

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

3

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

5

6 1 – Department of Psychology, University of Chicago, Chicago, IL 60637

7 2 – Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, MD 20814

8 3 – University of Cincinnati College of Medicine, Cincinnati, OH 45267

9 * indicates equal contribution

10

11 Corresponding author:

12 Wilma A. Bainbridge, wilma@uchicago.edu

13 (773) 702-3189

14 Department of Psychology, University of Chicago

15 5848 South University Ave, 303 Beecher Hall, Chicago, IL 60637

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

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

37

Introduction

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

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 (Oliva & Torralba, 2006; Fei-Fei et al., 2007). In exploring and understanding that scene, we are not only guided 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 (Loftus & Mackworth, 1978; De Graef et al., 1990; Henderson et al., 1999; 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 (Cornelissen & Võ, 2017) when encoded incidentally. 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; Pedzek et al., 1989; Hollingworth et al., 2001). 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 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

62 specifically asked *how* that memory is impacted: what specific aspects of a scene's memory are
63 altered—memory for the entire image, the manipulated object, or the background scene?

64 In the current study, we compare the underlying visual memory representations for
65 inconsistent and consistent scenes using a visual recall drawing task. Previous work assessing
66 the influences of scene semantics on memory have relied on verbal recall or visual recognition
67 tasks (Friedman, 1979; Pezedk et al., 1989; Hollingworth & Henderson, 1998; Cornelissen &
68 Vő, 2017). However, these types of measures may provide limited information about the nature
69 of a memory – only revealing whether an item is remembered or not, but not what specific visual
70 content of that memory drives its recollection. Recent work has discovered that using drawing-
71 based recall to quantify memory can reveal more fine-grained information than verbal
72 recollection, and requires no assumptions about matched foil images, often necessary for visual
73 recognition tasks (Bainbridge et al., 2019). Drawings have also been known to reveal valuable
74 insight about the memories of children (e.g., Bruck et al., 2000; Otgaar et al., 2016), artists (e.g.,
75 Vogt & Magnussen, 2007; Perdreau & Cavanagh, 2015), and patient groups (e.g., Rey, 1941;
76 Corkin, 2002). Thus, a drawing task may reveal subtler differences between memories for
77 consistent and inconsistent scenes than was possible to capture in previous work utilizing
78 verbal- or recognition-based tasks. Further, drawings can be objectively quantified through
79 online crowd-sourced scoring to reveal a wide range of information, including object detail,
80 spatial accuracy, and inclusion of false additional objects (Bainbridge et al., 2019). With such a
81 task, we can thus examine not only whether an inconsistent object is remembered better than a
82 consistent object, but how inconsistency impacts memory for other objects in the scenes, their
83 spatial relations, and the scene overall.

84 With these measures, we examined several questions for how scene semantics might
85 influence memory representations. First, we consider how consistency affects recall of the
86 manipulated object in the scene. One possibility is that inconsistent objects are distinctive and
87 easier to remember (Friedman, 1979; Pedzek et al., 1989; Hollingworth, Williams & Henderson,

2001). 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 & Vö, 2017). A third possibility is that we may observe no memory difference between inconsistent and consistent objects (Cornelissen & Vö, 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 & Vö, 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 & Vö, 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

114 during encoding, we find that recall patterns cannot be explained by fixation patterns or image
115 saliency during encoding.

116

117 **Methods**

118 **Participants**

119 Thirty adults (9 male, 21 female; age $M=24.8$ years, $SD=4.5$) were recruited from the
120 local Washington D.C. area for participation in this within-subjects experiment. This sample size
121 was determined *a priori*, to match the same sample size collected in a previous, similar drawing-
122 based experiment that measured high detail from memory drawings with only fifteen participants
123 drawing any given image (Bainbridge et al., 2019). The current study also includes far fewer
124 images to hold in memory (twelve versus thirty), thus we anticipate more drawings will be
125 produced from memory per image, resulting in higher power per image than Bainbridge et al.,
126 2019. However, as this drawing recall methodology is very new, we hope the current study will
127 serve as a basis upon which to conduct power analyses for future drawing studies. Participants
128 were healthy native English speakers with corrected or normal vision, with the exception of
129 participants with high-prescription glasses, who were not recruited, to avoid calibration issues
130 with the head-mounted eye tracker. No participants or trials were excluded. All participants
131 consented following the guidelines of the National Institutes of Health (NIH) Institutional Review
132 Board (NCT00001360, 93M-0170) and were compensated for their participation.

133 1,725 online scorers were recruited from online crowd-sourcing task platform Amazon
134 Mechanical Turk (AMT), acknowledging their participation following the guidelines of the NIH
135 Office of Human Subjects Research Protections (OHSRP), and were also compensated for their
136 participation. The number of online scorers per task was selected to be identical to prior online
137 scoring studies of drawings (Bainbridge et al., 2019; Bainbridge et al., 2021).

138

139 **Stimuli**



140

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

144

145 Stimulus images were created from twelve distinctive scene images from different scene
 146 categories, half indoor (bathroom, bedroom, classroom, kitchen, laboratory, laundry room) and
 147 half outdoor (campsite, construction site, neighborhood street, playground, swimming pool,
 148 backyard). The original object and scene images came from publicly available photographs on
 149 Google Images, found by searches of the scene category and object names. Adobe Photoshop
 150 was used to naturally add an object to each image (referred to throughout as the “manipulated
 151 object”) that was either consistent or inconsistent with the scene semantics (Figure 1). The
 152 scene images were paired, and these object manipulations were conducted within the pairs, so
 153 that the consistent object in a given image was also used as the inconsistent object in its paired
 154 image, and vice versa. For example, in the consistent condition, a lab scene contained a
 155 microscope and a pool scene contained a beach ball (Figure 1). In the inconsistent condition,
 156 the lab scene had a beach ball and the pool scene had a microscope. The consistent and
 157 inconsistent object were placed at the same size and in the same location within a given scene,
 158 and shadowing and lighting were added to each object to integrate it naturally with the
 159 surrounding scene. This resulted in a set of 24 stimuli, comprising of a consistent and

160 inconsistent version of each of the twelve scenes (and, similarly, each of twelve objects had a
161 consistent image and an inconsistent image). To confirm that we successfully manipulated
162 scene consistency, all images were rated online by Amazon Mechanical Turk (AMT) workers
163 (N=15 per image; N=67 total) on a 5-point Likert Scale on how typical (“normal”) it was for that
164 object to be in the scene (1 = Very Abnormal, 5 = Very Normal). As expected, consistent objects
165 were rated to be significantly more normal than inconsistent objects (Consistent: M=4.4 SD=1.1;
166 Inconsistent: M=1.6 SD=1.1; Wilcoxon signed rank test: $Z=2.20$, $p=0.028$, effect size $r=0.64$).
167 During the main experiment, each participant saw 12 images (one of each background scene),
168 with half consistent images and half inconsistent images. Which images were consistent or
169 inconsistent was counterbalanced across participants, so each of the 24 images was ultimately
170 seen by fifteen participants, akin to Bainbridge and colleagues (2019).

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

180

181 **Experimental Procedures**

182 At the beginning of the experiment, participants were told to carefully examine each
183 image as they would be later tested on their memory. Participants were informed that this was a
184 memory task, so that they were motivated to actively encode the image details for long-term
185 memory. However, participants were unaware of the nature of the memory task, and were

186 unlikely to expect a drawing task, so they could not employ strategies specifically honed for
187 drawing as a task. The experiment was split into four phases (Figure 2).

188 The first phase was a study phase, in which participants viewed each of the images for
189 10 seconds while their eye movements were tracked with a head-mounted EyeLink 1000 Plus
190 eye-tracking device. Participants studied twelve images in total, determined as the average
191 number of scenes recalled by participants in a prior memory drawing study ($M=12.1$ images,
192 Bainbridge et al., 2019). We anticipated that twelve images per participant would maximize the
193 power of our study; more images would likely result in a low recall rate for any given image,
194 while fewer images could be too easy and reduce the experimental power. Between the
195 presentation of each stimulus image, a fixation cross was displayed to the right of the image on
196 the screen, in order to avoid biasing eye movements to the center. After the participant fixated
197 on the cross, the next stimulus was then displayed. Each image was displayed at 1200 x 800
198 pixels on a 1920 x 2000 resolution 24-inch screen.

