

1 Feature and order manipulations in a free recall task affect memory  
2 for current and future lists

3 Jeremy R. Manning<sup>1,\*</sup>, Emily C. Whitaker<sup>1</sup>, Paxton C. Fitzpatrick<sup>1</sup>,  
Madeline R. Lee<sup>1</sup>, Allison M. Frantz<sup>1</sup>, Bryan J. Bollinger<sup>1</sup>,  
Darya Romanova<sup>1</sup>, Campbell E. Field<sup>1</sup>, and Andrew C. Heusser<sup>1,2</sup>

<sup>1</sup>Dartmouth College

<sup>2</sup>Akili Interactive

\*Corresponding author: jeremy.r.manning@dartmouth.edu

4 **Abstract**

5 We perceive, interpret, and remember ongoing experiences through the lens of our prior  
6 experiences. Inferring that we are in one type of situation versus another can lead us to interpret  
7 the same physical experience differently. In turn, this can affect how we focus our attention,  
8 form expectations about what will happen next, remember what is happening now, draw on  
9 our prior related experiences, and so on. To study these phenomena, we asked participants  
10 to perform simple word list-learning tasks. Across different experimental conditions, we held  
11 the set of to-be-learned words constant, but we manipulated how incidental visual features  
12 changed across words and lists, along with the orders in which the words were studied. We  
13 found that these manipulations affected not only how the participants recalled the manipulated  
14 lists, but also how they recalled later (randomly ordered) lists. Our work shows how structure  
15 in our ongoing experiences can influence how we remember both our current experiences and  
16 unrelated subsequent experiences.

17 **Keywords:** episodic memory, free recall, incidental features, implicit priming, temporal  
18 **order**

## 19 Introduction

20 Experience is subjective: different people who encounter identical physical experiences  
21 can take away very different meanings and memories. One reason is that our moment-by-  
22 moment subjective experiences are shaped in part by the idiosyncratic prior experiences,  
23 memories, goals, thoughts, expectations, and emotions that we bring with us into the  
24 present moment. These factors collectively define a *context* for our experiences (Manning,  
25 2020).

26 The contexts we encounter help us to construct *situation models* (Manning et al., 2015;  
27 Radvansky and Copeland, 2006; Ranganath and Ritchey, 2012; Zwaan et al., 1995; Zwaan  
28 and Radvansky, 1998) or *schemas* (Baldassano et al., 2018; Masís-Obando et al., 2022;  
29 Tse et al., 2007) that describe how experiences are likely to unfold based on our prior  
30 experiences with similar contextual cues. For example, when we enter a sit-down restau-  
31 rant, we might expect to be seated at a table, given a menu, and served food. Priming  
32 someone to expect a particular situation or context can also influence how they resolve  
33 potential ambiguities in their ongoing experiences, including in ambiguous movies and  
34 narratives (Rissman et al., 2003; Yeshurun et al., 2017).

35 Our understanding of how we form situation models and schemas, and how they  
36 interact with our subjective experiences and memories, is constrained in part by substantial  
37 differences in how we study these processes. Situation models and schemas are most often  
38 studied using “naturalistic” stimuli such as narratives and movies (Nastase et al., 2020;  
39 Zwaan et al., 1995; Zwaan and Radvansky, 1998). In contrast, our understanding of how  
40 we organize our memories has been most widely informed by more traditional paradigms  
41 like free recall of random word lists (Kahana, 2012, 2020). In free recall, participants study  
42 lists of items and are instructed to recall the items in any order they choose. The orders  
43 in which words come to mind can provide insights into how participants have organized

44 their memories of the studied words. Because random word lists are unstructured by  
45 design, it is not clear if, or how, non-trivial situation models might apply to these stimuli.  
46 Nevertheless, there are *some* commonalities between memory for word lists and memory  
47 for real-world experiences.

48 Like remembering real-world experiences, remembering words on a studied list re-  
49 quires distinguishing the current list from the rest of one's experience. To model this  
50 fundamental memory capability, cognitive scientists have posited a special context repre-  
51 sentation that is associated with each list. According to early theories (e.g. Anderson and  
52 Bower, 1972; Estes, 1955) context representations are composed of many features which  
53 fluctuate from moment to moment, slowly drifting through a multidimensional feature  
54 space. During recall, this representation forms part of the retrieval cue, enabling us to  
55 distinguish list items from non-list items. Understanding the role of context in memory  
56 processes is particularly important in self-cued memory tasks, such as free recall, where  
57 the retrieval cue is "context" itself (Howard and Kahana, 2002a). Conceptually, the same  
58 general processes might be said to describe how real-world contexts evolve during natural  
59 experiences. However, this is still an open area of study (Manning, 2020, 2021).

60 Over the past half-century, context-based models have had impressive success at ex-  
61 plaining many stereotyped behaviors observed during free recall and other list-learning  
62 tasks (Estes, 1955; Glenberg et al., 1983; Howard and Kahana, 2002a; Kimball et al., 2007;  
63 Polyn and Kahana, 2008; Polyn et al., 2009; Raaijmakers and Shiffrin, 1980; Sederberg  
64 et al., 2008; Shankar and Howard, 2012; Sirotin et al., 2005). These phenomena include  
65 the well known recency and primacy effects (superior recall of items from the end and,  
66 to a lesser extent, from the beginning of the study list), as well as semantic and temporal  
67 clustering effects (Howard and Kahana, 2002b; Kahana et al., 2008). The contiguity effect  
68 is an example of temporal clustering, which is perhaps the dominant form of organization

69 in free recall. This effect can be seen in people’s tendencies to successively recall items that  
70 occupied neighboring positions in the studied list (Kahana, 1996). There are also striking  
71 effects of semantic clustering (Bousfield, 1953; Bousfield et al., 1954; Jenkins and Russell,  
72 1952; Manning and Kahana, 2012; Romney et al., 1993), whereby the recall of a given item  
73 is more likely to be followed by recall of a similar or related item than a dissimilar or  
74 unrelated one. In general, people organize memories for words along a wide variety of  
75 stimulus dimensions. As formalized by models like the *Context Maintenance and Retrieval*  
76 *Model* (Polyn et al., 2009), the stimulus features associated with each word (e.g. the word’s  
77 meaning, size of the object the word represents, the letters that make up the word, font  
78 size, font color, location on the screen, etc.) are incorporated into the participant’s mental  
79 context representation (Manning, 2020; Manning et al., 2015, 2011, 2012; Smith and Vela,  
80 2001). During a memory test, any of these features may serve as a memory cue, which in  
81 turn leads the participant to recall in succession words that share stimulus features.

82 A key mystery is whether (and how) the sorts of situation models and schemas that  
83 people use to organize their memories of real-world experiences might map onto the  
84 clustering effects that reflect how people organize their memories for word lists. On  
85 one hand, both situation models and clustering effects reflect statistical regularities in  
86 ongoing experiences. Our memory systems exploit these regularities when generating  
87 inferences about the unobserved past and yet-to-be-experienced future (Bower et al., 1979;  
88 Momennejad et al., 2017; Ranganath and Ritchey, 2012; Schapiro and Turk-Browne, 2015;  
89 Xu et al., 2023). On the other hand, the rich structures of real-world experiences and other  
90 naturalistic stimuli that enable people to form deep and meaningful situation models and  
91 schemas have no obvious analogs in simple word lists. Often, lists in free recall studies are  
92 explicitly *designed* to be devoid of exploitable temporal structure, for example, by sorting  
93 the words in a random order (Kahana, 2012).

94 We designed an experimental paradigm to explore how people organize their mem-  
95 ories for simple stimuli (word lists) whose temporal properties change across different  
96 “situations,” analogous to how the content of real-world experiences change across dif-  
97 ferent real-world situations. We asked participants to study and freely recall a series of  
98 word lists (Fig. 1). In the different conditions in our experiment, we varied the lists’  
99 appearances and presentation orders in different ways. The studied items (words) were  
100 designed to vary along three general dimensions: semantic (word *category* and physical  
101 *size* of the referent), lexicographic (word *length* and *first letter*), and visual (font *color* and  
102 the onscreen *location* of each word). We used two control conditions as a baseline; in  
103 these control conditions all of the lists were sorted randomly, but we manipulated the  
104 presence or absence of the visual features. In two conditions, we manipulated whether  
105 the words’ appearances were fixed or variable within each list. In six conditions, we asked  
106 participants to first study and recall eight lists whose items were sorted by a target feature  
107 (e.g., word category), and then study and recall an additional eight lists whose items had  
108 the same features, but that were sorted in a random temporal order. We were interested  
109 in how these manipulations affected participants’ recall behaviors on early (manipulated)  
110 lists, as well as how order manipulations on early lists affected recall behaviors on later  
111 (randomly ordered) lists. Finally, in an *adaptive* experimental condition we used partici-  
112 pants’ recall behaviors on early lists to manipulate, in real-time, the presentation orders  
113 of subsequent lists. In this adaptive condition, we varied the agreement between how  
114 participants preferred to organize their memories of the studied items versus the orders  
115 in which the items were presented.

## 116 **Materials and methods**

### 117 **Participants**

118 We enrolled a total of 491 members of the Dartmouth College community across 11 exper-  
119 imental conditions. The conditions included two controls (feature rich and reduced), two  
120 visual manipulation conditions [reduced (early) and reduced (late)], six order manipula-  
121 tion conditions (category, size, length, first letter, color, and location), and a final adaptive  
122 condition. Each of these conditions is described in the *Experimental design* subsection  
123 below.

124 Participants either received course credit or a one-time \$10 payment for enrolling in  
125 our study. We asked each participant to fill out a demographic survey that included  
126 questions about their age, gender, ethnicity, race, education, vision, reading impairments,  
127 medications or recent injuries, coffee consumption on the day of testing, and level of  
128 alertness at the time of testing. All components of the demographics survey were optional.  
129 One participant elected not to fill out any part of the demographic survey, and all other  
130 participants answered some or all of the survey questions.

131 We aimed to run (to completion) at least 60 participants in each of the two primary  
132 control conditions and in the adaptive condition. In all of the other conditions, we set a  
133 target enrollment of at least 30 participants. Because our data collection procedures en-  
134 tailed the coordinated efforts of 12 researchers and multiple testing rooms and computers,  
135 it was not feasible for individual experimenters to know how many participants had been  
136 run in each experimental condition until the relevant databases were synchronized at the  
137 end of each working day. We also over-enrolled participants for each condition to help  
138 ensure that we met our minimum enrollment targets even if some participants dropped  
139 out of the study prematurely or did not show up for their testing session. This led us to

140 exceed our target enrollments for several conditions. Nevertheless, we analyze all viable  
141 data in the present paper.

142 Participants were assigned to experimental conditions based loosely on their date of  
143 participation. (This aspect of our procedure helped us to more easily synchronize the ex-  
144 periment databases across multiple testing computers.) Of the 490 participants who opted  
145 to fill out the demographics survey, reported ages ranged from 17 to 31 years (mean: 19.1  
146 years; standard deviation: 1.356 years). A total of 318 participants reported their gender as  
147 female, 170 as male, and two participants declined to report their gender. A total of 442 par-  
148 ticipants reported their ethnicity as “not Hispanic or Latino,” 39 as “Hispanic or Latino,”  
149 and nine declined to report their ethnicity. Participants reported their races as White (345  
150 participants), Asian (120 participants), Black or African American (31 participants), Amer-  
151 ican Indian or Alaska Native (11 participants), Native Hawaiian or Other Pacific Islander  
152 (four participants), Mixed race (three participants), Middle Eastern (one participant), and  
153 Arab (one participant). A total of five participants declined to report their race. We note  
154 that several participants reported more than one of the above racial categories. Participants  
155 reported their highest degrees achieved as “Some college” (359 participants), “High school  
156 graduate” (117 participants), “College graduate” (seven participants), “Some high school”  
157 (five participants), “Doctorate” (one participant), and “Master’s degree” (one participant).  
158 A total of 482 participants reported no reading impairments, and eight reported having  
159 mild reading impairments. A total of 489 participants reported having normal color vision  
160 and one participant reported that they were red-green color blind. A total of 482 partic-  
161 ipants reported taking no prescription medications and having no recent injuries; four  
162 participants reported having ADHD, one reported having dyslexia, one reported having  
163 allergies, one reported a recently torn ACL/MCL, and one reported a concussion from  
164 several months prior. The participants reported consuming 0–3 cups of coffee prior to the

165 testing session (mean: 0.32 cups; standard deviation: 0.58 cups). Participants reported  
166 their current level of alertness, and we converted their responses to numerical scores as  
167 follows: “very sluggish” (-2), “a little sluggish” (-1), “neutral” (0), “a little alert” (1), and  
168 “very alert” (2). Across all participants, the full range of alertness levels were reported  
169 (range: -2–2; mean: 0.35; standard deviation: 0.89).

170 We dropped from our dataset the one participant who reported having abnormal color  
171 vision, as well as 38 participants whose data were corrupted due to technical failures while  
172 running the experiment or during the daily database merges. In total, this left usable data  
173 from 452 participants, broken down by experimental condition as follows: feature rich (67  
174 participants), reduced (61 participants), reduced (early) (42 participants), reduced (late)  
175 (41 participants), category (30 participants), size (30 participants), length (30 participants),  
176 first letter (30 participants), color (31 participants), location (30 participants), and adaptive  
177 (60 participants). The participant who declined to fill out their demographic survey  
178 participated in the location condition, and we verified verbally that they had normal color  
179 vision and no significant reading impairments.

## 180 **Experimental design**

181 Our experiment is a variant of the classic free recall paradigm that we term “*feature-rich free*  
182 *recall*.” In feature-rich free recall, participants study 16 lists, each comprised of 16 words  
183 that vary along a number of stimulus dimensions (Fig. 1). The stimulus dimensions include  
184 two semantic features related to the *meanings* of the words (semantic category, referent  
185 object size), two lexicographic features related to the *letters* that make up the words (word  
186 length in number of letters, identity of the word’s first letter), and two visual features  
187 that are independent of the words themselves (text color, presentation location). Each  
188 list contains four words from each of four different semantic categories, with two object



189 sizes reflected across all of the words. After studying each list, the participant attempts  
190 to recall as many words as they can from that list, in any order they choose. Because  
191 each individual word is associated with several well defined (and quantifiable) features,  
192 and because each list incorporates a diverse mix of feature values along each dimension,  
193 this allows us to estimate which features participants are considering or leveraging in  
194 organizing their memories.

## 195 **Stimuli**

196 The stimuli in our paradigm were 256 English words selected in a previous study (Ziman  
197 et al., 2018). The words all referred to concrete nouns, and were chosen from 15 unique se-  
198 mantic categories: body parts, building-related, cities, clothing, countries, flowers, fruits,  
199 insects, instruments, kitchen-related, mammals, (US) states, tools, trees, and vegetables.  
200 We also tagged each word according to the approximate size of the object the word referred  
201 to. Words were labeled as “small” if the corresponding object was likely able to “fit in  
202 a standard shoebox” or “large” if the object was larger than a shoebox. Most semantic  
203 categories comprised words that reflected both “small” and “large” object sizes, but sev-  
204 eral included only one or the other (e.g., all countries, US states, and cities are larger than  
205 a shoebox; mean number of different sizes per category: 1.33; standard deviation: 0.49).  
206 The numbers of words in each semantic category also varied from 12–28 (mean number of  
207 words per category: 17.07; standard deviation number of words: 4.65). We also identified  
208 lexicographic features for each word, including the words’ first letters and lengths (i.e.,  
209 number of letters). Across all categories, all possible first letters were represented except  
210 for ‘Q’ (average number of unique first letters per category: 11; standard deviation: 2  
211 letters). Word lengths ranged from 3–12 letters (average: 6.17 letters; standard deviation:  
212 2.06 letters).



