Differences in memory for what, where, and when components of recently formed episodes

John J. Sakon1 and Roozbeh Kiani1,2,3
1Center for Neural Science, New York University, New York, New York; 2Neuroscience Institute, NYU Langone Medical Center, New York, New York; and 3Department of Psychology, New York University, New York, New York

Abstract
An integral feature of human memory is the ability to recall past events. What distinguishes such episodic memory from semantic or associative memory is the joint encoding and retrieval of “what,” “where,” and “when” (WWW) for such events. Surprisingly, little work has addressed whether all three components of WWW are retrieved with equal fidelity when remembering episodes. To study this question, we created a novel task where human participants identified matched or mismatched still images sampled from recently viewed synthetic movies. The mismatch images only probe one of the three WWW components at a time, allowing us to separately test accuracies for each component of the episodes. Crucially, each WWW component in the movies is easily distinguishable in isolation, thereby making any differences in accuracy between components due to how they are joined in memory. We find that memory for “when” has the lowest accuracy, with it being the component most influenced by primacy and recency. Furthermore, the memory of “when” is most susceptible to interference due to changes in task load. These findings suggest that episodes are not stored and retrieved as a coherent whole but instead their components are either stored or retrieved differentially as part of an active reconstruction process.

NEW & NOTEWORTHY When we store and subsequently retrieve episodes, does the brain encode them holistically or in separate parts that are later reconstructed? Using a task where participants study abstract episodes and on any given trial are probed on the what, where, and when components, we find mnemonic differences between them. Accuracy for “when” memory is the lowest, as it is most influenced by primacy, recency, and interference, suggesting that episodes are not treated holistically by the brain.

INTRODUCTION
Episodic memories store past events that shape how we remember our lives. Each episodic memory unites disparate streams from our sensory cortices into a combined representation (1). At minimum, episodic memories bind “what,” “where,” and “when” (WWW) components (2, 3). Therefore, studies in animals (4–9) and humans (1, 2, 10–12) have focused on these key components when probing the presence of, and brain structures responsible for, episodic memory.

However, little is known about the strength of association and interactions between the WWW components of episodic memory (13–15). Put more simply: are the what, where, and when components of episodic memory handled with equal fidelity? One possibility is that an episodic memory is a holistic representation in which the WWW components are inseparably intertwined. In this case, encoding and retrieval of each component should invariably be at the same level as the other components. Alternatively, the WWW components may be stored separately and rejoined post hoc during retrieval to synthesize a coherent memory of a past event. Between these two extreme hypotheses, there is a spectrum in which the what, where, and when components of an event are stored or retrieved with various degrees of separability.

A number of studies have investigated this separability between the components of episodes (13–22). Early work that tasked humans to remember series of consonants that differed in either spatial (where) or temporal (when) order found limited differences in recall accuracy once phonemic
coding (saying the letters to remember them) was accounted for (13). A more recent study attempting to probe the separability of spatial versus temporal memory utilized objects with little resemblance to real-world objects to avoid such phonemic confounds (14). After serial presentation of the objects during a study phase, spatial and temporal memory were probed by altering the location and/or order of the objects during the test phase. A change in temporal order affected spatial memory accuracy when subjects were asked to focus on location memory, whereas a change in spatial relations did not affect temporal memory accuracy when subjects were asked to focus on temporal order (14). In addition, a series of studies where subjects saw or heard a sequence of items found better memory of temporal order of the items than of their presented location, but only when subjects were explicitly told to pay attention to serial order (15–17). Another study found better mnemonic accuracy for object or location memory when the task was related to that component (21). Yet other work probing person-location-object associations found dependence between these multi-element events, as after cuing one element of the set accuracy of naming the other two elements was higher than would be expected if the elements were only remembered as pairwise associations (20). This study supports the idea that what-where elements of episodes are remembered as more than just their independent associations but did not assess temporal memory as part of the episodes. A number of neuroimaging studies have also utilized tasks with WWW components, although these studies actively attempt to equilibrate behavior between components [e.g., space and time (23)] to compare region-specific neural responses or explicitly instruct participants to remember one WWW component [e.g., “Which object was viewed first?” (24, 25)]. In summary, although many studies have used tasks to probe the interactions between WWW components, few have attempted to assess the fundamental differences in memory for all three components without explicitly focusing subjects’ attention on particular ones during encoding (26) or retrieval (14–17, 24, 25).