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

204 The third phase was the drawing recall test phase. Participants were given sheets of
205 paper with a rectangular outline with dimensions matching those of the original images, and
206 were asked to draw as many images as they could remember in as much detail as possible.
207 Participants drew on a Wacom Paper Pro tablet, which allowed participants to draw with an
208 inked pen on paper while it simultaneously recorded pen strokes digitally in real time.
209 Participants were told to draw the images in any order. They were also given colored pencils if
210 they wanted to include color detail in their drawings, but were asked to include color only if they
211 specifically recalled it. They were instructed to add color after completing the pen drawing of all

212 the objects they recalled. Participants were also told they could label objects if they wished to
 213 clarify what they were. Participants were given as much time as they needed and took 27
 214 minutes on average for the recall phase (SD=8).

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

229

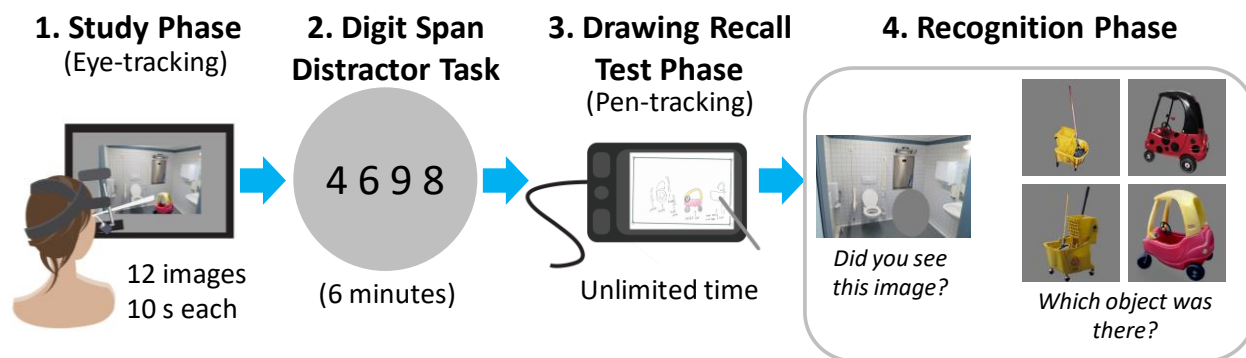


Figure 2 – The experimental procedures. The experiment consisted of four phases: 1) A study phase in which participants studied 12 randomly ordered images (6 consistent, 6 inconsistent) for 10s each while

233 their fixations were tracked, 2) A digit span distractor task in which participants had to verbally recall digit
234 series, 3) A drawing recall test phase in which participants drew the studied scenes from memory while
235 their pen movements were tracked, and 4) A recognition phase in which participants had to separately
236 recognize the scene and the manipulated object.

237



238 **Online Scoring Procedures**

239 The resulting 275 drawings were scanned and uploaded to AMT, to crowd-source
240 worker ratings on several properties of the memory drawings. Specifically, four different sets of
241 measures were collected for each drawing. For all rating tasks, each AMT worker could
242 participate in as many trials as they wanted. Depictions of the four online tasks can be seen in
243 Figure 3, with the precise instructions given to AMT workers. AMT workers did not know the
244 origins of these drawings, the different image conditions (i.e., consistent versus inconsistent
245 scenes), nor the nature of the main drawing experiment. Drawings were also randomly mixed,
246 so that if an AMT worker participated on multiple trials, they would not know if they were scoring
247 drawings from the same (or different) person or conditions. Thus, AMT workers scored these
248 drawings blind to the conditions.

249

250 A)

Drawing Match


Is this a drawing of this image?






Definitely Not Probably Not Unsure
 Probably Definitely

Object Identification

Choose the objects in the drawing.

The photographs below have an object outlined in red. Click the picture if that object is in the drawing.




Object Location

Circle the object in the drawing!

Move and resize the circle to cover *only* the same object in the drawing as the one that is outlined in red.

False Additional Object

Spot the difference!

Write down any objects in the drawing that are **NOT** in the photo.




New objects:

251

252 B)



pool



beach ball



umbrella



planter



building

253

254 **Figure 3 – A) Example trials from the online experiments.** Shown are depictions of example trials from

255 the four online experiments: 1) Drawing Match Scoring, 2) Object Identification Scoring, 3) Object

256 Location Scoring, and 4) False Additional Object Scoring. **B) Up-close view of highlighted objects.** The
257 Object Identification Scoring and Object Location Scoring tasks required online participants to compare a
258 highlighted object with the drawings. Shown here are the close-up examples of objects highlighted in red
259 from “Object Identification” (3A). Outlines were created *a priori* using LabelMe (see Methods). Online
260 participants saw the full image when making responses (as in 3A).

261
262 **Drawing Match Scoring.** AMT workers rated how well each drawing matched one of
263 the images participants studied, providing a measure of diagnosticity of that drawing (Figure 3A,
264 top left). For a given trial, they were presented with a drawing with an image next to it and rated
265 on a 5-point Likert scale the likelihood that it was a drawing of that image (1= Definitely Not, 5 =
266 Definitely). Across trials, drawings were tested against each of the 12 images seen by the
267 participant who made the drawings. Twelve ratings were collected for each drawing-image pair
268 (with a total of 144 ratings per drawing across the 12 pairs), and 611 AMT workers participated
269 in total. The image with the highest match rating with a drawing (averaged across the 12 AMT
270 workers for that drawing-image pair) was selected as the corresponding image (“original image”)
271 for that drawing. To score overall image recall performance, each of the 12 images each
272 participant saw was given a binary score (1 = was drawn, 0 = was not drawn) as determined by
273 if a drawing was matched to it by the AMT scorers. Number of recalled images was calculated
274 as the sum of those 12 binary scores. If a participant made multiple drawings of the same
275 image, that image was only given a single score for being remembered. However, these
276 duplicate drawings were scored for other measures (below) and their objects present were
277 taken as the union across duplicates within participants (rather than us selecting a single
278 drawing to be scored for a given image). Seventeen participants drew multiple drawings of the
279 same image.

280 **Object Identification Scoring.** For each drawing-image pair resulting as the highest
281 match from the Drawing Match Scoring, AMT workers determined which objects from that image

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

298 ***Object Location and Size Scoring.*** AMT workers determined the locations and sizes of
299 each object present in the drawings. For a given trial, they were presented with a drawing and
300 its matched image with an object highlighted on it. On the drawing, they had to place and resize
301 an ellipse to encircle that specified object. Five AMT workers made ellipses for each object and
302 453 AMT workers participated in total. The final ellipse was determined by the median centroid
303 and radii in the x and y directions. This scoring was conducted for all objects determined to exist
304 in a given drawing, based on ratings in the Object Identification Scoring. One participant did not
305 draw any consistent scenes, and so they were not included in analyses comparing locations of
306 objects in consistent versus inconsistent scenes. One participant also did not draw any

307 inconsistent manipulated objects, and so an analysis comparing the location and size of
308 manipulated objects only included 28 of the 30 total participants.

309 ***False Additional Object Scoring.*** AMT workers determined the presence of additional
310 objects in the drawings that were not in the original images. For a given trial, they were
311 presented with a drawing and its corresponding image and had to write down all objects that
312 existed in the drawing but not the image. Fifteen AMT workers rated each image and 200 AMT
313 workers participated in total. Any objects listed by at least five workers were counted as false
314 alarms.

315

316 **Fixation, Pen-tracking, and Saliency Analyses**

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

324 Using the tablet recordings of the pen movements, we also calculated amount of time
325 spent drawing each manipulated object per participant. An in-lab scorer watched the video of
326 pen strokes created by the drawing tablet for each drawing. The start and end time of pen
327 strokes for the manipulated objects were noted for each image, for the first span of time in which
328 the object was drawn. Time spent coloring objects was not included, as participants were
329 instructed to add color after completing their drawing (and the tablet could not track colored
330 pens / pencils). Object drawing time also did not include any time spent returning to add details
331 to an object later. Total amount of time spent drawing the object was calculated as the
332 difference between the end time and the start time, normalized by total amount of time spent on

333 the drawing. Similarly, we calculated sequential drawing order per participant by assigning an
334 order to each object based on first pen stroke on that object. Drawing order was then
335 normalized by total number of objects in the drawing. One participant was removed from the
336 drawing time and drawing order analyses due to a technical glitch with the pen tablet software
337 (resulting in N=29 for these analyses).