**Figure 1: Feature-rich free recall.** After studying lists comprised of words that vary along several feature dimensions, participants verbally recall words in any order (microphone icon). Each experimental condition manipulates word features and/or presentation orders within and/or across lists. The rows display representative (illustrated) examples of items from the first list participants might encounter in each condition. The rectangles during the “Presentation phase” show illustrated screen captures during a series of word presentations. Each word appeared onscreen for 2 seconds, followed by 2 seconds of blank screen. The red microphone icons during the “Recall” phase denote the one minute verbal recall interval. The labels on the right (and corresponding groupings on the left) denote experimental condition labels.

213 We assigned the categorized words into a total of 16 lists with several constraints. First,  
214 we required that each list contained words from exactly four unique categories, each with  
215 exactly four exemplars from each category. Second, we required that (across all words  
216 on the list) at least one instance of both object sizes were represented. On average, each  
217 category was represented in 4.27 lists (standard deviation: 1.16 lists). Aside from these  
218 two constraints, we assigned each word to a unique list. After random assignment, each  
219 list contained words with an average of 11.13 unique starting letters (standard deviation:  
220 1.15 letters) and an average word length of 6.17 letters (standard deviation: 0.34 letters).

221 The above assignments of words to lists was performed once across all participants,  
222 such that every participant studied the same set of 16 lists. In every condition we random-  
223 ized the study order of these lists across participants. For participants in most conditions,  
224 on some or all of the lists, we also randomly varied two additional visual features associ-  
225 ated with each word: the presentation font color, and the word’s onscreen location. These  
226 attributes were assigned independently for each word (and for every participant). These  
227 visual features were varied for words in all lists and conditions except for the “reduced”  
228 condition (all lists), the first eight lists of the “reduced (early)” condition, and the last eight  
229 lists of the “reduced (late)” condition. In these latter cases, words were all presented in  
230 black at the center of the experimental computer’s display.

231 To select a random font color for each word, we drew three integers uniformly and  
232 at random from the interval [0,255], corresponding to the red (r), green (g), and blue  
233 (b) color channels for that word. To assign random presentation locations to each word,  
234 we selected two floating point numbers uniformly and at random (one for the word’s  
235 horizontal  $x$ -coordinate and the other for its vertical  $y$ -coordinate). The bounds of these  
236 coordinates were selected to cover the entire visible area of the display without cutting off  
237 any part of the words. The words were shown on 27-in (diagonal) Retina 5K iMac displays

238 (resolution:  $5120 \times 2880$  pixels).

239 Most of the experimental manipulations we carried out entailed presenting or sorting  
240 the presented words differently on the first eight lists participants studied (which we call  
241 *early* lists) versus on the final eight lists they studied (*late* lists). Since every participant  
242 studied exactly 16 lists, every list was either “early” or “late” depending on its order in  
243 the list study sequence.

#### 244 **Real-time speech-to-text processing**

245 Our experimental paradigm incorporates the Google Cloud Speech API speech-to-text en-  
246 gine (Halpern et al., 2016) to automatically transcribe participants’ verbal recalls into text.  
247 This allows recalls to be transcribed in real time—a distinguishing feature of the experi-  
248 ment; in typical verbal recall experiments, the audio data must be parsed and transcribed  
249 manually. In prior work, we used a similar experimental setup (equivalent to the “re-  
250 duced” condition in the present study) to verify that the automatically transcribed recalls  
251 were sufficiently close to human-transcribed recalls to yield reliable data (Ziman et al.,  
252 2018). This real-time speech processing component of the paradigm plays an important  
253 role in the “adaptive” condition of the experiment, as described below.

#### 254 **Random conditions (Fig. 1, top four rows)**

255 We used two “control” conditions to evaluate and explore participants’ baseline behaviors.  
256 We also used performance on these control conditions to help interpret performance in  
257 other “manipulation” conditions. In the first control condition, which we call the *feature*  
258 *rich* condition, we randomly shuffled the presentation order (independently for each  
259 participant) of the words on each list. In the second control condition, which we call the  
260 *reduced* condition, we randomized word presentations as in the feature rich condition.

261 However, rather than assigning each word a random color and location, we instead  
262 displayed all of the words in black and at the center of the screen.

263 We also designed two conditions where we varied the words' visual appearances across  
264 lists. In the *reduced (early)* condition, we followed the "reduced" procedure (presenting  
265 each word in black at the center of the screen) for early lists, and followed the "feature rich"  
266 procedure (presenting each word in a random color and location) for late lists. Finally, in  
267 the *reduced (late)* condition, we followed the feature rich procedure for early lists and the  
268 reduced procedure for late lists.

#### 269 **Order manipulation conditions (Fig. 1, middle six rows)**

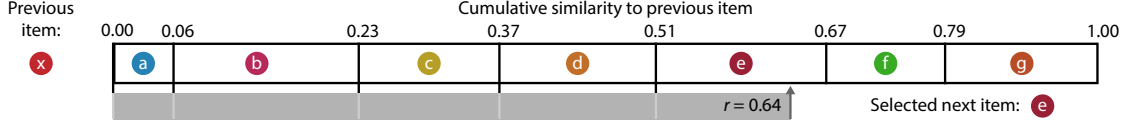
270 Each of six *order manipulation* conditions used a different feature-based sorting procedure  
271 to order words on early lists, where each sorting procedure relied on one relevant feature  
272 dimension. All of the irrelevant features varied freely across words on early lists, in that  
273 we did not consider irrelevant features in ordering the early lists. However, we note that  
274 some features were correlated—for example, some semantic categories of words referred  
275 to objects that tended to be a particular size, which meant that category and size were not  
276 fully independent. On late lists, the words were always presented in a randomized order  
277 (chosen anew for each participant). In all of the order manipulation conditions, we varied  
278 words' font colors and onscreen locations, as in the feature rich condition.

279 **Defining feature-based distances.** Sorting words according to a given relevant feature  
280 requires first defining a distance function for quantifying the dissimilarity between each  
281 pair of features. This function varied according to the type of feature under consideration.  
282 Semantic features (category and size) are *categorical*. For these features, we defined a  
283 binary distance function: two words were considered to "match" (i.e., have a distance of  
284 0) if their labels were the same (i.e., both from the same semantic category or both of the

285 same size). If two words' labels were different for a given feature, we defined the words  
 286 to have a distance of 1 for that feature. Lexicographic features (length and first letter)  
 287 are *discrete*. For these features we defined a discrete distance function. Specifically, we  
 288 defined the distance between two words as either the absolute difference between their  
 289 lengths, or the absolute distance between their starting letters in the English alphabet,  
 290 respectively. For example, two words that started with the same letter would have a "first  
 291 letter" distance of 0, and a pair of words starting with 'J' and 'A' would have a first letter  
 292 distance of 9. Because words' lengths and letters' positions in the alphabet are always  
 293 integers, these discrete distances always take on integer values. Finally, the visual features  
 294 (color and location) are *continuous* and *multivariate*, in that each "feature" is defined by  
 295 multiple (positive) real values. We defined the "color" and "location" distances between  
 296 two words as the Euclidean distances between their  $(r, g, b)$  color or  $(x, y)$  location vectors,  
 297 respectively. Therefore, the color and location distance measures always take on non-  
 298 negative real values (upper-bounded at 441.67 for color, or 27 in for location, reflecting the  
 299 distances between the corresponding maximally different vectors).

300 **Constructing feature-sorted lists.** Given a list of words, a relevant feature, and each  
 301 word's value(s) for that feature, we developed a stochastic algorithm for (noisily) sorting  
 302 the words. The stochastic aspect of our sorting procedure enabled us to obtain unique  
 303 orderings for each participant. First, we choose a word uniformly and at random from  
 304 the set of words on the to-be-presented list. Second, we compute the distances between  
 305 the chosen word's feature(s) and the corresponding feature(s) of all yet-to-be-presented  
 306 words. Third, we convert these distances (between the previously presented word's  
 307 feature values,  $a$ , and the candidate word's feature values,  $b$ ) to similarity scores:

$$\text{similarity}(a, b) = \exp\{-\tau \cdot \text{distance}(a, b)\}, \quad (1)$$



**Figure 2: Generating stochastic feature-sorted lists.** For a given feature dimension (e.g., color), we compute the similarity (Eqn. 1) between the feature value(s) of the previous item,  $x$ , and all yet-to-be-presented items ( $a$ – $g$ ). Next, we normalize these similarity scores so that they sum to 1. We lay, in sequence, a set of “sticks,” one for each candidate item, whose lengths are equal to these normalized similarity scores. To select the next to-be-presented item, we draw a random number,  $r$ , from the uniform distribution bounded between 0 and 1 (inclusive). The identity of the next item is given by the stick adjacent to an indicator that moves distance  $r$  (starting from 0) along the sequence of sticks. In this case, the next to-be-presented item is  $e$ . Note that each item’s chances of selection is proportional to its similarity to the previous item, along the given feature dimension (e.g., color).

where  $\tau = 1$  in our implementation. We note that increasing the value of  $\tau$  would amplify the influence of similarity on order, and decreasing the value of  $\tau$  would diminish the influence of similarity on order. Also note that this approach requires  $\tau > 0$ . Finally, we computed a set of normalized similarity values by dividing the similarities by their sum:

$$\text{similarity}_{\text{normalized}}(a, b) = \frac{\text{similarity}(a, b)}{\sum_{i=1}^n \text{similarity}(a, i)}, \quad (2)$$

where in the denominator,  $i$  takes on each of the  $n$  feature values of the to-be-presented words. The resulting set of normalized similarity scores sums to 1.

As illustrated in Figure 2, we use these normalized similarity scores to construct a sequence of “sticks” that we lay end to end in a line. Each of the  $n$  sticks corresponds to a single to-be-presented word, and the stick lengths are proportional to the relative similarities between each word’s feature value(s) and the feature value(s) of the just-presented word. We choose the next to-be-presented word by moving an indicator along the set of sticks, by a distance chosen uniformly and at random on the interval  $[0, 1]$ . We select the word associated with the stick lying next to the indicator to be presented next. This process continues iteratively (re-computing the similarity scores and stochastically choosing the

322 next to-be-presented word using the just-presented word) until all of the words have been  
323 presented. The result is an ordered list that tends to change gradually along the selected  
324 feature dimension (for example “sorted” lists, see Fig. 1, *Order manipulation* lists).

### 325 **Adaptive condition**

326 We designed the *adaptive* experimental condition to study the effect on memory of lists  
327 that matched (or mismatched) the ways participants “naturally” organized their memories.  
328 Like the other conditions, all participants in the adaptive condition studied a total of 16  
329 lists, in a randomized order. We varied the words’ colors and locations for every word  
330 presentation, as in the feature rich and order manipulation conditions.

331 All participants in the adaptive condition began the experiment by studying a set of  
332 four *initialization* lists. Words and features on these lists were presented in a randomized  
333 order (computed independently for each participant). These initialization lists were used  
334 to estimate each participant’s “memory fingerprint,” defined below. At a high level,  
335 a participant’s memory fingerprint describes how they prioritize or consider different  
336 semantic, lexicographic, and/or visual features when they organize their memories.

337 Next, participants studied a sequence of 12 lists in three batches of four lists each. These  
338 batches came in three types: *random*, *stabilize*, and *destabilize*. The batch types determined  
339 how words on the lists in that batch were ordered. Lists in each batch were always  
340 presented consecutively (e.g., a participant might receive four random lists, followed  
341 by four stabilize lists, followed by four destabilize lists). The batch orders were evenly  
342 counterbalanced across participants: there are six possible orderings of the three batches,  
343 and 10 participants were randomly assigned to each ordering sub-condition.

344 Lists in the random batches were sorted randomly (as on the initialization lists and in  
345 the feature rich condition). Lists in the stabilize and destabilize batches were sorted in ways



346 that either matched or mismatched each participant’s memory fingerprint, respectively.  
347 Our procedures for estimating participants’ memory fingerprints and ordering the stabilize  
348 and destabilize lists are described next.

349 **Feature clustering scores (uncorrected).** Feature clustering scores describe participants’  
350 tendencies to recall similar presented items together in their recall sequences, where  
351 “similarity” considers one given feature dimension (e.g., category, color, etc.). We base  
352 our main approach to computing clustering scores on analogous temporal and semantic  
353 clustering scores developed by Polyn et al. (2009). Computing the clustering score for  
354 one feature dimension starts by considering the corresponding feature values from the  
355 first word the participant recalled correctly from the just-studied list. Next, we sort all  
356 not-yet-recalled words in ascending order according to their feature-based distance to the  
357 just-recalled item (see *Defining feature-based distances*). We then compute the percentile rank  
358 of the observed next recall. We average these percentile ranks across all of the participant’s  
359 recalls for the current list to obtain a single uncorrected clustering score for the list, for the  
360 given feature dimension. We repeated this process for each feature dimension in turn to  
361 obtain a single uncorrected clustering score for each list, for each feature dimension.

362 **Temporal clustering score (uncorrected).** Temporal clustering describes a participant’s  
363 tendency to organize their recall sequences by the learned items’ encoding positions. For  
364 instance, if a participant recalled the lists’ words in the exact order they were presented (or  
365 in exact reverse order), this would yield a score of 1. If a participant recalled the words in  
366 a random order, this would yield an expected score of 0.5. For each recall transition (and  
367 separately for each participant), we sorted all not-yet-recalled words according to their  
368 absolute lag (that is, distance away in the list). We then computed the percentile rank of  
369 the next word the participant recalled. We took an average of these percentile ranks across

all of the participant’s recalls to obtain a single (uncorrected) temporal clustering score for the participant.

**Permutation-corrected feature clustering scores.** Suppose that two lists contain unequal numbers of items of each size. For example, suppose that list *A* contains all “large” items, whereas list *B* contains an equal mix of “large” and “small” items. For a participant recalling list *A*, any correctly recalled item will necessarily match the size of the previous correctly recalled item. In other words, successively recalling several list *A* items of the same size is essentially meaningless, since *any* correctly recalled list *A* word will be large. In contrast, successively recalling several list *B* items of the same size *could* be meaningful, since (early in the recall sequence) the yet-to-be-recalled items come from a mix of sizes. However, once all of the small items on list *B* have been recalled, the best possible next matching recall will be a large item. All subsequent correct recalls must also be large items—so for those later recalls it becomes difficult to determine whether the participant is successively recalling large items because they are organizing their memories according to size, or (alternatively), whether they are simply recalling the yet-to-be-recalled items in a random order. In general, the precise order and blend of feature values expressed in a given list, the order and number of correct recalls a participant makes, the number of intervening presentation positions between successive recalls, and so on, can all affect the range of clustering scores that are possible to observe for a given list. An uncorrected clustering score therefore conflates participants’ actual memory organization with other “nuisance” factors.

