et al., 2016). Other work has shown that artists are able to produce more memory information than nonartists (Perdreau & Cavanagh, 2015; Vogt & Magnussen, 2016), potentially suggesting that different strategies may boost performance on the task. Thus, further investigation into the limitations as well as potential for drawing as a memory task will be highly important in future work.

Finally, our results also suggest attention-based and fixation-based models may be insufficient models for recall. Here, we are successfully able to replicate findings suggesting that individuals fixate inconsistent objects during perception (De Graef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978; Malcolm & Henderson, 2010). We also observe significantly higher fixation times on objects that are recalled versus those that are forgotten. However, we do not observe that this effect is modulated by the consistency of the object. We also do not observe correlations between eye-tracking patterns during study and pen-tracking patterns during recall. In terms of computer-vision-based visual saliency metrics, we are able to replicate prior work showing that they can successfully model eye movements on an image (Harel et al., 2007; Kümmerer et al., 2016). We find that these saliency measures are not different between inconsistent and consistent versions of an object, suggesting consistency effects are not strongly driven by visual differences between conditions. That being said, while we had counterbalanced consistent–inconsistent pairs across participants, it is still possible the inconsistent images may have been more visually striking (e.g., a colorful beach ball in a monochromatic laboratory) and driven fixation behavior. Indeed, we wonder if semantically consistent objects tend to share low-level visual features, making it difficult to create equally salient inconsistent images. However, if image saliency were to drive recall performance, we would expect to observe a relationship between fixations during encoding and pen movements during recall, which we do not find. Thus, our findings are likely due to semantically driven differences in memory rather than visually driven differences. Prior work has found key differences between saliency-based predictions and recall, such as a lower visual field bias for object recall not present in saliency models (Bainbridge et al., 2019) as well as an inability for saliency models to capture semantically meaningful portions of an image (Bylinskii et al., 2016; Henderson & Hayes, 2017). The current work highlights a need for image-based metrics aimed at making predictions specific to scene memory, accounting for semantic

abstraction of the scene as well as what objects and features are memorable (Bainbridge, 2019). Future work could examine scenes with graded levels of inconsistency, in order to create more nuanced models that can account for both semantic inconsistency as well as visual saliency.

In sum, this study reveals a multi-pronged impact of scene semantics on visual memory representations. While semantic inconsistencies result in highly atypical images that are remembered overall, these inconsistencies disrupt memory for local object detail in the scenes.

Acknowledgements We thank Anna Corriveau for her help digitizing the drawings from the study, and Adam Dickter for his help with the eye tracker system. This research was supported by the Intramural Research Program of the National Institutes of Health (ZIA-MH-002909), under National Institute of Mental Health Clinical Study Protocol 93-M-1070 (NCT00001360).

References

- Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review*, *14*, 332–337. <https://doi.org/10.3758/BF03194073>
- Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember. In K. Federmeier & D. Beck (Eds.), *Psychology of learning and motivation* (Vol. 70, 1–27). Elsevier. <https://doi.org/10.1016/bs.plm.2019.02.001>
- Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*, Article 5. <https://doi.org/10.1038/s41467-018-07830-6>
- Bainbridge, W. A., Pounder, Z., Eardley, A. F., & Baker, C. I. (2021). Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. *Cortex*, *135*, 159–172. <https://doi.org/10.1016/j.cortex.2020.11.014>
- Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164–169. <https://doi.org/10.1111/1467-8721.00192>
- Bruck, M., Melnyk, L., & Ceci, S. J. (2000). Draw it again Sam: The effect of drawing on children's suggestibility and source monitoring ability. *Journal of Experimental Child Psychology*, *77*(3), 169–196. <https://doi.org/10.1006/jecp.1999.2560>
- Bylinskii, Z., Recansens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where should saliency models look next? *European Conference on Computer Vision* (pp. 809–824). Springer.
- Corkin, S. (2002). What's new with the amnesic patient HM? *Nature Reviews Neuroscience*, *3*(2), 153–160. <https://doi.org/10.1038/nm726>
- Cornelissen, T. H. W., & Vö, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object–scene inconsistencies modulate ongoing gaze behavior. *Attention, Perception, & Psychophysics*, *79*, 154–168. <https://doi.org/10.3758/s13414-016-1203-7>
- Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*(3), 393–401. <https://doi.org/10.3758/BF03193280>
- De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329.
- Draschkow, D., Wolfe, J. M., & Vö, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build