338 To compute image saliency scores, we used two state-of-the-art computer vision
339 algorithms designed to predict human fixation time: DeepGaze II (Kümmerer, Wallis, & Bethge,
340 2016) and Graph-Based Visual Saliency (Harel, Koch, & Perona, 2007). Both models aim to
341 predict human fixations of an image, but DeepGaze II is a more recent approach that utilizes a
342 wide range of feature types (i.e., both low-level and high-level visual information) and is trained
343 on human fixation data, while Graph-Based Visual Saliency (GBVS) is a more established
344 method commonly tested by attention researchers, that relies solely on image-computable low-
345 level visual features. Specifically, DeepGaze II predicts fixation time based on features from the
346 VGG-19 deep neural network for object identification combined with a readout network trained
347 for saliency prediction based on human fixations (Kümmerer et al., 2016). In contrast, GBVS is
348 a model that identifies visually dissimilar regions of an image (Harel et al., 2007). We were
349 curious to see if these two models would perform differently in their predictions of recall and
350 fixation behavior, given current debates on the success of these models in predicting scene
351 semantics (Hayes & Henderson, 2019; Pedziwiatr et al., 2021; Henderson et al., 2021). For both
352 metrics, we obtained saliency heatmaps for each of the stimulus images (see Figure 8). Object-
353 based saliency was then calculated as the average saliency across the pixels of that object,
354 normalized by the average saliency of the entire image.

355 Finally, to generate heatmaps of recall for each image, we calculated a recall score for
356 each object, calculated as the number of participants who drew that object, divided by the
357 number of participants who drew the image containing that object. This allows us to create a

358 heatmap of how well different objects in an image were remembered, that can be directly
359 compared to heatmaps formed from fixation patterns, pen movements, or saliency measures.

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

366

367 **Data Analyses**

368 For most analyses, we conducted paired samples t-tests within subjects to compare the
369 above metrics between participants' drawings of consistent scenes versus inconsistent scenes.
370 We first tested these metrics for normality using a Kolmogorov-Smirnov goodness-of-fit test, and
371 found none of these were significantly different from a normal distribution (all $p > 0.05$). For
372 metrics with limited ranges (e.g., Likert scales, number of drawings made), we instead
373 conducted non-parametric paired samples Wilcoxon signed rank tests. Effect sizes are included
374 with all significant statistical tests.

375

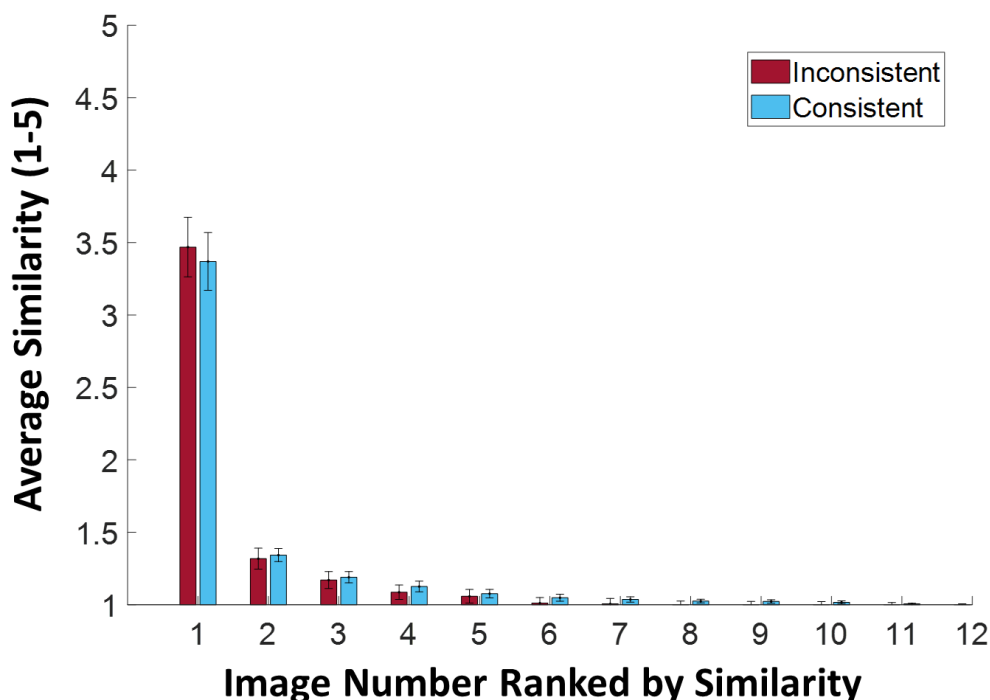
376 **Results**

377

378 **Drawings are highly diagnostic of their images**

379 The first step is identifying what image each drawing corresponds to. Further, given the
380 range of people's drawing abilities and memory, can a separate group of participants tell what
381 image a drawing represents? AMT workers saw individual drawings matched with each of the
382 12 images studied by participants in the main experiment, and judged the likelihood that the
383 drawing was of that image on a scale of 1 (Definitely Not) to 5 (Definitely). Overall, it was clear

384 to AMT workers what images matched the drawings, with only a single image getting a score
 385 above 3 on average (Figure 4). For all further analyses, the highest rated image was taken as
 386 the corresponding image for each drawing. Importantly, there was no significant difference in
 387 ratings between consistent and inconsistent images (Consistent: $M=3.9$, $SD=0.5$; Inconsistent:
 388 $M=3.7$, $SD=0.6$; Wilcoxon signed rank test: $Z=1.70$, $p=0.090$). This means that both
 389 semantically consistent and inconsistent drawings could be matched with their original images,
 390 and were equally diagnostic of their original image. However, this rating of diagnosticity serves
 391 as a relatively coarse metric, as several different features could contribute to being able to
 392 successfully match a drawing to an image (i.e., the manipulated object, properties of the
 393 background scene). A closer look at the content in these drawings may reveal key differences
 394 between consistent and inconsistent images.



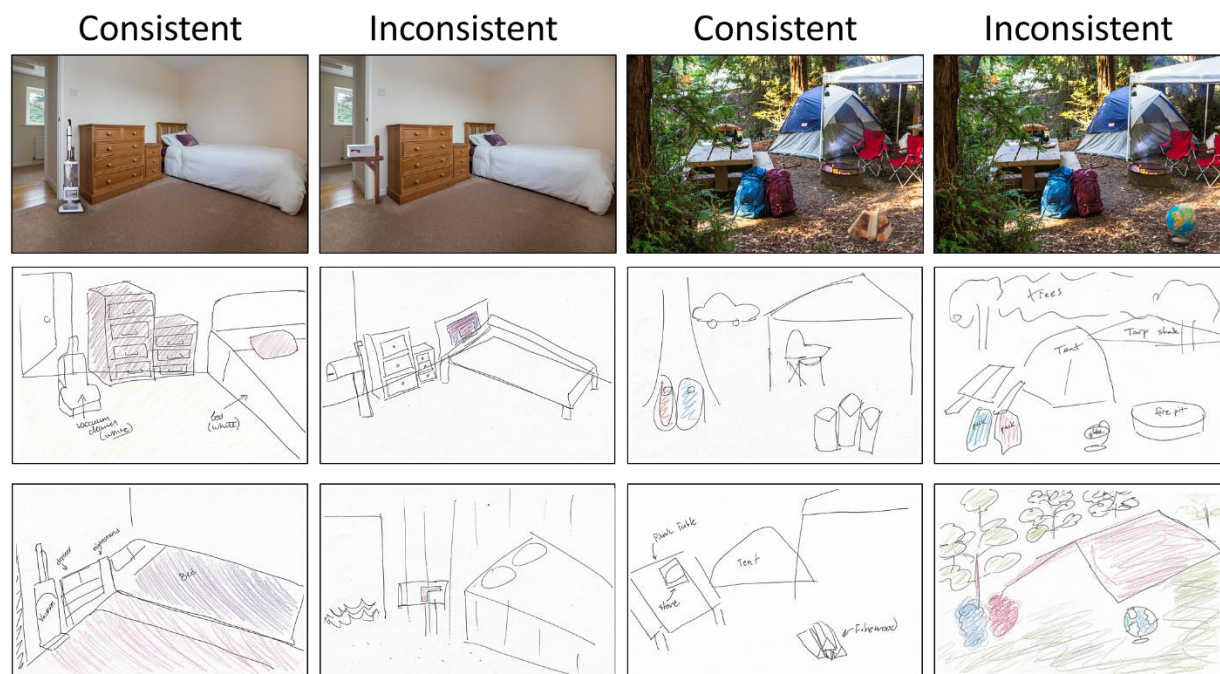
395
 396 **Figure 4 – Diagnosticity of the drawings across images for the two conditions.** The average ratings
 397 made by online scorers of the similarity of participants' drawings to each of the 12 images they saw,
 398 ranked from highest to lowest. Similarity here was assessed on a scale of 1 (low) to 5 (high) with a

399 question asking how likely it was that a drawing was of a given image (the “Drawing Match Scoring”
400 experiment). Red bars indicate drawings made of inconsistent images, while blue bars indicate drawings
401 made of consistent images. The high spike for image #1 and quick drop-off for images #2-12 (all
402 averaging below a rating of 1.5) for both conditions indicates that it was clear to AMT scorers that a given
403 drawing was highly similar to only one image and dissimilar from all others. In other words, drawings were
404 highly diagnostic of their images. There was no significant difference in diagnosticity between consistent
405 and inconsistent images. Error bars indicate standard error of the mean.