Following our prior work (Heusser et al., 2017), we used a permutation-based correction procedure to help isolate the behavioral aspects of clustering that we were most interested in. After computing the uncorrected clustering score (for the given list and observed recall sequence), we compute a “null” distribution of  $n$  additional clustering

395 scores after randomly shuffling the order of the recalled words (we use  $n = 500$  in the  
396 present study). This null distribution represents an approximation of the range of cluster-  
397 ing scores one might expect to observe by “chance,” given that a hypothetical participant  
398 was *not* truly clustering their recalls, but where the hypothetical participant still studied  
399 and recalled exactly the same items (with the same features) as the true participant. We  
400 define the *permutation-corrected clustering score* as the percentile rank of the observed un-  
401 corrected clustering score in this estimated null distribution. In this way, a corrected score  
402 of 1 indicates that the observed score was greater than any clustering score one might  
403 expect by chance—in other words, good evidence that the participant was truly clustering  
404 their recalls along the given feature dimension. We applied this correction procedure to  
405 all of the clustering scores (feature and temporal) reported in this paper.

406 **Memory fingerprints.** We define each participant’s *memory fingerprint* as the set of their  
407 permutation-corrected clustering scores across all dimensions we tracked in our study,  
408 including their six feature-based clustering scores (category, size, length, first letter, color,  
409 and location) and their temporal clustering score. Conceptually, a participant’s memory  
410 fingerprint describes their tendency to order in their recall sequences (and, presumably,  
411 organize in memory) the studied words along each dimension. To obtain stable estimates  
412 of these fingerprints for each participant, we averaged their clustering scores across lists.  
413 We also tracked and characterized how participants’ fingerprints changed across lists (e.g.,  
414 Figs. 6, S8).

415 **Online “fingerprint” analysis.** The presentation orders of some lists in the adaptive  
416 condition of our experiment (see *Adaptive condition*) were sorted according to participants’  
417 *current* memory fingerprint, estimated using all of the lists they had studied up to that point  
418 in the experiment. Because our experiment incorporated a speech-to-text component, all

419 of the behavioral data for each participant could be analyzed just a few seconds after the  
420 conclusion of the recall intervals for each list. We used the Quail Python package (Heusser  
421 et al., 2017) to apply speech-to-text algorithms to the just-collected audio data, aggregate  
422 the data for the given participant, and estimate the participant’s memory fingerprint  
423 using all of their available data up to that point in the experiment. Two aspects of our  
424 implementation are worth noting. First, because memory fingerprints are computed  
425 independently for each list and then averaged across lists, the already-computed memory  
426 fingerprints for earlier lists could be cached and loaded as needed in future computations.  
427 This meant that our computations pertaining to updating our estimate of a participant’s  
428 memory fingerprint only needed to consider data from the most recent list. Second, each  
429 element of the null distributions of uncorrected fingerprint scores (see *Permutation-corrected*  
430 *feature clustering scores*) could be estimated independently from the others. This enabled  
431 us to make use of the testing computers’ multi-core CPU architectures by considering (in  
432 parallel) elements of the null distributions in batches of eight (i.e., the number of CPU  
433 cores on each testing computer). Taken together, we were able to compress the relevant  
434 computations into just a few seconds of computing time. The combined processing time for  
435 the speech-to-text algorithm, fingerprint computations, and permutation-based ordering  
436 procedure (described next) easily fit within the inter-list intervals, where participants  
437 paused for a self-paced break before moving on to study and recall the next list.

438 **Ordering “stabilize” and “destabilize” lists by an estimated fingerprint.** In the adap-  
439 tive condition of our experiment, the presentation orders for *stabilize* and *destabilize* lists  
440 were chosen to either maximally or minimally (respectively) comport with participants’  
441 memory fingerprints. Given a participant’s memory fingerprint and a to-be-presented set  
442 of items, we designed a permutation-based procedure for ordering the items. First, we  
443 dropped from the participant’s fingerprint the temporal clustering score. For the remain-

ing feature dimensions, we arranged the clustering scores in the fingerprint into a template  
 vector,  $f$ . Second, we computed  $n = 2500$  random permutations of the to-be-presented  
 items. These permutations served as candidate presentation orders. We sought to select  
 the specific order that most (or least) closely matched  $f$ . Third, for each random permu-  
 tation, we computed the (permutation-corrected) “fingerprint,” treating the permutation  
 as though it were a potential “perfect” recall sequence. (We did not include temporal  
 clustering scores in these fingerprints, since the temporal clustering score for every per-  
 mutation is always equal to 1.) This yielded a “simulated fingerprint” vector,  $\hat{f}_p$  for each  
 permutation  $p$ . We used these simulated fingerprints to select a specific permutation,  $i$ ,  
 that either maximized (for stabilize lists) or minimized (for destabilize lists) the correlation  
 between  $\hat{f}_i$  and  $f$ .

#### Computing low-dimensional embeddings of memory fingerprints

Following some of our prior work (Heusser et al., 2021, 2018; Manning et al., 2022),  
 we use low-dimensional embeddings to help visualize how participants’ memory fin-  
 gerprints change across lists (Figs. 6A, S8A). To compute a shared embedding space  
 across participants and experimental conditions, we concatenated the full set of across-  
 participant average fingerprints (for all lists and experimental conditions) to create a large  
 matrix with number-of-lists (16)  $\times$  number-of-conditions (10, including the adaptive con-  
 dition) rows and seven columns (one for each feature clustering score, plus an additional  
 temporal clustering score column). We used principal components analysis to project  
 the seven-dimensional observations into a two-dimensional space (using the two prin-  
 cipal components that explained the most variance in the data). For two visualizations  
 (Figs. 6B, and S8B), we computed an additional set of two-dimensional embeddings for the  
 average fingerprints across lists within a given list grouping (i.e., early or late). For those

visualizations, we averaged across the rows (for each condition and group of lists) in the combined fingerprint matrix prior to projecting it into the shared two-dimensional space. This yielded a single two-dimensional coordinate for each *list group* (in each condition), rather than for each individual list. We used these embeddings solely for visualization. All statistical tests were carried out in the original (seven-dimensional) feature spaces.

## Analyses

### Probability of $n^{\text{th}}$ recall curves

Probability of first recall curves (Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924) reflect the probability that an item will be recalled first, as a function of its serial position during encoding. To carry out this analysis, we initialized (for each participant) a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then, for each list, we found the index of the word that was recalled first, and we filled in that position in the matrix with a 1. Finally, we averaged over the rows of the matrix to obtain a 1 by 16 array of probabilities, for each participant. We used an analogous procedure to compute probability of  $n^{\text{th}}$  recall curves for each participant. Specifically, we filled in the corresponding matrices according to the  $n^{\text{th}}$  recall on each list that each participant made. When a given participant had made fewer than  $n$  recalls for a given list, we simply excluded that list from our analysis when computing that participant's curve(s). The probability of first recall curve corresponds to a special case where  $n = 1$ .

### Lag-conditional response probability curve

The lag-conditional response probability (lag-CRP) curve (Kahana, 1996) reflects the probability of recalling a given item after the just-recalled item, as a function of their relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item was

presented immediately after the previously recalled item, and a lag of  $-3$  indicates that a recalled item came three items before the previously recalled item. For each recall transition (following the first recall), we computed the lag between the just-recalled word's presentation position and the next-recalled word's presentation position. We computed the proportions of transitions (between successively recalled words) for each lag, normalizing for the total numbers of possible transitions. In carrying out this analysis, we excluded all incorrect recalls and successive repetitions (i.e., recalling the same word twice in a row). This yielded, for each list, a 1 by number-of-lags ( $-15$  to  $+15$ ; 30 lags in total, excluding lags of 0) array of conditional probabilities. We averaged these probabilities across lists to obtain a single lag-CRP for each participant. Because transitions at large absolute lags are rare, these curves are typically displayed using range restrictions (Kahana, 2012).

### **Serial position curve**

Serial position curves (Murdock, 1962) reflect the proportion of participants who remember each item as a function of the items' serial positions during encoding. For each participant, we initialized a number-of-lists (16) by number-of-words-per-list (16) matrix of 0s. Then, for each correct recall, we identified the presentation position of the word and entered a 1 into that position (row: list; column: presentation position) in the matrix. This resulted in a matrix whose entries indicated whether or not the words presented at each position, on each list, were recalled by the participant (depending on whether the corresponding entries were set to 1 or 0). Finally, we averaged over the rows of the matrix to yield a 1 by 16 array representing the proportion of words at each position that the participant remembered.

## 514 Identifying event boundaries

515 We used the distances between feature values for successively presented words (see *Defin-*  
516 *ing feature-based distances*) to estimate “event boundaries” where the feature values changed  
517 more than usual (DuBrow and Davachi, 2016; Ezzyat and Davachi, 2011; Manning et al.,  
518 2016; Radvansky and Copeland, 2006; Swallow et al., 2011, 2009). For each list, for each  
519 feature dimension, we computed the distribution of distances between the feature values  
520 for successively presented words. We defined event boundaries (e.g., Fig. 3B) as occurring  
521 between any successive pair of words whose distances along the given feature dimension  
522 were greater than one standard deviation above the mean for that list. Note that, because  
523 event boundaries are defined for each feature dimension, each individual list may contain  
524 several sets of event boundaries, each at different moments in the presentation sequence  
525 (depending on the feature dimension of interest).

## 526 Results

527 While holding the set of words (and the assignments of words to lists) constant, we  
528 manipulated two aspects of participants’ experiences of studying each list. We sought to  
529 understand the effects of these manipulations on participants’ memories for the studied  
530 words. First, we added two additional sources of visual variation to the individual word  
531 presentations: font color and onscreen location. Importantly, these visual features were  
532 independent of the meaning or semantic content of the words (e.g., word category, size  
533 of the referent, etc.) and of the lexicographic properties of the words (e.g., word length,  
534 first letter, etc.). We wondered whether this additional word-independent information  
535 might facilitate recall (e.g., by providing new potential ways of organizing or retrieving  
536 memories of the studied words) or impair recall (e.g., by distracting participants with



537 irrelevant information). Second, we manipulated the orders in which words were studied  
538 (and how those orderings changed over time). We wondered whether presenting the same  
539 list of words with different appearances (e.g., by manipulating font size and onscreen  
540 location) or in different orders (e.g., sorted along one feature dimension versus another)  
541 might serve to influence how participants organized their memories of the words. We also  
542 wondered whether some order manipulations might be temporally “sticky” by influencing  
543 how *future* lists were remembered.

544 To obtain a clean preliminary estimate of the consequences on memory of randomly  
545 varying the font colors and locations of presented words (versus holding the font color  
546 fixed at black, and holding the display locations fixed at the center of the display) we  
547 compared participants’ performance on the *feature rich* and *reduced* experimental conditions  
548 (see *Random conditions*, Fig. S1). In the feature rich condition the words’ colors and  
549 locations varied randomly across words, and in the reduced condition words were always  
550 presented in black, at the center of the display. Aggregating across all lists for each  
551 participant, we found no difference in recall accuracy (i.e., the proportions of correctly  
552 recalled words) for feature rich versus reduced lists ( $t(126) = -0.290, p = 0.772$ ). However,  
553 participants in the feature rich condition clustered their recalls substantially more along  
554 every dimension we examined (temporal clustering:  $t(126) = 10.624, p < 0.001$ ; semantic  
555 category clustering:  $t(126) = 10.077, p < 0.001$ ; size clustering:  $t(126) = 11.829, p < 0.001$ ;  
556 word length clustering:  $t(126) = 10.639, p < 0.001$ ; first letter clustering:  $t(126) = 7.775, p <$   
557  $0.001$ ; see *Permutation-corrected feature clustering scores* for more information about how we  
558 quantified each participant’s clustering tendencies.) Taken together, these comparisons  
559 suggest that adding new features changes how participants organize their memories of  
560 studied words, even when those new features are independent of the words themselves  
561 and even when the new features vary randomly across words. We found no evidence

that those additional uninformative features were distracting (in terms of their impact on memory performance), but they did affect participants' recall dynamics (measured via their clustering scores).

We also wondered whether adding these incidental visual features to later lists (after the participants had already studied impoverished lists), or removing the visual features from later lists (after the participants had already studied visually diverse lists) might affect memory performance. In other words, we sought to test for potential effects of changing the "richness" of participants' experiences over time. All participants studied and recalled a total of 16 lists; we defined *early* lists as the first eight lists and *late* lists as the last eight lists each participant encountered. To help interpret our results, we compared participants' memories on early versus late lists in the above feature rich and reduced conditions. Participants in both conditions remembered more words on early versus late lists (feature rich:  $t(66) = 4.553, p < 0.001$ ; reduced:  $t(60) = 2.434, p = 0.018$ ). Participants in the feature rich (but not reduced) conditions exhibited more temporal clustering on early versus late lists (feature rich:  $t(66) = 2.318, p = 0.024$ ; reduced:  $t(60) = 0.929, p = 0.357$ ). And participants in both conditions exhibited more semantic (category and size) clustering on early versus late lists (feature rich, category:  $t(66) = 3.805, p < 0.001$ ; feature rich, size:  $t(66) = 2.190, p = 0.032$ ; reduced, category:  $t(60) = 2.856, p = 0.006$ ; reduced, size:  $t(60) = 2.947, p = 0.005$ ). Participants in the reduced (but not feature rich) conditions exhibited more lexicographic clustering on early versus late lists (feature rich, word length:  $t(66) = 0.161, p = 0.872$ ; feature rich, first letter:  $t(66) = 0.410, p = 0.683$ ; reduced, word length:  $t(60) = 3.528, p = 0.001$ ; reduced, first letter:  $t(60) = 2.275, p = 0.026$ ). Taken together, these comparisons suggest that even when the presence or absence of incidental visual features is stable across lists, participants still exhibit some differences in their performance and memory organization tendencies for early versus late lists.

587 With these differences in mind, we next compared participants' memories on early ver-  
 588 sus late lists for two additional experimental conditions (see *Random conditions*, Fig. S1). In  
 589 a *reduced (early)* condition, we held the visual features constant on early lists, but allowed  
 590 them to vary randomly on late lists. In a *reduced (late)* condition, we allowed the visual fea-  
 591 tures to vary randomly on early lists, but held them constant on late lists. Given our above  
 592 findings that (a) participants tended to remember more words and exhibit stronger cluster-  
 593 ing effects on feature rich (versus reduced) lists, and (b) participants tended to remember  
 594 more words and exhibit stronger clustering effects on early (versus late) lists, we expected  
 595 these early versus late differences to be enhanced in the reduced (early) condition and  
 596 diminished in the reduced (late) condition. However, to our surprise, participants in *nei-*  
 597 *ther* condition exhibited reliable early versus late differences in accuracy (reduced (early):  
 598  $t(41) = 1.499, p = 0.141$ ; reduced (late):  $t(40) = 1.462, p = 0.152$ ), temporal clustering (re-  
 599 duced (early):  $t(41) = 0.998, p = 0.324$ ; reduced (late):  $t(40) = 1.099, p = 0.278$ ), nor feature-  
 600 based clustering (reduced (early), category:  $t(41) = 0.753, p = 0.456$ ; reduced (early), size:  
 601  $t(41) = 0.721, p = 0.475$ ; reduced (early), length:  $t(41) = 0.493, p = 0.625$ ; reduced (early),  
 602 first letter:  $t(41) = 0.780, p = 0.440$ ; reduced (late), category:  $t(40) = -0.086, p = 0.932$ ;  
 603 reduced (late), size:  $t(40) = 0.746, p = 0.460$ ; reduced (late), length:  $t(40) = 1.476, p = 0.148$ ;  
 604 reduced (late), first letter:  $t(40) = 0.966, p = 0.340$ ). We hypothesized that adding or remov-  
 605 ing the variability in the visual features was acting as a sort of "event boundary" between  
 606 early and late lists. In prior work, we (and others) have found that memories formed just  
 607 after event boundaries can be enhanced (e.g., due to less contextual interference between  
 608 pre- and post-boundary items; Flores et al., 2017; Gold et al., 2017; Manning et al., 2016;  
 609 Pettijohn et al., 2016).