- better memories. *Journal of Vision*, 14(8), 10. <https://doi.org/10.1167/14.8.10>
- Draschkow, D., & Vö, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, 7, Article 16471. <https://doi.org/10.1167/14.8.10>
- Farah, M. J., & Hammond, K. M. (1988). Visual and spatial mental imagery: Dissociable systems of representations. *Cognitive Psychology*, 20, 439–462. [https://doi.org/10.1016/0010-0285\(88\)90012-6](https://doi.org/10.1016/0010-0285(88)90012-6)
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 1–29. <https://doi.org/10.1167/7.1.10>
- Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108, 316–355. <https://doi.org/10.1037/0096-3445.108.3.316>
- Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics*, 77, 1239–1251. <https://doi.org/10.3758/s13414-015-0859-8>
- Hannigan, S. L., & Reinitz, M. T. (2003). Migration of objects and inferences across episodes. *Memory & Cognition*, 31(3), 434–444. <https://doi.org/10.3758/BF03194401>
- Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19, 545–552.
- Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review*, 26, 1683–1689. <https://doi.org/10.3758/s13423-019-01642-5>
- Helbing, J., Draschkow, D., & Vö, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, 196, 104147.
- Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1(10), 743–747. <https://doi.org/10.1167/18.6.10>
- Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1), 210–228.
- Henderson, J., 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. *OSF Preprints*.
- Hock, H.S., & Schmelzkopf, K. F. (1980). The abstraction of schematic representations from photographs of real-world scenes. *Memory & cognition*, 8, 543–554. <https://doi.org/10.3758/BF03213774>
- Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. *Visual Cognition*, 17(1/2), 273–291.
- Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127, 398–415.
- Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, 8(4), 761–768.
- Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, 1(6), 217–222. [https://doi.org/10.1016/S1364-6613\(97\)01067-X](https://doi.org/10.1016/S1364-6613(97)01067-X)
- Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). *DeepGaze II: Reading fixations from deep features trained on object recognition*. arXiv: 1610.01563[cs.CV].
- Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4), 565–572.
- Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11), 8. <https://doi.org/10.1167/9.11.8>
- Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, 10(2), 1–11.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23–36.
- Otgaar, H., van Ansem, R., Pauw, C., & Horselenberg, R. (2016). Improving children’s interviewing methods? The effects of drawing and practice on children’s memories for an event. *Journal of Police and Criminal Psychology*, 31(4), 279–287. <https://doi.org/10.1007/s11896-016-9190-0>
- Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 587–595. <https://doi.org/10.1037/0278-7393.15.4.587>
- Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, 206, Article 104465. <https://doi.org/10.1016/j.cognition.2020.104465>
- Perdreau, F., & Cavanagh, P. (2015). Drawing experts have better visual memory while drawing. *Journal of Vision*, 15(5), 5. <https://doi.org/10.1167/15.5.5>
- Rey, A. (1941). L’examen psychologique dans les cas d’encéphalopathie traumatique. (Les problèmes) [Psychological examination in cases of traumatic encephalopathy]. *Archives de psychologie*, 28, 286–340.
- Roggeman, C., Klingberg, T., Feenstra, H. E. M., Compte, A., & Almeida, R. (2014). Trade-off between capacity and precision in visuospatial working memory. *Journal of Cognitive Neuroscience*, 26(2), 211–222. https://doi.org/10.1162/jocn_a_00485
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1/3), 157–173.
- Staresina, B. P., Duncan, K. D., & Davachi, L. (2011). Perirhinal and parahippocampal cortices differentially contribute to later recollection of object- and scene-related event details. *Journal of Neuroscience*, 31, 8739–8747.
- Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4), 766–786. <https://doi.org/10.1037/0033-295X.113.4.766>
- Vö, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210. <https://doi.org/10.1016/j.copsyc.2019.03.009>
- Vö, M. L.-H., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, 1339(1), 72. <https://doi.org/10.1111/nyas.12667>
- Vogt, S., & Magnussen, S. (2016). Expertise in Pictorial Perception: Eye-Movement Patterns and Visual Memory in Artists and Laymen. *Perception*, 36(1), 91–100.
- Zhao, Q., & Koch, S. (2013). Learning saliency-based visual attention: A review. *Signal Processing*, 93(6), 1401–1407. <https://doi.org/10.1016/j.sigpro.2012.06.014>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.