406

407 **More inconsistent scenes are recalled than consistent scenes**

408 The memory drawing experiment resulted in 275 total drawings, with 126 drawings of
409 consistent images, and 149 drawings of inconsistent images (Figure 5). This reflects a general
410 tendency to recall inconsistent images over consistent images (Chi-squared test for proportions:
411 $\chi^2=3.85$, $p=0.050$, effect size $\phi=0.12$). Each participant on average drew 9.2 drawings from
412 memory out of the 12 that they studied (SD=2.16, Min=5, Max=12). Of those drawings,
413 participants drew more inconsistent images than consistent images from memory (Wilcoxon
414 signed rank test: $Z=2.10$, $p=0.036$, $r=0.38$), drawing on average 5.0 inconsistent images
415 (SD=1.5) and 4.2 consistent images (SD=1.5). Thus, memory for inconsistent images overall
416 was better than that for consistent images.



417

418 **Figure 5 – Example drawings for four of the stimulus images.** Two example drawings each for the
 419 consistent and inconsistent bedroom scene, and the consistent and inconsistent camp scene. Each
 420 drawing was taken from a different participant, showcasing the impressive level of both object and spatial
 421 detail in the memory drawings for a range of people. The key question in this study is whether there are
 422 differences in memory detail between drawings for the consistent and the inconsistent scenes.

423

424 **More objects are recalled in consistent scenes than inconsistent scenes**

425 Next, we looked at the amount of detail available in each drawing by having AMT
 426 workers judge which objects from the original image were included in each drawing. Each image
 427 contained on average 39.3 objects (with the same number of objects across the consistent and
 428 inconsistent versions of a given scene), and participants drew on average 9.0 objects per image
 429 (SD=2.9), or 77.6 objects on average across the experiment. Participants drew a significantly
 430 higher proportion of the objects in consistent drawings versus inconsistent drawings
 431 (Consistent: M=23.4%, SD=6.9%; Inconsistent: M=19.8%, SD=8.1%; Paired t-test, excluding
 432 the manipulated object: $t(28)=2.56$, $p=0.016$, $d=0.52$). We then looked to see whether there

433 were differences in the tendency to draw the manipulated object (the consistent or inconsistent
434 object). We found no significant difference between consistent or inconsistent drawings in the
435 proportion containing the manipulated object (Consistent: $M=53.8\%$, $SD=25.8\%$; Inconsistent:
436 $M=65.0\%$, $SD=28.1\%$; $t(28)=1.37$, $p=0.181$). This indicates that while semantically inconsistent
437 objects were recalled just as frequently as their consistent counterparts, there was reduced
438 memory for objects in the inconsistent background scenes than consistent ones.

439

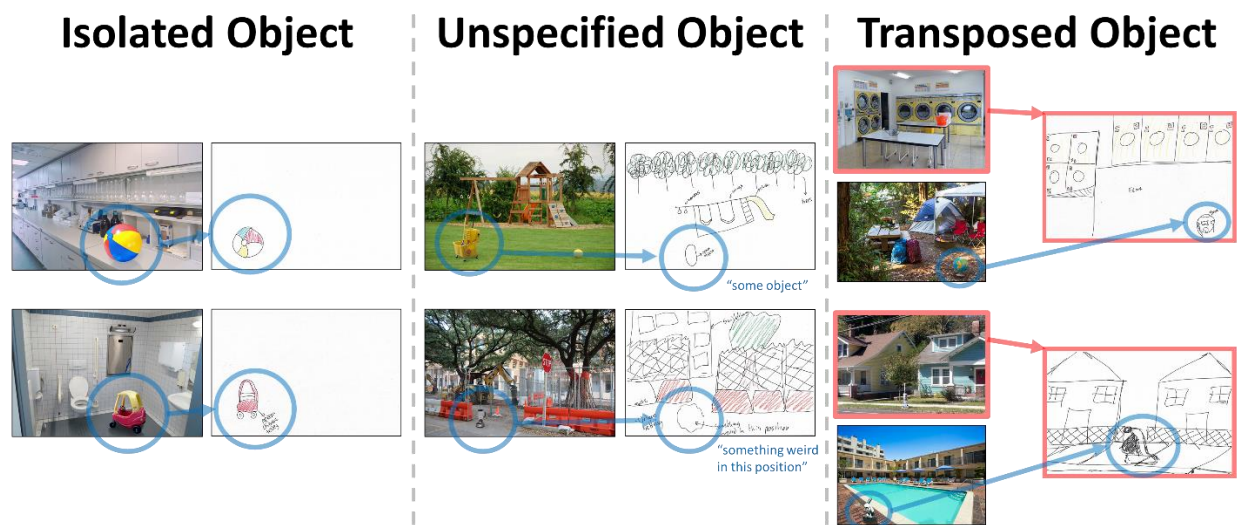
440 **The nature of object errors in consistent versus inconsistent scenes**

441 We then investigated whether there were differences between conditions in the types of
442 object errors that were made in the drawings. Overall, participants included relatively few false
443 additional objects in their drawings, only drawing 25 objects that did not exist in the consistent
444 images (across the 126 drawings from all participants), and 21 objects that did not exist in the
445 inconsistent images (across 149 drawings). Within participants, there was no significant
446 difference in the number of false additional objects they drew for inconsistent scenes versus
447 consistent scenes ($t(29)=0.64$, $p=0.526$). Thus, differences in scene semantics did not appear to
448 induce differences in false memories in these drawings.

449 However, participants made intriguing errors with the manipulated object in their
450 drawings (see Figure 6). In 18 drawings (13 inconsistent, 5 consistent), participants made
451 drawings of only the manipulated object, unable to recall the surrounding background scene. In
452 6 drawings (6 inconsistent, 0 consistent), participants drew a detailed scene and included a
453 circle with an unspecified object; they remembered that the manipulated object was there, but
454 not what it was. Participants were not explicitly instructed to draw such “fuzzy” objects, so these
455 occurred spontaneously by the participant. Finally, in 16 drawings (13 inconsistent, 3
456 consistent), participants transposed the manipulated object so that it was in a different scene
457 they had viewed. All of these errors occurred significantly more frequently for inconsistent than
458 consistent scenes (Chi-squared test of proportions, Isolated Object: $\chi^2=7.11$, $p=0.008$, $\phi=0.63$;

459 Unspecified Object: $\chi^2=12.00$, $p=5.32 \times 10^{-4}$, $\phi=1.41$; Transposed Object: $\chi^2=12.50$, $p=4.07 \times$
 460 10^{-4} , $\phi=0.88$). These results imply that a disruption of scene semantics may result in a looser
 461 binding in memory of that inconsistent object with its encompassing scene.

462



463

464 **Figure 6 – Examples of memory errors made by participants for the manipulated object.** Each
 465 example is taken from a different participant. We identified three types of errors: 1) drawing the object in
 466 isolation (left), 2) drawing a detailed scene with a circle noting recollection of an unspecified object at that
 467 location but not its identity, and 3) transposing the object to a different scene. For the unspecified object
 468 errors, the text labeling the circle is included in larger font. These errors occurred overwhelmingly more
 469 often when objects were inconsistent with their background scenes ($p<0.01$ for all error types).