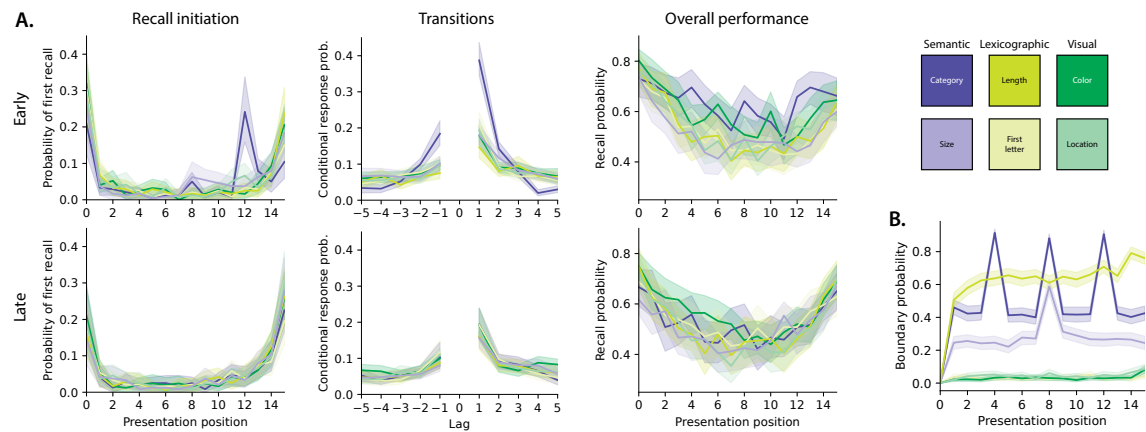
610 We found that *adding* incidental visual features on later lists that had not been present  
 611 on early lists (as in the reduced (early) condition) served to enhance recall performance

612 relative to conditions where all lists had the same blends of features (accuracy for feature  
 613 rich versus reduced (early):  $t(107) = -2.230, p = 0.028$ ; reduced versus reduced (early):  
 614  $t(101) = -2.045, p = 0.043$ ; also see Fig. S3A). However, *subtracting* irrelevant visual fea-  
 615 tures on later lists that *had* been present on early lists (as in the reduced (late) condition) did  
 616 not appear to impact recall performance (accuracy for feature rich versus reduced (late):  
 617  $t(106) = -0.638, p = 0.525$ ; reduced versus reduced (late):  $t(100) = -0.407, p = 0.685$ ).  
 618 These comparisons suggest that recall accuracy has a directional component: accuracy is  
 619 affected differently by removing features later that had been present earlier versus adding  
 620 features later that had *not* been present earlier. In contrast, we found that participants  
 621 exhibited more temporal and feature-based clustering when we added incidental visual  
 622 features to *any* lists (comparisons of clustering on feature rich versus reduced lists are  
 623 reported above; temporal clustering in reduced versus reduced (early) and reduced ver-  
 624 sus reduced (late) conditions:  $ts \leq -9.780, ps < 0.001$ ; feature-based clustering in reduced  
 625 versus reduced (early) and reduced versus reduced (late) conditions:  $ts \leq -5.443, ps$   
 626  $< 0.001$ ). Temporal and feature-based clustering were not reliably different in the feature  
 627 rich, reduced (early), and reduced (late) conditions (temporal clustering in feature rich  
 628 versus reduced (early) and feature rich versus reduced (late) conditions:  $ts \geq -1.434, ps$   
 629  $\geq 0.154$ ; feature-based clustering in feature rich versus reduced (early) and feature rich  
 630 versus reduced (late) conditions:  $ts \geq -1.359, ps > 0.177$ ).

631 Taken together, our findings thus far suggest that adding item features that change  
 632 over time, even when they vary randomly and independently of the items, can enhance  
 633 participants' overall memory performance and can also enhance temporal and feature-  
 634 based clustering. To the extent that the number of item features that vary from moment  
 635 to moment approximates the "richness" of participants' experiences, our findings sug-  
 636 gest that participants remember "richer" stimuli better and organize richer stimuli more

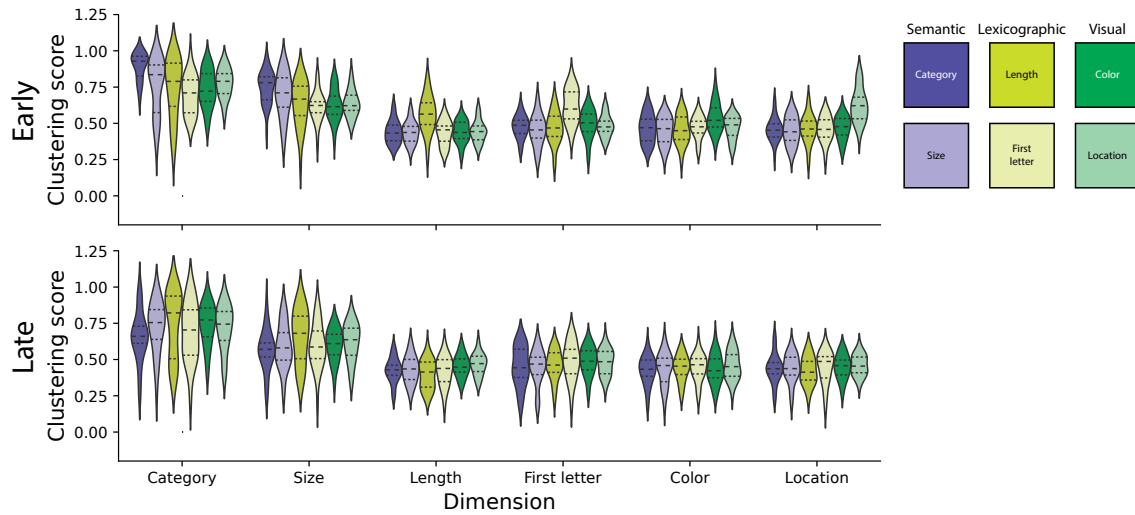
reliably in their memories. Next, we turn to examine the memory effects of varying the temporal ordering of different stimulus features. We hypothesized that changing the orders in which participants were exposed to the words on a given list might enhance (or diminish) the relative influence of different features. For example, presenting a set of words alphabetically might enhance participants' attention to the studied items' first letters, whereas sorting the same list of words by semantic category might instead enhance participants' attention to the words' semantic attributes. Importantly, we expected these order manipulations to hold even when the variation in the total set of features (across words) was held constant across lists (e.g., unlike in the reduced (early) and reduced (late) conditions, where variations in visual features were added or removed from a subset of the lists participants studied).

Across each of six order manipulation conditions, we sorted early lists by one feature dimension but randomly ordered the items on late lists (see *Order manipulation conditions*; features: category, size, length, first letter, color, and location). Participants in the category-ordered condition showed an increase in memory performance on early lists (accuracy, relative to early feature rich lists;  $t(95) = 3.034, p = 0.003$ ). Participants in the color-ordered condition also showed a trending increase in memory performance on early lists (again, relative to early feature rich lists:  $t(96) = 1.850, p = 0.067$ ). Participants' performances on early lists in all of the other order manipulation conditions were indistinguishable from performance on the early feature rich lists ( $|t|s < 1.013, ps > 0.314$ ). Participants in both of the semantically ordered conditions exhibited stronger temporal clustering on early lists (versus early feature rich lists; category:  $t(95) = 8.508, p < 0.001$ ; size:  $t(95) = 2.429, p = 0.017$ ). Participants in the length-ordered condition tended to exhibit *less* temporal clustering on early lists relative to early feature rich lists ( $t(95) = -1.666, p = 0.099$ ), whereas participants in the first letter-ordered condi-



**Figure 3: Recall dynamics in feature rich free recall (order manipulation conditions).** **A.** Behavioral plots. **Left panels.** The probabilities of initiating recall with each word are plotted as a function of presentation position. **Middle panels.** The conditional probabilities of recalling each word are plotted as a function of the relative position (Lag) to the words recalled just-prior. **Right panels.** The overall probabilities of recalling each word are plotted as a function of presentation position. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals (calculated across participants). Top panels display the recall dynamics for early (order manipulation) lists in each condition (color). Bottom panels display the recall dynamics for late (randomly ordered) lists. See Figures S1 and S2 for analogous plots for the random and adaptive conditions. **B.** Proportion of event boundaries (see *Identifying event boundaries*) for each condition's feature of focus, plotted as a function of presentation position.

662 tion exhibited stronger temporal clustering on early lists ( $t(95) = 2.587, p = 0.011$ ). Partici-  
 663 pants in the visually ordered conditions exhibited more similar performance on early lists,  
 664 relative to early feature rich lists (color:  $t(96) = -1.064, p = 0.290$ ; we found a trending  
 665 enhancement for participants in the location-ordered condition:  $t(95) = 1.682, p = 0.096$ ).  
 666 We also compared feature-based clustering on early lists across the order manipulation  
 667 and feature rich conditions. Since these results were similar across both semantic con-  
 668 ditions (category and size), both lexicographic conditions (length and first letter), and  
 669 both visual conditions (color and location), here we aggregate data from conditions that  
 670 manipulated each of these three feature groupings in our comparisons, to simplify the  
 671 presentation. On early lists, participants in the semantically ordered conditions exhibited  
 672 stronger semantic clustering relative to participants in the feature rich condition (category:  
 673  $t(125) = 2.524, p = 0.013$ ; size:  $t(125) = 3.510, p = 0.001$ ), but showed no reliable differences  
 674 in lexicographic (length:  $t(125) = 0.539, p = 0.591$ ; first letter:  $t(125) = -0.587, p = 0.558$ )  
 675 or visual (color:  $t(125) = -0.579, p = 0.564$ ; location:  $t(125) = -0.346, p = 0.730$ ) clustering.  
 676 Similarly, participants in the lexicographically ordered conditions exhibited stronger (rela-  
 677 tive to feature rich participants) lexicographic clustering (length:  $t(125) = 3.426, p = 0.001$ ;  
 678 first letter:  $t(125) = 3.236, p = 0.002$ ) on early lists, but showed no reliable differences in  
 679 semantic (category:  $t(125) = -1.078, p = 0.283$ ; size:  $t(125) = -0.310, p = 0.757$ ) or visual  
 680 (color:  $t(125) = -0.209, p = 0.835$ ; location:  $t(125) = -0.004, p = 0.997$ ) clustering. And  
 681 participants in the visually ordered conditions exhibited stronger visual clustering (again,  
 682 relative to feature rich participants, and on early lists; color:  $t(126) = 2.099, p = 0.038$ ;  
 683 location:  $t(126) = 4.392, p < 0.001$ ), but showed no reliable differences in semantic (cate-  
 684 gory:  $t(126) = 0.204, p = 0.839$ ; size:  $t(126) = -0.093, p = 0.926$ ) or lexicographic (length:  
 685  $t(126) = 0.714, p = 0.476$ ; first letter:  $t(126) = 0.820, p = 0.414$ ) clustering. Taken together,  
 686 these order manipulation results suggest several broad patterns (Figs. 3A, 4). First, most of



**Figure 4: Memory “fingerprints” (order manipulation conditions).** The across-participant distributions of clustering scores for each feature type ( $x$ -coordinate) are displayed for each experimental condition (color), separately for order manipulation (early, top) and randomly ordered (late, bottom) lists. See Figures S5 and S6 for analogous plots for the random and adaptive conditions.

the order manipulations we carried out did *not* reliably affect overall recall performance. Second, most of the order manipulations increased participants’ tendencies to temporally cluster their recalls. Third, all of the order manipulations enhanced participants’ clustering of each condition’s target feature (i.e., semantic manipulations enhanced semantic clustering, lexicographic manipulations enhanced lexicographic clustering, and visual manipulations enhanced visual clustering) while leaving clustering along other feature dimensions roughly unchanged (i.e., semantic manipulations did not affect lexicographic or visual clustering, and so on).

When we closely examined the sequences of words participants recalled from early order-manipulated lists (Fig. 3A, top panel), we noticed several differences from the dynamics of participants’ recalls of randomly ordered lists (Figs. S1, S7). One difference is that participants in the category condition (dark purple curves, Fig. 3) most often initiated recall with the fourth-from-last item (*Recall initiation*, top left panel), whereas participants



700 who recalled randomly ordered lists tended to initiate recall with either the first or last list  
701 items (Fig. S1, top left panel). We hypothesized that the participants might be “clumping”  
702 their recalls into groups of items that shared category labels. Indeed, when we com-  
703 pared the positions of feature changes in the study sequence (Fig. 3B; see *Identifying event*  
704 *boundaries*) with the positions of items participants recalled first, we noticed a striking  
705 correspondence in both semantic conditions. Specifically, on category-ordered lists, the  
706 category labels changed every four items on average (dark purple peaks in Fig. 3B), and  
707 participants also seemed to display an increased tendency (relative to other order manipu-  
708 lation and random conditions) to initiate recall of category-ordered lists with items whose  
709 study positions were integer multiples of four. Similarly, for size-ordered lists, the size la-  
710 bels changed every eight items on average (light purple peaks in Fig. 3B), and participants  
711 also seemed to display an increased tendency to initiate recall of size-ordered lists with  
712 items whose study positions were integer multiples of eight. A second striking difference  
713 is that participants in the category condition exhibited a much steeper lag-CRP (Fig. 3A,  
714 top middle panel) than participants in other conditions. (This is another expression of  
715 participants’ increased tendencies to temporally cluster their recalls on category-ordered  
716 lists, as we reported above.) Taken together, these order-specific idiosyncrasies suggest  
717 a hierarchical set of influences on participants’ memories. At longer timescales, “event  
718 boundaries” (to use the term loosely) can be induced across lists by adding or removing  
719 incidental visual features. At shorter timescales, “event boundaries” can be induced across  
720 items (within a single list) by adjusting how item features change throughout the list.

721 The above comparisons between memory performance on early lists in the order ma-  
722 nipulation versus feature rich conditions highlight how sorted lists are remembered differ-  
723 ently from random lists. We also wondered how sorting lists along each feature dimension  
724 influenced memory relative to sorting lists along the other feature dimensions. Partici-

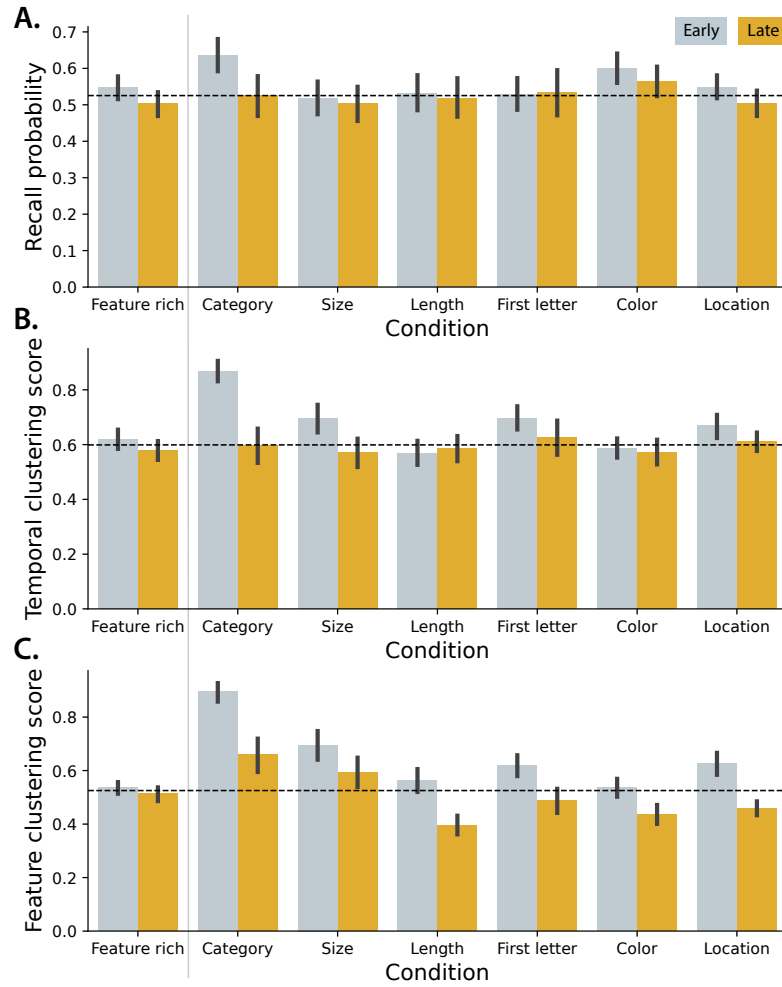
pants trended towards remembering early lists that were sorted semantically better than  
 lexicographically sorted lists ( $t(118) = 1.936, p = 0.055$ ). Participants also remembered  
 visually sorted lists better than lexicographically sorted lists ( $t(119) = 2.145, p = 0.034$ ).  
 However, participants showed no reliable differences in recall for semantically versus  
 visually sorted lists ( $t(119) = 0.113, p = 0.910$ ). Participants temporally clustered semanti-  
 cally sorted lists more strongly than either lexicographically ( $t(118) = 5.572, p < 0.001$ ) or  
 visually ( $t(119) = 6.215, p < 0.001$ ) sorted lists, but did not show reliable differences in tem-  
 poral clustering on lexicographically versus visually sorted lists ( $t(119) = 0.189, p = 0.850$ ).  
 Participants also showed reliably more semantic clustering on semantically sorted lists  
 than lexicographically (category:  $t(118) = 3.492, p = 0.001$ , size:  $t(118) = 3.972, p < 0.001$ )  
 or visually (category:  $t(119) = 2.702, p = 0.008$ , size:  $t(119) = 4.230, p < 0.001$ ) sorted  
 lists; more lexicographic clustering on lexicographically sorted lists than semantically  
 (length:  $t(118) = 3.112, p = 0.002$ ; first letter:  $t(118) = 3.686, p < 0.001$ ) or visually (length:  
 $t(119) = 3.024, p = 0.003$ ; first letter:  $t(119) = 2.644, p = 0.009$ ) sorted lists; and more visual  
 clustering on visually sorted lists than semantically (color:  $t(119) = -2.659, p = 0.009$ ;  
 location:  $t(119) = -4.604, p < 0.001$ ) or lexicographically (color:  $t(119) = -2.366, p = 0.020$ ;  
 location:  $t(119) = -4.265, p < 0.001$ ) sorted lists. In summary, sorting lists by different  
 features appeared to have slightly different effects on overall memory performance and  
 temporal clustering. Participants also tended to cluster their recalls along a given fea-  
 ture dimension more when the studied lists were (versus were not) sorted along that  
 dimension.

Beyond affecting how we process and remember *ongoing* experiences, what is happen-  
 ing to us now can also affect how we process and remember *future* experiences. Within  
 the framework of our study, we wondered: if early lists are sorted along different feature  
 dimensions, might this affect how people remember later (random) lists? In exploring this

question, we considered both group-level effects (i.e., effects that tended to be common across individuals) and participant-level effects (i.e., effects that were idiosyncratic across individuals).

At the group level, there seemed to be almost no lingering impact of sorting early lists on memory for later lists. To simplify the presentation, we report these null results in aggregate across the three feature groupings. Relative to memory performance on late feature rich lists, participants' memory performance in all six order manipulation conditions showed no reliable differences (semantic:  $t(125) = 0.487, p = 0.627$ ; lexicographic:  $t(125) = 0.878, p = 0.382$ ; visual:  $t(126) = 1.437, p = 0.153$ ). Nor did we observe any reliable differences in temporal clustering on late lists (relative to late feature rich lists; semantic:  $t(125) = 0.146, p = 0.884$ ; lexicographic:  $t(125) = 0.923, p = 0.358$ ; visual:  $t(126) = 0.525, p = 0.601$ ). Aside from a slightly increased tendency for participants to cluster words by their length on late visual order manipulation lists (more than late feature rich lists;  $t(126) = 2.199, p = 0.030$ ), we observed no reliable differences in any type of feature clustering on late order manipulation condition lists versus late feature rich lists ( $|t| \leq 1.234, p \geq 0.220$ ).

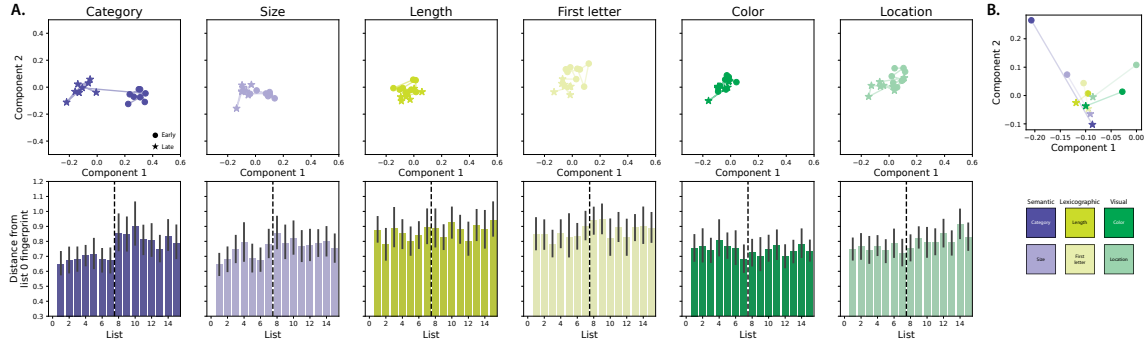
We also looked for more subtle group-level patterns. For example, perhaps sorting early lists by one feature dimension could affect how participants cluster *other* features (on early and/or late lists) as well. We defined participants' *memory fingerprints* as the set of their temporal and feature clustering scores (see *Memory fingerprints*). A participant's memory fingerprint describes how they tend to retrieve memories of the studied items, perhaps searching in parallel through several feature spaces (or along several representational dimensions). To gain insights into the dynamics of how participants' clustering scores tended to change over time, we computed the average (across participants) fingerprint from each list, from each order manipulation condition (Fig. 6). We projected these



**Figure 5: Recall probability and clustering scores on early and late lists.** The bar heights display the average (across participants) recall probabilities (A.), temporal clustering scores (B.), and feature clustering scores (C.) for early (gray) and late (gold) lists. For the feature rich bars (left), the feature clustering scores are averaged across features. For the order manipulation conditions, feature clustering scores are displayed for the focused-on feature for each condition (e.g., category clustering scores are displayed for the category condition, and so on). All panels: error bars denote bootstrap-estimated 95% confidence intervals. The horizontal dotted lines denote the average values (across all lists and participants) for the feature rich condition.

775 fingerprints into a two-dimensional space to help visualize the dynamics (top panels; see  
776 *Computing low-dimensional embeddings of memory fingerprints*). We found that participants’  
777 average fingerprints tended to remain relatively stable on early lists, and exhibited a  
778 “jump” to another stable state on later lists. The sizes of these jumps varied somewhat  
779 across conditions (the Euclidean distances between fingerprints in their original high di-  
780 mensional spaces are displayed in the bottom panels). We also averaged the fingerprints  
781 across early and late lists, respectively, for each condition (Fig. 6B). We found that par-  
782 ticipants’ fingerprints on early lists seem to be influenced by the order manipulations  
783 for those lists (see the locations of the circles in Fig. 6B). There also seemed to be some  
784 consistency across different features within a broader type. For example, both semantic  
785 feature conditions (category and size; purple markers) diverge in a similar direction from  
786 the group; both lexicographic feature conditions (length and first letter; yellow markers)  
787 diverge in a similar direction; and both visual conditions (color and location; green) also  
788 diverge in a similar direction. But on late lists, participants’ fingerprints seem to return  
789 to a common state that is roughly shared across conditions (i.e., the stars in that panel are  
790 clumped together).

791 When we examined the data at the level of individual participants (Figs. 7 and 8), a  
792 clearer story emerged. Within each order manipulation condition, participants exhibited  
793 a range of feature clustering scores on both early and late lists (Fig. 7A, B). Across every  
794 order manipulation condition, participants who exhibited stronger feature clustering (for  
795 their condition’s manipulated feature) recalled more words. This trend held overall across  
796 conditions and participants (early:  $r(179) = 0.537, p < 0.001$ ; late:  $r(179) = 0.492, p < 0.001$ )  
797 as well as for each condition individually for early ( $r_s \geq 0.386$ , all  $p_s \leq 0.035$ ) and late  
798 ( $r_s \geq 0.462$ , all  $p_s \leq 0.010$ ) lists. We found no evidence of a condition-level trend; for  
799 example, the conditions where participants tended to show stronger clustering scores

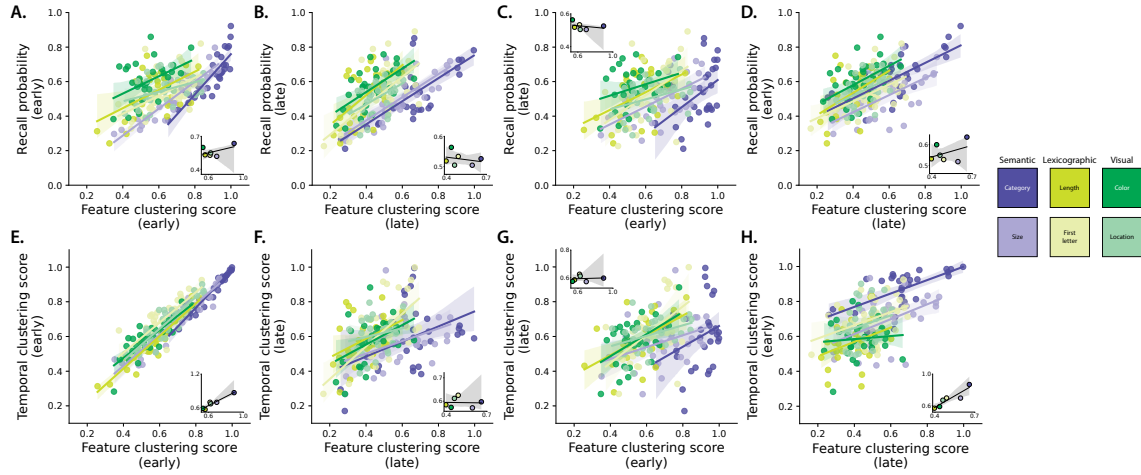


**Figure 6: Memory fingerprint dynamics (order manipulation conditions).** **A.** Each column (and color) reflects an experimental condition. In the top panels, each marker displays a 2D projection of the (across-participant) average memory fingerprint for one list. Order manipulation (early) lists are denoted by circles and randomly ordered (late) lists are denoted by stars. All of the fingerprints (across all conditions and lists) are projected into a common space. The bar plots in the bottom panels display the Euclidean distances of the per-list memory fingerprints to the list 0 fingerprint, for each condition. Error bars denote bootstrap-estimated 95% confidence intervals. The dotted vertical lines denote the boundaries between early and late lists. **B.** In this panel, the fingerprints for early (circle) and late (star) lists are averaged across lists and participants before projecting the fingerprints into a (new) 2D space. See Figure S8 for analogous plots for the random conditions.

were not correlated with the conditions where participants remembered more words (early:  $r(4) = 0.526, p = 0.284$ ; late:  $r(4) = -0.257, p = 0.623$ ; see insets of Fig. 7A and B). We observed carryover associations between feature clustering and recall performance (Fig. 7C, D). Participants who showed stronger feature clustering on early lists tended to recall more items on late lists (across conditions:  $r(179) = 0.492, p < 0.001$ ; all conditions individually:  $r_s \geq 0.462$ , all  $p_s \leq 0.010$ ). Participants who recalled more items on early lists also tended to show stronger feature clustering on late lists (across conditions:  $r(179) = 0.280, p < 0.001$ ; all non-visual conditions:  $r_s \geq 0.445$ , all  $p_s \leq 0.014$ ; color:  $r(29) = 0.298, p = 0.103$ ; location:  $r(28) = 0.354, p = 0.055$ ). Neither of these effects showed condition-level trends (early feature clustering versus late recall probability:  $r(4) = -0.299, p = 0.565$ ; early recall probability versus late feature clustering:  $r(4) = 0.400, p = 0.432$ ). We also looked for associations between feature clustering and temporal clustering. Across every order manipulation condition, participants who exhibited stronger feature clustering also

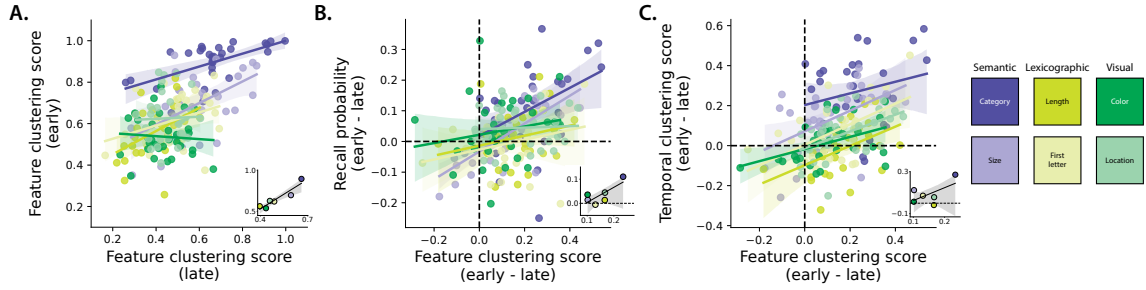
813 exhibited stronger temporal clustering. For early lists (Fig. 7E), this trend held overall  
 814 ( $r(179) = 0.924, p < 0.001$ ), for each condition individually (all  $r_s \geq 0.822$ , all  $p_s < 0.001$ ),  
 815 and across conditions ( $r(4) = 0.964, p = 0.002$ ). For late lists (Fig. 7F), the results were more  
 816 variable (overall:  $r(179) = 0.348, p < 0.001$ ; all non-visual conditions:  $r_s \geq 0.382$ , all  $p_s$   
 817  $\leq 0.037$ ; color:  $r(29) = 0.453, p = 0.011$ ; location:  $r(28) = 0.190, p = 0.314$ ; across-conditions:  
 818  $r(4) = -0.036, p = 0.945$ ). While less robust than the carryover associations between feature  
 819 clustering and recall performance, we also observed some carryover associations between  
 820 feature clustering and temporal clustering (Fig. 7G, H). Participants who showed stronger  
 821 feature clustering on early lists trended towards showing stronger temporal clustering  
 822 on later lists (overall:  $r(179) = 0.301, p < 0.001$ ; for individual conditions: all  $r_s \geq 0.297$ ,  
 823 all  $p_s \leq 0.111$ ; across conditions:  $r(4) = 0.107, p = 0.840$ ). And participants who showed  
 824 stronger temporal clustering on early lists trended towards showing stronger feature  
 825 clustering on later lists (overall:  $r(179) = 0.579, p < 0.001$ ; all non-visual conditions:  $r_s$   
 826  $\geq 0.323$ , all  $p_s \leq 0.082$ ; visual conditions:  $r_s \geq 0.089$ , all  $p_s \leq 0.632$ ; across conditions:  
 827  $r(4) = 0.916, p = 0.010$ ). Taken together, the results displayed in Figure 7 show that  
 828 participants who were more sensitive to the order manipulations (i.e., participants who  
 829 showed stronger feature clustering for their condition's feature on early lists) remembered  
 830 more words and showed stronger temporal clustering. These associations also appeared  
 831 to carry over across lists, even when the items on later lists were presented in a random  
 832 order.