470

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

472 While there are differences in object memory based on scene semantics, are there also
 473 differences in spatial accuracy for the objects in the drawings? AMT workers indicated the size
 474 and location of each object in the drawing by drawing an ellipse on the drawing. With that
 475 ellipse, we calculated mean location error (centroid x and y error) and size error (radius x and y
 476 error) for each object. In both conditions, low amounts of spatial error were found, although
 477 errors were larger in magnitude in the Y-direction than the X-direction. Errors of object location

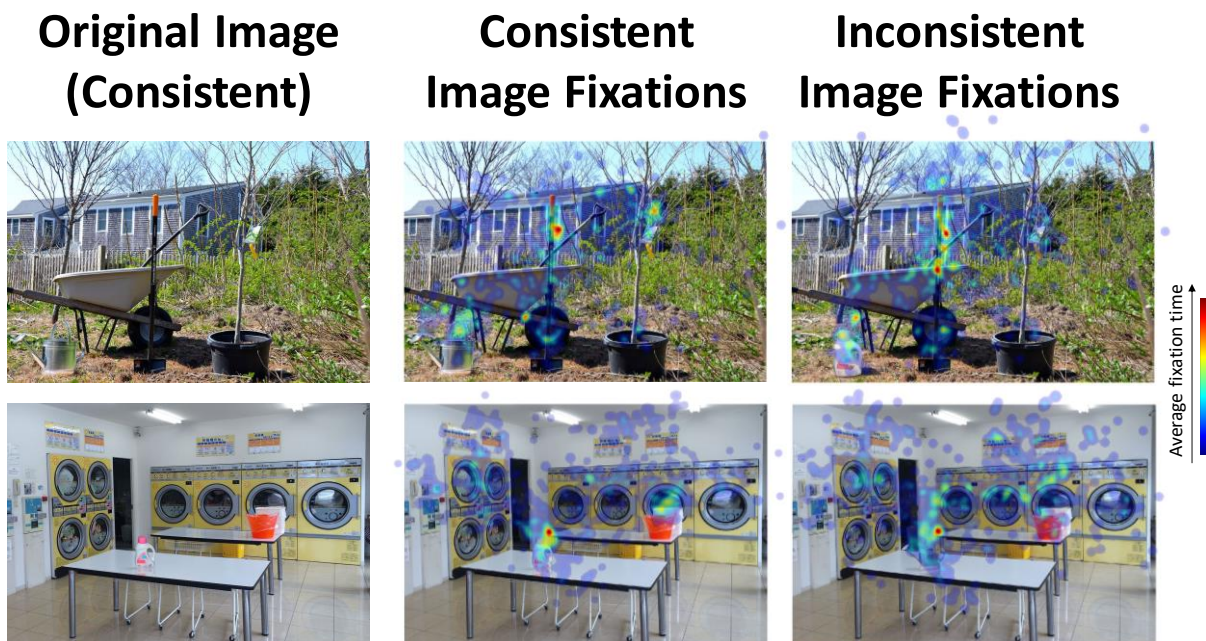
478 were transpositions of less than 11% of the size of the entire image (X-direction: Consistent
479 M=2.2%, Inconsistent M=0.4%, Y-direction: Consistent M=9.3%, Inconsistent M=10.8%). Errors
480 in size were on average less than 4% of an image's pixels (Width: Consistent M=2.2%,
481 Inconsistent M=2.0%; Height: Consistent M=2.9%, Inconsistent M=3.7%). For the manipulated
482 object, there was no significant difference between consistent and inconsistent drawings in
483 spatial accuracy, neither in terms of location accuracy (X-direction: $t(27)=0.94$, $p=0.357$; Y-
484 direction: $t(27)=1.24$, $p=0.226$), nor object size (Width: $t(27)=0.51$, $p=0.613$; Height: $t(27)=1.34$,
485 $p=0.191$). There were also no differences between conditions in accuracy for location or size of
486 the other objects in the scene (X-location: $t(28)=1.18$, $p=0.249$; Y-location: $t(28)=0.93$, $p=0.361$;
487 Width: $t(28)=1.24$, $p=0.227$; Height: $t(28)=1.43$, $p=0.163$). These results indicate that
488 manipulations of object semantics do not appear to affect spatial accuracy in memory.

489

490 **Comparing eye fixations, visual saliency, and recall**

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

497



498

499 **Figure 7 – Example average fixation heatmaps.** Heatmaps showing examples of average fixation time

500 (averaged across participants) for the consistent and inconsistent versions of two paired scenes where

501 their objects were swapped (i.e., a watering can or laundry detergent in a backyard scene or a laundry

502 scene). Red indicates higher total fixation time on average, and blue indicates lower total fixation time.

503 For the backyard scene, the inconsistent detergent bottle causes more fixations than the consistent

504 watering can. For the laundry scene, both the watering can and detergent bottle elicit more fixations.

505 Fixation heatmaps are generated here for visualization purposes and were creating using EyeLink's Data

506 Viewer. However, analyses were conducted at the level of individual fixations, without smoothing.

507

508 We looked at fixation time for each participant for each object during the study phase,

509 where they viewed each image for 10 s. On average, participants fixated for significantly longer

510 on the inconsistent manipulated object than the consistent manipulated object (Inconsistent:

511 $M=1271.7$ ms, $SD=553.3$ ms; Consistent: 725.4 ms, $SD=365.4$ ms; $t(26)=5.44$, $p=1.05 \times 10^{-5}$,512 $d=1.17$; three participants were not measured as fixating on the manipulated object). For the

513 other objects in the scenes, participants spent numerically more time looking at them in the

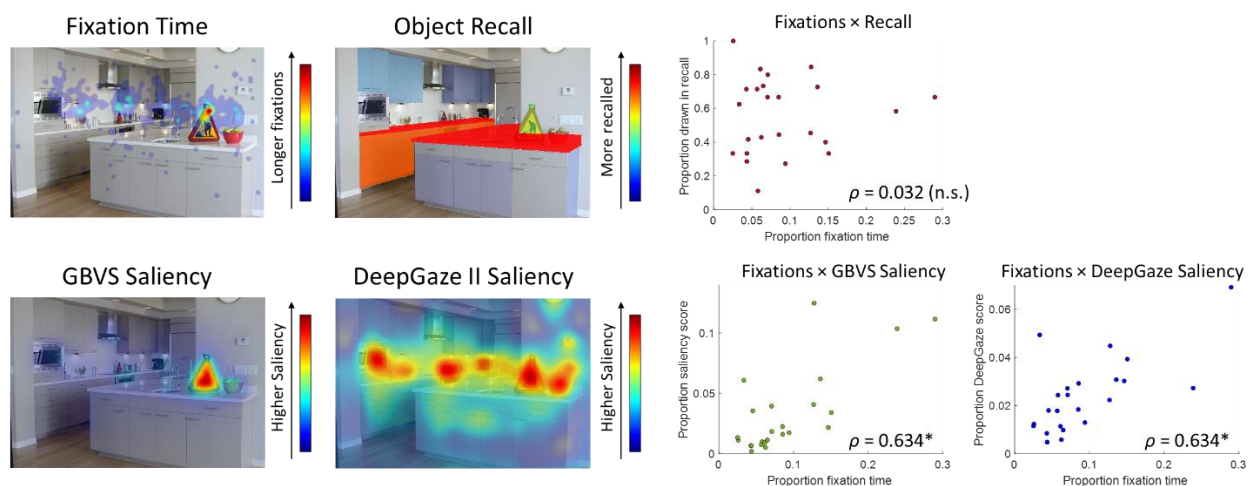
514 consistent condition than the inconsistent condition, but this difference was not statistically

515 significant (Inconsistent: $M=6439.8$ ms total across all other objects, $SD=1126.8$ ms; Consistent:
516 $M=6956.5$ ms, $SD=1101.5$ ms; $t(29)=1.97$, $p=0.059$). There was also no difference in time spent
517 fixating non-object regions of the image (Inconsistent: $M=1592.2$ ms, $SD=1823.8$ ms;
518 Consistent: $M=1592.2$ ms, $SD=1682.1$ ms; $t(29)=0.16$, $p=0.871$). Thus there is no clear
519 evidence that increased fixations on the inconsistent object detracted from fixations on other
520 objects, preventing their encoding into memory.

521 We then looked to see whether current state-of-the-art visual saliency algorithms
522 DeepGaze II and GBVS could predict these fixation times (see Methods). Collapsing across
523 conditions, we found a significant correlation between DeepGaze-predicted saliency and fixation
524 time on the manipulated objects (Spearman's rank correlation: $\rho=0.634$, $p=0.001$) as well as the
525 same correlation for GBVS-predicted saliency and fixation time ($\rho=0.634$, $p=0.001$). However,
526 there was no significant difference in either saliency score measure between inconsistent and
527 consistent objects in the same scene (DeepGaze: $t(11)=0.48$, $p=0.643$; GBVS: $t(11)=1.42$,
528 $p=0.182$). These results indicate that visual saliency may be able to partially account for fixation
529 durations, but it does not show a clear relationship to semantic consistency; we discuss these
530 implications later in the Discussion.

531 Next, we investigated whether these metrics could predict the proportion of people who
532 recalled the manipulated object (Figure 8). We observed no significant correlation between
533 mean fixation time across participants and recall proportion for each manipulated object
534 ($\rho=0.22$, $p=0.308$). As a secondary analysis, we conducted an ANOVA across all manipulated
535 objects, to see whether fixation time differed based on two factors: 1) whether than object was
536 in a consistent or inconsistent scene, and 2) whether that object was recalled or not. For fixation
537 time, we replicated our significant effect of consistency, where inconsistent objects were fixated
538 longer ($F(1,229)=14.30$, $p=1.99 \times 10^{-4}$, $\eta^2=0.06$). We also observed significantly higher fixations
539 for objects that were recalled than those that were forgotten ($F(1,229)=4.25$, $p=0.001$, $\eta^2=0.02$;
540 Recalled: $M=1308.5$ ms, $SD=1153.3$ ms; Forgotten: $M=923.5$ ms, $SD=862.9$ ms), although we

541 observed no significant interaction ($F(1,229)=0.39, p=0.53$). Thus, fixation times do show some
 542 relationship to recall success, although this does not appear to be modulated by the consistency
 543 of an object with its scene. We also investigated the relationship of computational visual
 544 saliency to recall performance. We observed no significant correlation between recall proportion
 545 and DeepGaze-predicted saliency ($\rho=0.137, p=0.524$), nor GBVS-predicted saliency ($\rho=0.201,$
 546 $p=0.346$). For DeepGaze saliency, an ANOVA showed no significant difference between
 547 consistent and inconsistent objects ($p=0.477$), nor recalled or forgotten objects ($p=0.525$), nor a
 548 statistical interaction ($p=0.284$). Similarly, an ANOVA for GBVS saliency showed no significant
 549 difference between consistent and inconsistent scenes ($p=0.244$), nor recalled or forgotten
 550 objects ($p=0.714$), nor a statistical interaction ($p=0.455$). Thus, it does not appear that
 551 inconsistent objects were much more visually salient than consistent objects, and importantly,
 552 image-based saliency cannot account for differences in memory performance between the
 553 consistent and inconsistent objects.



554
 555 **Figure 8 – Comparison of fixation time during study, visual saliency, and recall success.** (Left) For
 556 each stimulus image, we looked at four types of information: 1) eye fixation times across the pixels of the
 557 image, 2) proportion of participants recalling each object in the image, 3) visual saliency of the image
 558 calculated using Graph-Based Visual Saliency (Harel et al., 2007), and 4) visual saliency of the image
 559 calculated using DeepGaze II (Kümmerer et al., 2016). (Right) Scatterplots of average fixation time with

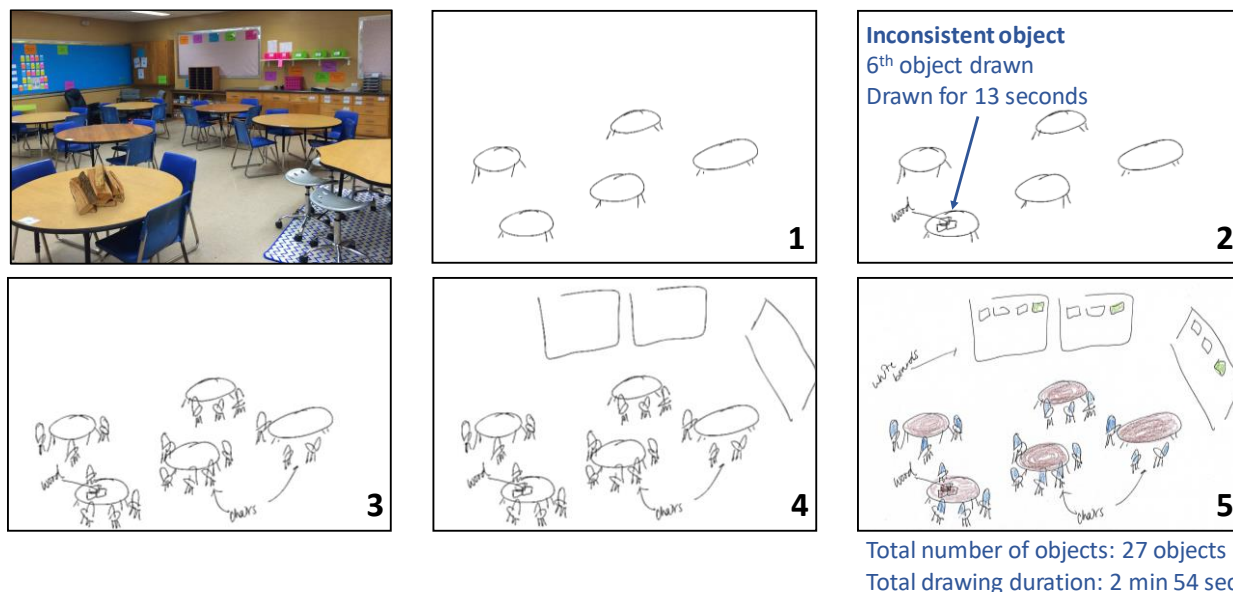
560 the three other metrics (recall proportion, GBVS visual saliency, and DeepGaze II visual saliency). Each
561 point represents one of the 24 stimulus images, and indicates the average score for the manipulated
562 object. While saliency metrics were significantly correlated with fixation time, no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

565

566 **Comparing the temporal order of recall for consistent and inconsistent scenes**

567 In conjunction with recording eye movements during study of the images, we also
568 recorded real-time pen movements during recall of the images (Figure 9). Participants did not
569 spend a significantly different amount of time drawing the inconsistent versus consistent
570 manipulated objects (Inconsistent: $M=16.08$ s, $SD=7.75$; Consistent: $M=15.85$ s, $SD=11.26$;
571 $t(27)=0.07$, $p=0.942$), nor a significantly different amount of time drawing inconsistent versus
572 consistent images (Inconsistent: $M=2.03$ min, $SD=0.62$; Consistent: $M=2.28$ min, $SD=0.89$;
573 $t(27)=1.53$, $p=0.139$). There was also no significant difference in the order in which inconsistent
574 versus consistent objects were drawn ($t(25)=1.01$, $p=0.323$). It was thus not the case that
575 inconsistent objects were drawn for longer or drawn earlier. There was also no significant
576 correlation at the level of the participant between amount of time spent fixating the manipulated
577 object and amount of time drawing the manipulated object (Spearman's rank correlation:
578 $\rho=0.21$, $p=0.130$). Similarly, there was no significant correlation between fixation order and
579 drawing order for the manipulated object ($\rho=0.09$, $p=0.449$). Thus, time spent during the
580 drawing recall phase does not reveal clear differences between the inconsistent and consistent
581 images, nor a clear relationship to fixations during study.

582



583

584 **Figure 9 – Example of pen tracking output and recall order analyses.** For each drawing, the pen
 585 tablet recorded a video of the pen strokes in order. This figure shows 5 ordered example frames from one
 586 drawing video of an inconsistent classroom scene (with a pile of logs). For the manipulated object, a
 587 scorer noted the order in which the object was drawn (normalized by total number of objects), and the
 588 length of time it was drawn for (normalized by total time spent on the drawing).