833 If participants show different sensitivities to order manipulations, how do their be-  
 834 haviors carry over to later lists? We found that participants who showed strong feature  
 835 clustering on early lists often tended to show strong feature clustering on late lists (Fig. 8A;  
 836 overall across participants and conditions:  $r(179) = 0.592, p < 0.001$ ; non-visual feature  
 837 conditions: all  $r_s \geq 0.350$ , all  $p_s \leq 0.058$ ; color:  $r(29) = -0.071, p = 0.704$ ; location:



**Figure 7: Interactions between feature clustering, recall probability, and contiguity.** A. Recall probability versus feature clustering scores for order manipulation (early) lists. B. Recall probability versus feature clustering for randomly ordered (late) lists. C. Recall probability on late lists versus feature clustering on early lists. D. Recall probability on early lists versus feature clustering on late lists. E. Temporal clustering scores (contiguity) versus feature clustering scores on early lists. F. Temporal clustering scores versus feature clustering scores on late lists. G. Temporal clustering scores on late lists versus feature clustering scores on early lists. H. Temporal clustering scores on early lists versus feature clustering scores on late lists. **All panels.** Each dot in the main scatterplots denotes the average scores for one participant. The colored regression lines are computed across participants. The inset displays condition-averaged results, where each dot reflects a single condition and the regression line is computed across experimental conditions. All error ribbons denote bootstrap-estimated 95% confidence intervals.

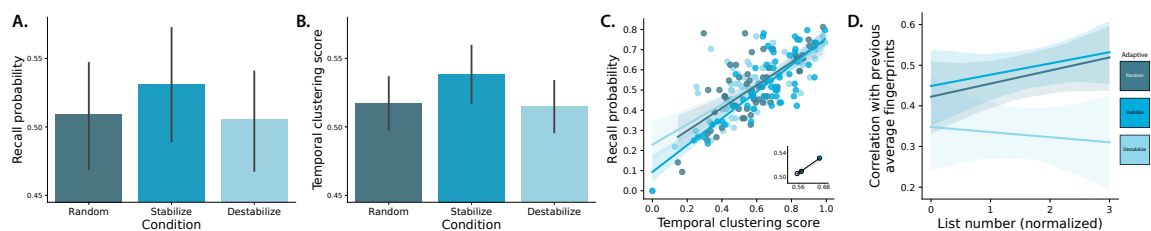




**Figure 8: Feature clustering carryover effects.** **A.** Feature clustering scores for order manipulation (early) versus randomly ordered (late) lists. **B.** Accuracy differences (on early versus late lists) versus feature clustering “carryover” (defined as the differences between the average clustering scores on early and late lists). **C.** Temporal clustering differences (on early versus late lists) versus feature clustering carryover. **All panels.** Each dot in the main scatterplots denotes the average scores for one participant. The colored regression lines are computed across participants. The inset displays condition-averaged results, where each dot reflects a single condition and the regression line is computed across experimental conditions. All error ribbons denote bootstrap-estimated 95% confidence intervals.

838  $r(28) = 0.032, p = 0.868$ ; across conditions:  $r(4) = 0.934, p = 0.006$ ). Although participants  
839 tended to show weaker feature clustering on late lists (Fig. 6) on *average*, the associations  
840 between early and late lists for individual participants suggests that some influence of  
841 early order manipulations may linger on late lists. We found that participants who exhib-  
842 ited larger carryover in feature clustering (i.e., continued to show strong feature clustering  
843 on late lists) for the semantic order manipulations (but not other manipulations) also  
844 tended to show a larger improvement in recall (Fig. 8B; overall:  $r(179) = 0.378, p < 0.001$ ;  
845 category:  $r(28) = 0.419, p = 0.021$ ; size:  $r(28) = 0.737, p < 0.001$ ; non-semantic condi-  
846 tions: all  $rs \leq 0.252$ , all  $ps \geq 0.179$ ; across conditions:  $r(4) = 0.773, p = 0.072$ ) on late  
847 lists, relative to early lists. Participants who exhibited larger carryover in feature cluster-  
848 ing also tended to show stronger temporal clustering on late lists (relative to early lists)  
849 for all but the category condition (Fig. 8C; overall:  $r(179) = 0.434, p < 0.001$ ; category:  
850  $r(28) = 0.229, p = 0.223$ ; all non-category conditions: all  $rs \geq 0.448$ , all  $ps \leq 0.012$ ; across  
851 conditions:  $r(4) = 0.598, p = 0.210$ ).

We suggest two potential interpretations of these findings. First, it is possible that some participants are more “malleable” or “adaptable” with respect to how they organize incoming information. When presented with list of items sorted along *any* feature dimension, they will simply adopt that feature as a dominant dimension for organizing those items and subsequent (randomly ordered) items. This flexibility in memory organization might afford such participants a memory advantage, explaining their strong recall performance. An alternative interpretation is that each participant comes into our study with a “preferred” way of organizing incoming information. If they happen to be assigned to an order manipulation condition that matches their preferences, then they will appear to be “sensitive” to the order manipulation and also exhibit a high degree of carryover in feature clustering from early to late lists. These participants might demonstrate strong recall performance not because of their inherently superior memory abilities, but rather because the specific condition they were assigned to happened to be especially easy for them, given their pre-experimental tendencies. To help distinguish between these interpretations, we designed an *adaptive* experimental condition (see *Adaptive condition*). The primary manipulation in the adaptive condition is that participants each experience three key types of lists. On *random* lists, words are ordered randomly (as in the feature rich condition). On *stabilize* lists, the presentation order is adjusted to be maximally similar to the current estimate of the participant’s memory fingerprint (see *Online “fingerprint” analysis*). Third, on *destabilize* lists, the presentation order is adjusted to be *minimally* similar to the current estimate of the participant’s memory fingerprint (see *Ordering “stabilize” and “destabilize” lists by an estimated fingerprint*). The orders in which participants experienced each type of list were counterbalanced across participants to help reduce the influence of potential list-order effects. Because the presentation orders on stabilize and destabilize lists are adjusted to best match each participant’s (potentially unique) memory fingerprint, the



**Figure 9: Adaptive free recall.** **A.** Average probability of recall (taken across words, lists, and participants) for lists from each adaptive condition. **B.** Average temporal clustering scores for lists from each adaptive condition. **C.** Recall probability versus temporal clustering scores by participant (main panel; each participant contributes one dot per condition) and averaged within condition (inset; each dot represents a single condition). **D.** Per-list correlations between the current list’s fingerprint and the average fingerprint computed from all previous lists. The normalized list numbers ( $x$ -axis) denote the number of lists of the same type that the participant had experienced at the time of the current list. All panels: Colors denote the sorting type (condition) for each list. Error bars and ribbons denote bootstrap-estimated 95% confidence intervals. For additional details about participants’ behavior and performance during the adaptive conditions, see Figure S2.

adaptive condition removes uncertainty about whether participants’ assigned conditions might just “happen” to match their preferred ways of organizing their memories.

Participants’ fingerprints on stabilize and random lists tended to become (numerically) slightly more similar to their average fingerprints computed from the previous lists they had experienced, and their fingerprints on destabilize lists tended to become numerically less similar (Fig. 9D). Overall, we found that participants tended to be better at remembering words on stabilize lists relative to words on both random ( $t(59) = 1.740, p = 0.087$ ) and destabilize ( $t(59) = 1.714, p = 0.092$ ) lists (Fig. 9A). Participants showed no reliable differences in their memory performance on destabilize versus random lists ( $t(59) = -0.249, p = 0.804$ ). Participants also exhibited stronger temporal clustering on stabilize lists, relative to random ( $t(59) = 3.554, p = 0.001$ ) and destabilize ( $t(59) = 4.045, p < 0.001$ ) lists (Fig. 9B). We found no reliable differences in temporal clustering for items on random versus destabilize lists ( $t(59) = -0.781, p = 0.438$ ).

As in the other experimental manipulations, participants in the adaptive condition exhibited substantial variability with respect to their overall memory performance and

892 their clustering tendencies (Fig. 9C). We found that individual participants who exhibited  
893 strong temporal clustering scores also tended to recall more items. This held across  
894 subjects, aggregating across all list types ( $r(178) = 0.721, p < 0.001$ ), and for each list type  
895 individually (all  $r_s \geq 0.683$ , all  $p_s \leq 0.001$ ). Taken together, the results from the adaptive  
896 condition suggest that each participant comes into the experiment with their own unique  
897 memory organization tendencies, as characterized by their memory fingerprint. When  
898 participants study lists whose items come pre-sorted according to their unique preferences,  
899 they tend to remember more and show stronger temporal clustering.

## 900 Discussion

901 We asked participants to study and freely recall word lists. The words on each list (and  
902 the total set of lists) were held constant across participants. For each word, we considered  
903 (and manipulated) two semantic features (category and size) that reflected aspects of the  
904 *meanings* of the words, along with two lexicographic features (word length and first letter),  
905 which reflected characteristics of the words' *letters*. These semantic and lexicographic  
906 features are intrinsic to each word. We also considered and manipulated two additional  
907 visual features (color and location) that affected the *appearance* of each studied item, but  
908 could be varied independently of the words' identities. Across different experimental  
909 conditions, we manipulated how the visual features varied across words (within each  
910 list), along with the orders of each list's words. Although the participants' task (verbally  
911 recalling as many words as possible, in any order, within one minute) remained constant  
912 across all of these conditions, and although the set of words they studied from each list  
913 remained constant, our manipulations substantially affected participants' memories. The  
914 impact of some of the manipulations also affected how participants remembered *future*  
915 lists that were sorted randomly.

## 916 **Recap: visual feature manipulations**

917 We found that participants in our feature rich condition (where we varied words' ap-  
918 pearances) recalled similar proportions of words to participants in a reduced condition  
919 (where appearance was held constant across words). However, varying the words' ap-  
920 pearances led participants to exhibit much more temporal and feature-based clustering.  
921 This suggests that even seemingly irrelevant elements of our experiences can affect how  
922 we remember them.

923 When we held the within-list variability in participants' visual experiences fixed across  
924 lists (in the feature rich and reduced conditions), they remembered more words from early  
925 lists than from late lists. For feature rich lists, they also showed stronger clustering for early  
926 versus late lists. However, when we *varied* participants' visual experiences across lists (in  
927 the "reduced (early)" and "reduced (late)" conditions), these early versus late accuracy  
928 and clustering differences disappeared. Abruptly changing how incidental visual features  
929 varied across words seemed to act as a sort of "event boundary" that partially reset how  
930 participants processed and remembered post-boundary lists. Within-list clustering also  
931 increased in these manipulations, suggesting that the "within-event" words were being  
932 more tightly associated with each other.

933 When we held the visual features constant during early lists, but then varied words'  
934 appearances in later lists (i.e., the reduced (early) condition), participants' overall memory  
935 performance improved. However, this impact was directional: when we *removed* visual  
936 features from words in late lists that had been present in early lists (i.e., the reduced (late)  
937 condition), we saw no memory improvement.

## 938 **Recap: order manipulations**

939 When we (stochastically) sorted early lists along different feature dimensions, we found  
940 several impacts on participants' memories. Sorting early lists semantically (by word cat-  
941 egory) enhanced participants' memories for those lists, but the effects on performance of  
942 sorting along other feature dimensions were inconclusive. However, each order manipu-  
943 lation substantially affected how participants *organized* their memories of words from the  
944 ordered lists. When we sorted lists semantically, participants displayed stronger semantic  
945 clustering; when we sorted lists lexicographically, they displayed stronger lexicographic  
946 clustering; and when we sorted lists visually, they displayed stronger visual clustering.  
947 Clustering along the unmanipulated feature dimensions in each of these cases was un-  
948 changed.

949 The order manipulations we examined also appeared to induce, in some cases, a  
950 tendency to "clump" similar words within a list. This was most apparent on semantically  
951 ordered lists, where the probability of initiating recall with a given word seemed to follow  
952 groupings defined by feature change points.

953 We also examined the impact of early list order manipulations on memory for late  
954 lists. At the group level, we found little evidence for lingering "carryover" effects of  
955 these manipulations: participants in the order manipulation conditions showed similar  
956 memory performance and clustering on late lists to participants in the corresponding  
957 control (feature rich) condition. At the level of individual participants, however, we  
958 found several meaningful patterns.

959 Participants who showed stronger feature clustering on early (order-manipulated) lists  
960 tended to better remember late (randomly ordered) lists. Participants who remembered  
961 early lists better also tended to show stronger feature clustering (along their condition's  
962 feature dimension) on late lists (even though the words on those late lists were presented

963 in a random order). We also observed some (weaker) carryover effects of temporal cluster-  
964 ing. Participants who showed stronger feature clustering (along their condition's feature  
965 dimension) on early lists tended to show stronger temporal clustering on late lists. And  
966 participants who showed stronger temporal clustering on early lists also tended to show  
967 stronger feature clustering on late lists. Essentially, these order manipulations appeared to  
968 affect each participant differently. Some participants were sensitive to our manipulations,  
969 and those participants' memory performance was impacted more strongly, both for the  
970 ordered lists and for future (random) lists. Other participants appeared relatively insen-  
971 sitive to our manipulations, and those participants showed little carryover effects on late  
972 lists.