589

590 **Post-drawing recognition task performance**

591 Finally, we also tested participants' memory for the items using a visual recognition task
 592 following the drawing recall task. When tested for their recognition of each background scene
 593 (with the manipulated object concealed by a gray circle), participants had very high recognition
 594 accuracy regardless of whether the scene was originally consistent or inconsistent (Inconsistent:
 595 Mean hit rate=92.2%, SD=11.4%; Consistent: M=88.9%, SD=13.4%), with no significant
 596 difference between the groups ($t(29)=1.00$, $p=0.326$). There was also no significant difference in
 597 false recognitions of matched foil images from the same scene category (Inconsistent: Mean
 598 false alarm rate=11.2%, SD=14.8%; Consistent: M=10.0%, SD=11.2%; $t(29)=0.45$, $p=0.656$).
 599 Participants then were presented with four possible objects to fill in the obscured part of the
 600 image: 1) the inconsistent exemplar, 2) the consistent exemplar, 3) a different exemplar image

601 from the inconsistent object category, 4) a different exemplar image from the consistent object
602 category. There was no significant difference between inconsistent and consistent scenes in
603 participants being able to choose the correct item out of the four options (Inconsistent: Mean hit
604 rate=28.2%, SD=20.9%; Consistent: M=26.5%, SD=14.6%; $t(28)=0.39$, $p=0.697$). There were
605 also no significant differences in the types of errors made by participants. Participants across
606 groups were equally likely to choose an object that appeared in another image (Inconsistent:
607 M=27.1%, SD=19.1%; Consistent: M=23.5%, SD=18.5%; $t(28)=0.79$, $p=0.434$), or an object of
608 the correct category but an incorrect exemplar (Inconsistent: M=18.3%, SD=16.6%; Consistent:
609 M=24.8%, SD=18.3%; $t(28)=1.34$, $p=0.192$). Thus, while we measured differences in recall
610 performance in the drawing task, clear differences in recognition performance did not appear.
611 That being said, participants reported this object recognition task was difficult (occurring after
612 the relatively effortful drawing recall task and with very closely-matched foil objects), and
613 performance was relatively low.

614

615

Discussion

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

624 These results provide important evidence on the impact of object-scene semantics on
625 memory (Friedman, 1979; Pedzek et al., 1979; Hollingworth et al., 2001; Draschkow & Vö,
626 2017; Cornelissen & Vö, 2017). Using drawing as a memory output allows for a fine-grained

627 look at how object-scene semantics influence memory representations, and we observe a
628 nuanced trade-off in which memory for the overall image is better, but memory for the objects
629 within it is worse. This dual result could account for the fact that some work had previously
630 observed diminished memory for inconsistent images (Draschkow & Vö, 2017) while others
631 observed improved memory (Friedman, 1979; Pedzek et al., 1989; Hollingworth et al., 2001). In
632 fact, both effects may be occurring simultaneously at different levels of stimulus information (i.e.,
633 the image, the objects, and the background scene). This simultaneous effect may be akin to the
634 trade-off of capacity and precision observed in visual working memory (Roggeman et al., 2014),
635 in which inconsistent scene semantics may result in higher capacity with less precision. Our
636 findings also provide evidence for the contextual guidance model suggesting two parallel
637 pathways for scene processing: one for gist-based global information and one for object-based
638 local information (Torralba et al., 2006; Vö & Wolfe, 2015), as well as fuzzy-trace theory, which
639 posits parallel storage and dissociable retrieval of verbatim versus gist information, which has
640 been shown to parsimoniously account for several other findings of false memories (Brainerd &
641 Reyna, 2002). While scene-object inconsistencies may result in a distinctive scene with boosted
642 memory for the gist of the scene, they may prevent the ability to use a scene template to fill in
643 local, precise object details (Malcolm & Henderson, 2009; Hollingworth, 2009). A disruption of
644 the scene-object semantics may also result in looser binding of an object to its scene, resulting
645 in a “spotlighting” on the inconsistent object (Cornelissen & Vö, 2017), and a tendency to
646 migrate objects across memory episodes (Hannigan & Reinitz, 2003). Within memory,
647 semantically inconsistent objects may impair abstraction of the scene from a schema template
648 (Hock & Schmelkopf, 1980; Intraub, 1997), resulting in a loss of schema-coherent details. While
649 the current study focused on the consistency of a single object with its greater scene,
650 investigations of recall for more complex semantic manipulations (e.g., manipulating the
651 semantic relationships of the objects to each other) may provide further insight on how
652 semantics during perception influence the memory representation for a scene.

653 Considering the role of scene-object consistencies on memory has important real world
654 implications in how we design scenes, and how we test memory. Some of the first seminal work
655 looking at scene consistency and memory tested memory for real graduate student offices
656 (Pezdek et al., 1989), and recent work has brought questions about scene memory into virtual
657 reality (Helbing et al., 2020). It will be exciting to see whether our findings can help guide the
658 design of real-world scenes, based on what aspects we wish to be memorable (Bainbridge,
659 2019): a key object, all objects, or the gist of the scene. In some cases, one may want to
660 enhance a specific object even at the cost of surrounding objects being forgotten, while in other
661 cases, the goal will be to make an entire landscape memorable. Drawing is also a task that has
662 been historically used as a clinical tool to measure patient groups (Rey, 1941; Corkin, 2002),
663 and recent work has applied these same drawing quantification techniques to aphantasia, a
664 condition of absent visual imagery (Bainbridge et al., 2021). The current task manipulating
665 scene grammar could potentially reveal insight into groups with differing abilities at visual,
666 semantic, or mnemonic processing, such as individuals across the lifespan.

667 While we observed differences in recall for consistent and inconsistent scenes, we also
668 observed several similarities between memory representations for consistent and inconsistent
669 scenes. Between these two conditions, recalled drawings tended to be equally diagnostic, have
670 equally high spatial accuracy (in terms of both object location and size), and equally rare
671 numbers of additional objects inserted into the drawings. We also did not observe differences
672 between the two conditions in visual recognition performance (although this could be due to the
673 difficulty of the recognition task). Thus, while scene semantics may influence some aspects of a
674 memory (e.g., memory for other objects in an image), it may have less of a sway on other
675 aspects of that memory (e.g., spatial accuracy). Indeed, various work has suggested differences
676 in how object and spatial information may be coded in memory (Farah & Hammond, 1988;
677 Staresina et al., 2011; Bainbridge, et al., 2021). While the current work investigates scene
678 semantics, other work has suggested that scene *syntax*—the spatial arrangement of semantically

679 consistent objects within a scene—as similarly meaningful organizational principles for scenes
680 (Võ et al., 2019). An experiment manipulating scene syntax rather than semantics (e.g., moving
681 a consistent object to an inconsistent location) may result in higher spatial error but preserved
682 object accuracy in memory.

683 While the current study serves as important evidence towards a dissociation of scene
684 versus object memory for inconsistent scenes, some caveats of this work motivate future
685 studies. We utilized stimulus images that were manipulated to appear natural regardless of the
686 object consistency. However, future studies could explore similar methods using stimuli that
687 prioritize systematic manipulation of the images (e.g., keeping object size, location, lighting, and
688 shadowing consistent across all images), rather than naturalness of the stimuli. Further, while
689 we decided *a priori* on a sample size validated from prior work (Bainbridge et al., 2019), some
690 results showed small to medium effect sizes or null effects that would be valuable to replicate in
691 follow-up research. A future study could also increase the number of images per participant in
692 order to look at influences of saliency and fixation patterns on within-participant recall
693 performance; however, we do note that participants may not be able to recall many more
694 images. Also, as this task required detailed memorization of the scene, all visual information
695 was highly task-relevant. However, prior research has shown that some tasks such as visual
696 search can result in higher recall performance than explicit memorization (Draschkow et al.,
697 2014). It would be interesting to see if an incidental study task would drive even stronger
698 differences between object and scene recollection. Relatedly, it would be interesting to see if
699 different explicit instructions (such as telling participants to indicate vaguely remembered
700 objects) would influence the information present in memory drawings. Finally, there are still
701 many open questions about how drawing as a recall task itself may influence remembered
702 information. Some work in children has shown that drawing of a memory can increase
703 accurately recalled information, but it also has a tendency to increase false memories (Bruck et
704 al., 2000; Otgaar et al., 2016). Other work has shown that artists are able to produce more

705 memory information than non-artists (Vogt & Magnussen, 2007; Perdreau & Cavanagh, 2015),
706 potentially suggesting that different strategies may boost performance on the task. Thus, further
707 investigation into the limitations as well as potential for drawing as a memory task will be highly
708 important in future work.

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

731 saliency models to capture semantically meaningful portions of an image (Bylinskii et al., 2016;
732 Henderson & Hayes, 2017). The current work highlights a need for image-based metrics aimed
733 at making predictions specific to scene memory, accounting for semantic abstraction of the
734 scene as well as what objects and features are memorable (Bainbridge, 2019). Future work
735 could examine scenes with graded levels of inconsistency, in order to create more nuanced
736 models that can account for both semantic inconsistency as well as visual saliency.

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

741

742 Acknowledgements

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

747

748 Open Practices Statement

749 The datasets generated by the current study are publicly available on the Open Science
750 Framework repository at
751 https://osf.io/e8qhm/?view_only=d1294e1fd68b44b3a4caff085403e527. The experiments were
752 not preregistered.

753

754 References

755

756 Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual
757 processes during object recognition. *Psychonomic Bulletin & Review*, *14*, 332-337.

758

759 Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember.
760 *Psychology of Learning and Motivation*, *70*, 1-27 (K. Federmeier & D. Beck, Eds.) Elsevier Inc.

761

762 Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free
763 recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*, 5.

764

765 Bainbridge, W.A. (2020). The resiliency of image memorability: A predictor of memory separate
766 from attention and priming. *Neuropsychologia*, *141*, 107408.

767

768 Bainbridge, W.A., Pounder, Z., Eardley, A. F., & Baker, C. I. (2021). Quantifying Aphantasia
769 through drawing: Those without visual imagery show deficits in object but not spatial memory.
770 *Cortex*, *135*, 159-172.

771

772 Bonnici, H. M., Chadwick, M. J., Kumaran, D., Hassabis, D., Weiskopf, N., & Maguire, E. A.
773 (2012). Multi-voxel pattern analysis in human hippocampal subfields. *Frontiers in Human*
774 *Neuroscience*, *6*, 290.

775

776 Brainerd, C.J., & Reyna, V.F. (2002). Fuzzy-trace theory and false memory. *Current Directions*
777 *in Psychological Science*, *11*(5), 164-169.

778

779 Brandt, S.A., & Stark, L.W. (1997). Spontaneous eye movements during visual imagery reflect
780 the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*(1), 27-38.

781

- 782 Bruck, M., Melnyk, L., & Ceci, S.J. (2000). Draw it again Sam: The effect of drawing on
783 children's suggestibility and source monitoring ability. *Journal of Experimental Child Psychology*,
784 77(3), 169-196.
785
- 786 Bylinskii, Z., Recansens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where should
787 saliency models look next? *European Conference on Computer Vision*, 809-824.
788
- 789 Corkin, S. (2002). What's new with the amnesic patient HM? *Nature Reviews Neuroscience*,
790 3(2), 153-160.
791
- 792 Cornelissen, T. H. W., & Võ, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant
793 object-scene inconsistencies modulate ongoing gaze behavior. *Attention, Perception, and*
794 *Psychophysics*, 79, 154-168.
795
- 796 Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*,
797 35(3), 393-401.
798
- 799 Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene
800 semantics interact with visual search to build better memories. *Journal of Vision*, 14, 8, 10.
801
- 802 Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects,
803 strengthens memories, and speeds search. *Scientific Reports*, 7, 16471.
804
- 805 Farah, M. J., & Hammond, K. M. (1988). Visual and spatial mental imagery: Dissociable
806 systems of representations. *Cognitive Psychology*, 20, 439-462.
807

- 808 Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-
809 world scene? *Journal of Vision*, 7(1), 1-29.
- 810
- 811 Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and
812 memory for gist. *Journal of Experimental Psychology: General*, 108, 316-355.
- 813
- 814 Greene, M. R., Botros, A. P., Beck, D. M, Fei-Fei, L. (2015). What you see is what you expect:
815 Rapid scene understanding benefits from prior experience. *Attention, Perception, and*
816 *Psychophysics*, 77, 1239-1251.
- 817
- 818 Hannigan, S. L., & Reinitz, M. T. (2003). Migration of objects and inferences across episodes.
819 *Memory & Cognition*, 31(3), 434-444.
- 820
- 821 Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural*
822 *Information Processing Systems*, 545-552.
- 823
- 824 Hayes, T.R., & Henderson, J.M. (2019). Scene semantics involuntarily guide attention during
825 visual search. *Psychonomic Bulletin & Review*, 26, 1683-1689.
- 826
- 827 Henderson, J.M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as
828 revealed by meaning maps. *Nature Human Behaviour*, 1(10), 743-747.
- 829
- 830 Henderson, J., Hayes, T., Peacock, C., & Rehrig, G. (2021). Meaning maps capture the density
831 of local semantic features in scenes: A reply to Pedziwiatr, Kümmerer, Wallis, Bethge & Teufel.
832 *OSF Preprints*.
- 833

- 834 Hock, H.S., & Schmelzkopf, K. F. (1980). The abstraction of schematic representations from
835 photographs of real-world scenes. *Memory & cognition*, 8, 543-554.
836
- 837 Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object
838 perception? *Journal of Experimental Psychology: General*, 127, 398-415.
839
- 840 Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually
841 specific information is retained in memory from previously attended objects in natural scenes.
842 *Psychonomic Bulletin and Review*, 8(4), 761-768.
843
- 844 Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene
845 context and memory for the binding of target object to scene location. *Visual Cognition*, 17(1-2),
846 273-291.
847
- 848 Hoover, M.A., & Richardson, D.C. (2008). When facts go down the rabbit hole: Contrasting
849 features and objecthood as indexes to memory. *Cognition*, 108(2), 533-542.
850
- 851 Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, 1(6),
852 217-222.
853
- 854 Jarosz, A.F., & Wiley, J. (2014). What are the odds? A practical guide to computing and
855 reporting Bayes factors. *The Journal of Problem Solving*, 7(1), 2.
856
- 857 Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in
858 memory retrieval. *Psychological Science*, 25(1) 236-242.
859

- 860 Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep
861 features trained on object recognition. *arXiv*: 1610.01563.
- 862
- 863 Laeng, B., Bloem, I. K., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The
864 role of gaze in mental imagery and memory. *Cognition*, 131(2), 263-283.
- 865
- 866 Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual
867 search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11), 8.
- 868
- 869 Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye
870 movements during real-world scene search. *Journal of Vision*, 10(2), 1-11.
- 871
- 872 Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in
873 recognition. *Progress in Brain Research*, 155, 23-36.
- 874
- 875 Otgaar, H., van Ansem, R., Pauw, C., & Horselenberg, R. (2016). Improving children's
876 interviewing methods? The effects of drawing and practice on children's memories for an event.
877 *Journal of Police and Criminal Psychology*, 31(4), 279-287.
- 878
- 879 Pedzek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-
880 world scenes: The role of consistency with schema expectation. *Journal of Experimental*
881 *Psychology: Learning, Memory, and Cognition*, 15(4), 587-595.
- 882
- 883 Pedziwiatr, M.A., Kümmerer, M., Wallis, T.S.A., Bethge, M., & Teufel, C. (2021). Meaning maps
884 and saliency models based on deep convolutional neural networks are insensitive to image
885 meaning when predicting human fixations. *Cognition*, 206, 104465.

886

887 Perdreau, F., & Cavanagh, P. (2015). Drawing experts have better visual memory while
888 drawing. *Journal of Vision*, 15(5), 5.

889

890 Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. (Les
891 problems). *Archives de psychologie*, 28, 286-340.

892

893 Roggeman, C., Klingberg, T., Feenstra, H.E.M., Compte, A., & Almeida, R. (2014). Trade-off
894 between capacity and precision in visuospatial working memory. *Journal of Cognitive
895 Neuroscience*, 26(2), 211-222.

896

897 Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and
898 web-based tool for image annotation. *International Journal of Computer vision*, 77(1-3), 157-
899 173.

900

901 Staresina, B. P., Duncan, K. D., & Davachi, L. (2011). Perirhinal and parahippocampal cortices
902 differentially contribute to later recollection of object- and scene-related event details. *Journal of
903 Neuroscience*, 31, 8739-8747.

904

905 Torralba, A., Oliva, A., Castelano, M. S., & Henderson, J. M. (2006). Contextual guidance of
906 eye movements and attention in real-world scenes: the role of global features in object search.
907 *Psychological Review*, 113(4), 766.

908

909 Võ, M. L.-H., Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the
910 New York Academy of Sciences*, 1339(1), 72.

911

- 912 Vő, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene
913 grammar guides attention and aids perception in real-world environments. *Current Opinion in*
914 *Psychology, 29*, 205-210.
- 915
- 916 Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values.
917 *Psychonomic Bulletin & Review, 14*(5), 779-704.
- 918
- 919 Zhao, Q., & Koch, S. (2013). Learning saliency-based visual attention: A review. *Signal*
920 *Processing, 93*(6), 1401-1407.