973     These results at the individual participant level suggested to us that either (a) some  
974 participants were more sensitive to *any* order manipulation, or (b) some participants might  
975 be more (or less) sensitive to manipulations along *particular* (e.g., preferred) feature dimen-  
976 sions. To help distinguish between these possibilities, we designed an adaptive condition  
977 whereby we attempted to manipulate whether participants studied words in an order that  
978 either matched or mismatched our estimate of how they would cluster or organize the  
979 studied words in memory (i.e., their idiosyncratic memory fingerprint). We found that  
980 when we presented words in orders that were consistent with participants' memory fin-  
981 gerprints, they remembered more words overall and showed stronger temporal clustering.  
982 This comports well with the second possibility described above. Specifically, each partici-  
983 pant seems to bring into the experiment their own idiosyncratic preferences and strategies  
984 for organizing the words in their memory. When we presented the words in an order  
985 consistent with each participant's idiosyncratic fingerprint, their memory performance  
986 improved. This might indicate that the participants were spending less cognitive effort  
987 "reorganizing" the incoming words on those lists, which freed up resources to devote to

988 encoding processes instead.

## 989 **Context effects on memory performance and organization**

990 In real-world experience, each moment's unique blend of contextual features (where we  
991 are, who we are with, what else we are thinking of at the time, what else we experience  
992 nearby in time, etc.) plays an important role in how we interpret, experience, and re-  
993 member that moment, and how we relate it to our other experiences (e.g., for review see  
994 Manning, 2020). What are the analogues of real-world contexts in laboratory tasks like  
995 the free recall paradigm employed in our study? In general, modern formal accounts of  
996 free recall (Kahana, 2020) describe context as comprising a mix of (a) features pertaining  
997 to or associated with each item and (b) other items and thoughts experienced nearby in  
998 time, e.g., that might still be "lingering" in the participant's thoughts at the time they  
999 study the item. Item features can include semantic properties (i.e., features related to the  
1000 item's meaning), lexicographic properties (i.e., features related to the item's letters), sen-  
1001 sory properties (i.e., feature related to the item's appearance, sound, smell, etc.), emotional  
1002 properties (i.e., features related to how meaningful the item is, whether the item evokes  
1003 positive or negative feelings, etc.), utility-related properties (e.g., features that describe  
1004 how an item might be used or incorporated into a particular task or situation), and more.  
1005 Essentially any aspect of the participant's experience that can be characterized, measured,  
1006 or otherwise described can be considered to influence the participant's mental context at  
1007 the moment they experience that item. Temporally proximal features include aspects of  
1008 the participant's internal or external experience that are *not* specifically occurring at the  
1009 moment they encounter an item, but that nonetheless influence how they process the item.  
1010 Thoughts related to percepts, goals, expectations, other experiences, and so on that might  
1011 have been cued (directly or indirectly) by the participant's recent experiences prior to the



current moment all fall into this category. Internally driven mental states, such as thinking about an experience unrelated to the experiment, also fall into this category.

Contextual features need not be intentionally or consciously perceived by the participant to affect memory, nor do they need to be relevant to the task instructions or the participant's goals. Incidental factors such as font color (Jones and Pyc, 2014), background color (Isarida and Isarida, 2007), inter-stimulus images (Chiu et al., 2021; Gershman et al., 2013; Manning et al., 2016), background sounds (Beaman and Jones, 1998; Sahakyan and Smith, 2014), secondary tasks (Masicampo and Sahakyan, 2014; Oberauer and Lewandowsky, 2008; Polyn et al., 2009), and more can all impact how participants remember, and organize in memory, lists of studied items.

Consistent with this prior work, we found that participants were sensitive to task-irrelevant visual features. We also found that changing the dynamics of those task-irrelevant visual features (in the reduced (early) and reduced (late) conditions) *also* affected participants' memories. This suggests that it is not only the contextual features themselves that affect memory, but also the *dynamics* of context—i.e., how the contextual features associated with each item change over time.

## **Priming effects on memory performance and organization**

When our ongoing experiences are ambiguous, we can draw on our past experiences, expectations, and other real, perceived, or inferred cues to help resolve these ambiguities. We may also be overtly or covertly “primed” to influence how we are likely to resolve ambiguities. For example, before listening to a story with several equally plausible interpretations, providing participants with “background” information beforehand can lead them towards one interpretation versus another (Yeshurun et al., 2017). More broadly, our conscious and unconscious biases and preferences can influence not only how we interpret

1036 high-level ambiguities, but even how we process low-level sensory information (Katabi  
1037 et al., 2023).

1038 In more simplified scenarios, like list-learning paradigms, the stimuli and tasks partic-  
1039 ipants encounter before studying a given list can influence what and how they remember.  
1040 For example, when participants are directed to suppress, disregard, or ignore “distracting”  
1041 stimuli early on in an experiment, participants often tend to remember those stimuli less  
1042 well when they are re-used as to-be-remembered targets later on in the experiment (Tip-  
1043 per, 1985). In general, participants’ memories can be influenced by exposing them to  
1044 a wide range of positive and negative priming factors before they encounter the to-be-  
1045 remembered information (Balota et al., 1992; Clayton and Chattin, 1989; Donnelly, 1988;  
1046 Flexser and Tulving, 1982; Gotts et al., 2012; Huang et al., 2004; Huber, 2008; Huber et al.,  
1047 2001; McNamara, 1994; Neely, 1977; Rabinowitz, 1986; Tulving and Schacter, 1991; Watkins  
1048 et al., 1992; Wiggs and Martin, 1998).

1049 The order manipulation conditions in our experiment show that participants can also be  
1050 primed to pick up on more subtle statistical structure in their experiences, like the dynamics  
1051 of how the presentation orders of stimuli vary along particular feature dimensions. These  
1052 order manipulations affected not only how participants remembered the manipulated  
1053 lists, but also how they remembered *future* lists with different (randomized) temporal  
1054 properties.

## 1055 **Expectation, event boundaries, and situation models**

1056 Our findings that participants’ current and future memory behaviors are sensitive to  
1057 manipulations in which features change over time, and how features change across items  
1058 and lists, suggest parallels with studies on how we form expectations and predictions,  
1059 segment our continuous experiences into discrete events, and make sense of different

scenarios and situations. Each of these real-world cognitive phenomena entail identifying statistical regularities in our experiences, and exploiting those regularities to gain insight, form inferences, organize or interpret memories, and so on. Our past experiences enable us to predict what is likely to happen in the future, given what happened “next” in our previous experiences that were similar to now (Barron et al., 2020; Brigard, 2012; Chow et al., 2016; Eichenbaum and Fortin, 2009; Gluck et al., 2002; Goldstein et al., 2021; Griffiths and Steyvers, 2003; Jones and Pashler, 2007; Kim et al., 2014; Manning, 2020; Tamir and Thornton, 2018; Xu et al., 2023).

When our expectations are violated, such as when our observations disagree with our predictions, we may perceive the “rules” or “situation” to have changed. *Event boundaries* denote abrupt changes in the state of our experience, for example, when we transition from one situation to another (Radvansky and Zacks, 2017; Zwaan and Radvansky, 1998). Crossing an event boundary can impair our memory for pre-boundary information and enhance our memory for post-boundary information (DuBrow and Davachi, 2013; Manning et al., 2016; Radvansky and Copeland, 2006; Sahakyan and Kelley, 2002). Event boundaries are also tightly associated with the notion of *situation models* and *schemas*—mental frameworks for organizing our understanding about the rules of how we and others are likely to behave, how events are likely to unfold over time, how different elements are likely to interact, and so on. For example, a situation model pertaining to a particular restaurant might set our expectations about what we are likely to experience when we visit that restaurant (e.g., what the building will look like, how it will smell when we enter, how crowded the restaurant is likely to be, the sounds we are likely to hear, etc.). Similarly, as mentioned in the *Introduction*, we might learn a schema describing how events are likely to unfold *across* any sit-down restaurant—e.g., open the door, wait to be seated, receive a menu, decide what to order, place the order, and so on. Situation models and schemas can

1085 help us to generalize across our experiences, and to generate expectations about how new  
1086 experiences are likely to unfold. When those expectations are violated, we can perceive  
1087 ourselves to have crossed into a new situation.

1088 In our study, we found that abruptly changing the “rules” about how the visual  
1089 appearances of words are determined, or about the orders in which words are presented,  
1090 can lead participants to behave similarly to what one might expect upon crossing an event  
1091 boundary. Adding variability in font color and presentation location for words on late  
1092 lists, after those visual features had been held constant on early lists, led participants to  
1093 remember more words on those later lists. One potential explanation is that participants  
1094 perceive an “event boundary” to have occurred when they encounter the first “late” list.  
1095 According to contextual change accounts of memory across event boundaries (e.g., Flores  
1096 et al., 2017; Gold et al., 2017; Pettijohn et al., 2016; Sahakyan and Kelley, 2002), this could  
1097 help to explain why participants in the reduced (early) condition exhibited better overall  
1098 memory performance. Specifically, their memory for late list items could benefit from less  
1099 interference from early list items, and the contextual features associated with late list items  
1100 (after the “event boundary”) might serve as more specific recall cues for those late items  
1101 (relative to if the boundary had not occurred).

## 1102 **Theoretical implications**

1103 Although most modern formal theories of episodic memory have been developed and  
1104 tested to explain memory for list-learning tasks (Kahana, 2020), a number of recent studies  
1105 suggest some substantial differences between memory for lists versus naturalistic stim-  
1106 uli (e.g., real-world experiences, narratives, films, etc.; Heusser et al., 2021; Lee et al., 2020;  
1107 Manning, 2021; Nastase et al., 2020). One reason is that naturalistic stimuli are often much  
1108 more engaging than the highly simplified list-learning tasks typically employed in the

1109 psychological laboratory, perhaps leading participants to pay more attention, exert more  
1110 effort, and stay more consistently motivated to perform well (Nastase et al., 2020). Another  
1111 reason is that the temporal unfoldings of events and occurrences in naturalistic stimuli  
1112 tend to be much more meaningful than the temporal unfoldings of items on typical lists  
1113 used in laboratory memory tasks. Real-world events exhibit important associations at a  
1114 broad range of timescales. For example, an early detail in a detective story may prove to  
1115 be a clue to solving the mystery later on. Further, what happens in one moment typically  
1116 carries some predictive information about what came before or after (Xu et al., 2023). In  
1117 contrast, the lists used in laboratory memory tasks are most often ordered randomly, by  
1118 design, to *remove* meaningful temporal structure in the stimulus (Kahana, 2012).

1119     On one hand, naturalistic stimuli provide a potential means of understanding how our  
1120 memory systems function in the circumstances we most often encounter in our everyday  
1121 lives. This implies that, to understand how memory works in the “real world,” we should  
1122 study memory for stimuli that reflect the relevant statistical structure of real-world expe-  
1123 riences. On the other hand, naturalistic stimuli can be difficult to precisely characterize or  
1124 model, making it difficult to distinguish whether specific behavioral trends follow from  
1125 fundamental workings of our memory systems, from some aspect of the stimulus, or from  
1126 idiosyncratic interactions or interference between participants’ memory systems and the  
1127 stimulus. This challenge implies that, to understand the fundamental nature of memory  
1128 in its “pure” form, we should study memory for highly simplified stimuli that can pro-  
1129 vide relatively unbiased (compared with real-world experiences) measures of the relevant  
1130 patterns and tendencies.

1131     The experiment we report in this paper was designed to help bridge some of this gap  
1132 between naturalistic tasks and more traditional list-learning tasks. We had people study  
1133 word lists similar to those used in classic memory studies, but we also systematically var-

1134 ied the lists' "richness" (by adding or removing visual features) and temporal structure  
1135 (through order manipulations that varied over time and across experimental conditions).  
1136 We found that participants' memory behaviors were sensitive to these manipulations.  
1137 Some of the manipulations led to changes that were common across people (e.g., more  
1138 temporal clustering when words' appearances were varied, enhanced memory for lists  
1139 following an "event boundary," more feature clustering on order-manipulated lists, etc.).  
1140 Other manipulations led to changes that were idiosyncratic (especially carryover effects  
1141 from order manipulations; e.g., participants who remembered more words on early order-  
1142 manipulated lists tended to show stronger feature clustering for their condition's feature  
1143 dimension on late randomly ordered lists, etc.). We also found that participants remem-  
1144 bered more words from lists that were sorted to align with their idiosyncratic clustering  
1145 preferences. Taken together, our results suggest that our memories are susceptible to ex-  
1146 ternal influences (i.e., to the statistical structure of ongoing experiences), but the effects of  
1147 past experiences on future memory are largely idiosyncratic across people.

## 1148 **Potential applications**

1149 Every participant in our study encountered exactly the same words, split into exactly the  
1150 same lists. But participants' memory performance, the orders in which they recalled the  
1151 words, and the effects of early list manipulations on later lists all varied according to how  
1152 we presented the to-be-remembered words.

1153 Our findings raise a number of exciting questions. For example, how far might these  
1154 manipulations be extended? In other words, might there be more sophisticated or clever  
1155 feature or order manipulations that one could implement to have stronger impacts on  
1156 memory? Are there limits to how much impact (on memory performance and/or or-  
1157 ganization) these sorts of manipulations can have? Are those limits universal across

1158 people, or are there individual differences (based on prior experiences, natural strate-  
1159 gies, neuroanatomy, etc.) that impose person-specific limits on the potential impact of  
1160 presentation-level manipulations on memory?

1161 Our findings indicate that the ways word lists are presented affects how people re-  
1162 member them. To the extent that word list memory reflects memory processes that are  
1163 relevant to real-world experiences, one could imagine potential real-world applications of  
1164 our findings. For example, we found that participants remembered more words when the  
1165 presentation order agreed with their memory fingerprints. If analogous fingerprints could  
1166 be estimated for classroom content, perhaps they could be utilized manually by teachers,  
1167 or even by automated content-presentation systems, to optimize how and what students  
1168 remember.

## 1169 **Concluding remarks**

1170 Our work raises deep questions about the fundamental nature of human learning. What  
1171 are the limits of our memory systems? How much does what we remember (and how we  
1172 remember) depend on how we learn or experience the to-be-remembered content? We  
1173 know that our expectations, strategies, situation models learned through prior experiences,  
1174 and more collectively shape how our experiences are remembered. But those aspects of  
1175 our memory are not fixed: when we are exposed to the same experience in a new way, it  
1176 can change how we remember that experience, and also how we remember, process, or  
1177 perceive *future* experiences.

## 1178 **Author contributions**

1179 Conceptualization: JRM and ACH. Methodology: JRM and ACH. Software: JRM, PCF,  
1180 CEF, and ACH. Analysis: JRM, PCF, and ACH. Data collection: ECW, PCF, MRL, AMF,

1181 BJB, DR, and CEF. Data curation and management: ECW, PCF, MRL, and ACH. Writing  
1182 (original draft): JRM. Writing (review and editing): ECW, PCF, MRL, AMF, BJB, DR, CEF,  
1183 and ACH. Supervision: JRM and ACH. Project administration: ECW and PCF. Funding  
1184 acquisition: JRM.

## 1185 **Data and code availability**

1186 All of the data analyzed in this manuscript, along with all of the code for carrying out the  
1187 analyses may be found at <https://github.com/ContextLab/FRFR-analyses>. Code for run-  
1188 ning the non-adaptive experimental conditions may be found at [https://github.com/Con-](https://github.com/ContextLab/efficient-learning-code)  
1189 [textLab/efficient-learning-code](https://github.com/ContextLab/efficient-learning-code). Code for running the adaptive experimental condition  
1190 may be found at <https://github.com/ContextLab/adaptiveFR>. We have also released an as-  
1191 sociated Python toolbox for analyzing free recall data, which may be found at [https://cdl-](https://cdl-quail.readthedocs.io/en/latest/)  
1192 [quail.readthedocs.io/en/latest/](https://cdl-quail.readthedocs.io/en/latest/).

## 1193 **Acknowledgements**

1194 We acknowledge useful discussions, assistance in setting up an earlier (unpublished)  
1195 version of this study, and assistance with some of the data collection efforts from Rachel  
1196 Chacko, Joseph Finkelstein, Sheherzad Mohyidin, Lucy Owen, Gal Perlman, Jake Rost,  
1197 Jessica Tin, Marisol Tracy, Peter Tran, and Kirsten Ziman. Our work was supported in part  
1198 by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility  
1199 of the authors and does not necessarily represent the official views of our supporting  
1200 organizations. The funders had no role in study design, data collection and analysis,  
1201 decision to publish, or preparation of the manuscript.



## References

- Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79(2):97–123.
- Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning and Motivation*, volume 2, pages 89–105. Academic Press, New York, NY.
- Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- Balota, D. A., Black, S. R., and Cheney, M. (1992). Automatic and attentional priming in young and older adults: reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):485–502.
- Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Progress in Neurobiology*, 192:101821–101834.
- Beaman, C. P. and Jones, D. M. (1998). Irrelevant sound disrupts order information in free recall as in serial recall. *The Quarterly Journal of Experimental Psychology Section A*, 51(3):615–636.
- Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49:229–240.
- Bousfield, W. A., Sedgewick, C. H., and Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology*, 67:111–118.
- Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2):177–220.

- 1224 Brigard, F. D. (2012). Predictive memory and the surprising gap. *Frontiers in Psychology*,  
1225 3(420):1–3.
- 1226 Chiu, Y.-C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., and Sahakyan, L. (2021). Sepa-  
1227 ration of item and context in item-method directed forgetting. *NeuroImage*, 235:117983.
- 1228 Chow, W.-Y., Momba, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory  
1229 retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.
- 1230 Clayton, K. and Chaitin, D. (1989). Spatial and semantic priming effects in tests of spa-  
1231 tial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1232 15(3):495–506.
- 1233 Donnelly, R. E. (1988). Priming effects in successive episodic tests. *Journal of Experimental*  
1234 *Psychology: Learning, Memory, and Cognition*, 14:256–265.
- 1235 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for  
1236 the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–  
1237 1286.
- 1238 DuBrow, S. and Davachi, L. (2016). Temporal binding within and across events. *Neurobi-*  
1239 *ology of Learning and Memory*, 134:107–114.
- 1240 Eichenbaum, H. and Fortin, N. J. (2009). The neurobiology of memory based predictions.  
1241 *Philosophical Transactions of the Royal Society of London Series B*, 364(1521):1183–1191.
- 1242 Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological*  
1243 *Review*, 62:145–154.
- 1244 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory?  
1245 *Psychological Science*, 22(2):243–252.

- 1246 Flexser, A. J. and Tulving, E. (1982). Priming and recognition failure. *Journal of Verbal*  
1247 *Learning and Verbal Behavior*, 21:237–248.
- 1248 Flores, S., Bailey, H. R., Eisenberg, M. L., and Zacks, J. M. (2017). Event segmentation  
1249 improves event memory up to one month later. *Journal of Experimental Psychology:*  
1250 *Learning, Memory, and Cognition*, 43(8):1183.
- 1251 Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013). Neural context  
1252 reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33(20):8590–  
1253 8595.
- 1254 Glenberg, A. M., Bradley, M. M., Kraus, T. A., and Renzaglia, G. J. (1983). Studies of the  
1255 long-term recency effect: support for a contextually guided retrieval theory. *Journal of*  
1256 *Experimental Psychology: Learning, Memory, and Cognition*, 12:413–418.
- 1257 Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
1258 prediction” task? individual variability in strategies for probabilistic category learning.  
1259 *Learning and Memory*, 9:408–418.
- 1260 Gold, D. A., Zacks, J. M., and Flores, S. (2017). Effects of cues to event segmentation on  
1261 subsequent memory. *Cognitive Research: Principles and Implications*, 2(1):1.
- 1262 Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder,  
1263 A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto,  
1264 C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A.,  
1265 Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. (2021). Thinking  
1266 ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*,  
1267 page doi.org/10.1101/2020.12.02.403477.

- 1268 Gotts, S. J., Chow, C. C., and Martin, A. (2012). Repetition priming and repetition sup-  
 1269 pression: A case for enhanced efficiency through neural synchronization. *Cognitive*  
 1270 *Neuroscience*, 3(3-4):227–237.
- 1271 Griffiths, T. L. and Steyvers, M. (2003). Prediction and semantic association. *Advances in*  
 1272 *Neural Information Processing Systems*, 15.
- 1273 Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml,  
 1274 M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages  
 1275 2338–2342.
- 1276 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail:  
 1277 a Python toolbox for analyzing and plotting free recall data. *Journal of Open Source*  
 1278 *Software*, 10.21105/joss.00424.
- 1279 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal  
 1280 behavioral and neural signatures of transforming naturalistic experiences into episodic  
 1281 memories. *Nature Human Behavior*, 5:905–919.
- 1282 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). HyperTools: a  
 1283 Python toolbox for gaining geometric insights into high-dimensional data. *Journal of*  
 1284 *Machine Learning Research*, 18(152):1–6.
- 1285 Howard, M. W. and Kahana, M. J. (2002a). A distributed representation of temporal  
 1286 context. *Journal of Mathematical Psychology*, 46:269–299.
- 1287 Howard, M. W. and Kahana, M. J. (2002b). When does semantic similarity help episodic  
 1288 retrieval? *Journal of Memory and Language*, 46:85–98.
- 1289 Huang, L., Holcombe, A. O., and Pashler, H. (2004). Repetition priming in visual search:  
 1290 episodic retrieval, not feature priming. *Memory and Cognition*, 32:12–20.

- 1291 Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental*  
1292 *Psychology: General*, 137(2):324–347.
- 1293 Huber, D. E., Shiffrin, R. M., Lyle, K. B., and Ruys, K. I. (2001). Perception and preference  
1294 in short-term word priming. *Psychological Review*, 108(1):149–182.
- 1295 Isarida, T. and Isarida, T. K. (2007). Environmental context effects of background color in  
1296 free recall. *Memory and Cognition*, 35(7):1620–1629.
- 1297 Jenkins, J. J. and Russell, W. A. (1952). Associative clustering during recall. *Journal of*  
1298 *Abnormal and Social Psychology*, 47:818–821.
- 1299 Jones, A. C. and Pyc, M. A. (2014). The production effect: costs and benefits in free recall.  
1300 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):300–305.
- 1301 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing  
1302 prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 1303 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*,  
1304 24:103–109.
- 1305 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press, New York,  
1306 NY.
- 1307 Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychol-*  
1308 *ogy*, 71:107–138.
- 1309 Kahana, M. J., Howard, M. W., and Polyn, S. M. (2008). Associative processes in episodic  
1310 memory. In Roediger III, H. L., editor, *Cognitive Psychology of Memory*, pages 476–490.  
1311 Elsevier, Oxford, UK.

- 1312 Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., and Yeshurun, Y. (2023). Deeper than  
1313 you think: partisanship-dependent brain responses in early sensory and motor brain  
1314 regions. *The Journal of Neuroscience*, pages doi.org/10.1523/JNEUROSCI.0895–22.2022.
- 1315 Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning  
1316 of memories by context-based prediction error. *Proceedings of the National Academy of*  
1317 *Sciences, USA*, In press.
- 1318 Kimball, D. R., Smith, T. A., and Kahana, M. J. (2007). The fSAM model of false recall.  
1319 *Psychological Review*, 114(4):954–993.
- 1320 Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases  
1321 of human memory. *Current Opinion in Behavioral Sciences*, 32:111–119.
- 1322 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1323 *Handbook of Human Memory*. Oxford University Press.
- 1324 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1325 function? *Psychological Review*, 128(4):711–725.
- 1326 Manning, J. R., Hulbert, J. C., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. A.  
1327 (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic*  
1328 *Bulletin and Review*, 23(5):1534–1542.
- 1329 Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free  
1330 recall. *Memory*, 20(5):511–517.
- 1331 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic  
1332 memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.

- 1333 Manning, J. R., Notaro, G. M., Chen, E., and Fitzpatrick, P. C. (2022). Fitness tracking  
1334 reveals task-specific associations between memory, mental health, and physical activity.  
1335 *Scientific Reports*, 12(13822):doi.org/10.1038/s41598-022-17781-0.
- 1336 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory pat-  
1337 terns in temporal lobe reveal context reinstatement during memory search. *Proceedings*  
1338 *of the National Academy of Sciences, USA*, 108(31):12893–12897.
- 1339 Manning, J. R., Sperling, M. R., Sharan, A., Rosenberg, E. A., and Kahana, M. J. (2012).  
1340 Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clus-  
1341 tering during memory search. *The Journal of Neuroscience*, 32(26):8871–8878.
- 1342 Masicampo, E. J. and Sahakyan, L. (2014). Imagining another context during encoding off-  
1343 sets context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory,*  
1344 *and Cognition*, 40(6):1772–1777.
- 1345 Masís-Obando, R., Norman, K. A., and Baldassano, C. (2022). Scheme representations in  
1346 distinct brain networks support narrative memory during encoding and retrieval. *eLife*,  
1347 11:e70445.
- 1348 McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental*  
1349 *Psychology: Learning, Memory, and Cognition*, 20:507–520.
- 1350 Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman,  
1351 S. J. (2017). The successor representation in human reinforcement learning. *Nature*  
1352 *Human Behavior*, 1:680–692.
- 1353 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental*  
1354 *Psychology: General*, 64:482–488.

- 1355 Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy  
1356 of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1357 Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhi-  
1358 bitionless spreading activation and limited-capacity attention. *Journal of Experimental*  
1359 *Psychology: General*, 106(3):226–254.
- 1360 Oberauer, K. and Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay,  
1361 temporal distinctiveness, or interference? *Psychological Review*, 115(3):544–576.
- 1362 Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., and Radvansky, G. A.  
1363 (2016). Event boundaries and memory improvement. *Cognition*, 148:136–144.
- 1364 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of  
1365 context. *Trends in Cognitive Sciences*, 12:24–30.
- 1366 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). Task context and organization in  
1367 free recall. *Neuropsychologia*, 47:2158–2163.
- 1368 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly*  
1369 *Journal of Experimental Psychology*, 17:132–138.
- 1370 Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of  
1371 associative memory. In Bower, G. H., editor, *The Psychology of Learning and Motivation:*  
1372 *Advances in Research and Theory*, volume 14, pages 207–262. Academic Press, New York,  
1373 NY.
- 1374 Rabinowitz, J. C. (1986). Priming in episodic memory. *Journal of Gerontology*, 41:204–213.
- 1375 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:  
1376 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.



- 1377 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition.  
1378 *Current Opinion in Behavioral Sciences*, 17:133–140.
- 1379 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior.  
1380 *Nature Reviews Neuroscience*, 13:713–726.
- 1381 Rissman, J., Eliassen, J. C., and Blumstein, S. E. (2003). An event-related fMRI investigation  
1382 of implicit semantic priming. *Journal of Cognitive Neuroscience*, 15(8):1160–1175.
- 1383 Romney, A. K., Brewer, D. D., and Batchelder, W. H. (1993). Predicting clustering from  
1384 semantic structure. *Psychological Science*, 4:28–34.
- 1385 Sahakyan, L. and Kelley, C. M. (2002). A contextual change account of the directed  
1386 forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,  
1387 28(6):1064–1072.
- 1388 Sahakyan, L. and Smith, J. R. (2014). A long time ago, in a context far, far away: Retro-  
1389 spective time estimates and internal context change. *Journal of Experimental Psychology:*  
1390 *Learning, Memory, and Cognition*, 40(1):86–93.
- 1391 Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclo-*  
1392 *pedic Reference*, 3:501–506.
- 1393 Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of  
1394 recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.
- 1395 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of  
1396 time. *Neural Computation*, 24:134–193.
- 1397 Sirotin, Y. B., Kimball, D. R., and Kahana, M. J. (2005). Going beyond a single list: modeling

1398 the effects of prior experience on episodic free recall. *Psychonomic Bulletin and Review*,  
1399 12(5):787–805.

1400 Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: a review and  
1401 meta-analysis. *Psychonomic Bulletin and Review*, 8(2):203–220.

1402 Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D., and Zacks, J. M. (2011).  
1403 Changes in events alter how people remember recent information. *Journal of Cognitive*  
1404 *Neuroscience*, 23(5):1052–1064.

1405 Swallow, K. M., Zacks, J. M., and Abrams, R. A. (2009). Event boundaries in perception  
1406 affect memory encoding and updating. *Journal of Experimental Psychology: General*,  
1407 138(2):236–257.

1408 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in*  
1409 *Cognitive Sciences*, 22(3):201–212.

1410 Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *The*  
1411 *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 37:571–  
1412 590.

1413 Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P.,  
1414 and Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–  
1415 82.

1416 Tulving, E. and Schacter, D. L. (1991). Priming and human memory systems. *Science*,  
1417 247:301–305.

1418 Watkins, P. C., Mathews, A., Williamson, D. A., and Fuller, R. D. (1992). Mood-congruent  
1419 memory in depression: emotional priming or elaboration? *Journal of Abnormal Psychol-*  
1420 *ogy*, 101(3):581–586.

- 1421 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American*  
1422 *Journal of Psychology*, 35:396–401.
- 1423 Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming.  
1424 *Current Opinion in Neurobiology*, 8(2):227–233.
- 1425 Xu, X., Zhu, Z., and Manning, J. R. (2023). The psychological arrow of time drives  
1426 temporal asymmetries in retrodicting versus predicting narrative events. *PsyArXiv*,  
1427 page doi.org/10.31234/osf.io/yp2qu.
- 1428 Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., and Hasson, U.  
1429 (2017). Same story, different story: the neural representation of interpretive frameworks.  
1430 *Psychological Science*, 28(3):307–319.
- 1431 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018).  
1432 Is automatic speech-to-text transcription ready for use in psychological experiments?  
1433 *Behavior Research Methods*, 50:2597–2605.
- 1434 Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation  
1435 models in narrative comprehension: an event-indexing model. *Psychological Science*,  
1436 6(5):292–297.
- 1437 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension  
1438 and memory. *Psychological Bulletin*, 123(2):162–185.