Therefore, we set out to design a task that satisfies two requirements: episodes with what, where, and when components on an equal footing and the ability to probe subjects on any of the individual components without explicitly asking them to focus on that component. In our task, which we call TRANSFER (transient snapshots from episodic recall), subjects first view a movie and are asked to memorize it (encoding phase) (see Fig. 1). The movie contains a series of shapes or “features” that sequentially appear and disappear in different locations, thereby requiring subjects to compare what is shown at different locations and times from memory. Subjects are then probed on the separate WWW components of their memory (retrieval phase) by classifying still images as matches or mismatches from movie content. Mismatch trials only probe one WWW component at a time, thereby allowing us to separately investigate participants’ accuracy for each component via the accuracies of each trial type.

A crucial design principle of the TRANSFER task is to make the sensory signals of the shape, location, and presentation time of features strong and easily distinguishable, thereby leaving no uncertainty about what, where, and when each feature was shown as subjects watched the movie. To put this another way, while watching the movie it would be trivial for subjects to identify that two consecutive features are different from each other (what), appeared in different locations (where), or had a particular order (when). Wittig and colleagues formally tested the ability of humans to do so using a match-mismatch task. For example, if one or two of a large number of possible features are consecutively presented, humans could correctly identify match or mismatch with 98% accuracy (27). Therefore, substantial inaccuracies for mismatch trials in the TRANSFER task must arise from memory mechanisms that underlie the encoding and retrieval of episodes.

Our null hypothesis is that each of the WWW components of episodes is stored and retrieved with equal fidelity. However, trials probing “when” memory showed the lowest accuracy between the mismatch trial types. We find that the mechanisms underlying this difference are due to the serial order of features in the movie, specifically stronger recency and primacy effects for the “when” component. In addition, increased task load on the “what” component by showing two similar features at different times in the movie most strongly affected the accuracy of “when.” Our results support the idea that episodes are not treated as a coherent whole by the brain (28, 29) but instead the WWW components are bound together such that some components can be remembered with more or less fidelity.

**METHODS**

Ten human subjects (5 male, 5 female; age 23.4 ± 4.6 yr, mean ± SD) with normal or corrected-to-normal vision participated in the experiment. All subjects were naive to the purpose of the experiment, provided written informed consent before participation, and were free to leave the study at any time. They received fixed monetary compensation per hour of their participation. We did not utilize online platforms as it would be infeasible to recruit subjects who would commit to tens of 1-h sessions or ensure consistency of data collection conditions (the minimum time a subject spent in our study was 17 h: 14 sessions of data collection and 3 h of training; see below for more information). All procedures were approved by the Institutional Review Board at New York University.

**Behavioral Task**

During the experiment subjects were seated in an adjustable chair in a dimly lit room with chin and forehead supported 57 cm in front of a cathode ray tube display monitor (20 in., EIZO FlexScan T966; refresh rate 75 Hz, screen resolution 1,600 × 1,200). Stimuli were presented with Psychophysics Toolbox (30) and MATLAB (The MathWorks, Inc.). Eye movements were monitored at 1 kHz with an infrared camera (EyeLink, SR Research).

An outline of the TRANSFER task is shown in Fig. 1A. Each trial started with the subject looking at a fixation point (FP, 0.15° radius) at the center of the screen. We define a block as the entire period from this fixation point until the final trial of the retrieval phase was finished, which includes the movie and all the retrieval trials associated with that movie. After a short delay, the encoding phase of the block began with the presentation of a movie centered on the
the sphere (45°, 135°, 225°, or 315° along the circumference). Subjects were required to maintain fixation within 1.5° from the FP throughout the movie. Each feature extended up to 2° from the edge of the sphere in 0.333 s, remained extended for 0.667 s, and then retracted in 0.333 s (total feature duration, 1.333 s). A 2-s delay showing only the sphere was presented after each feature, during which subjects maintained fixation but were permitted to blink to prevent dry eyes.

Features for each movie were chosen randomly from a pool of 470 unique features. This large pool ensured that most features were not repeated in a day (~400 features per day), minimizing interactions between movies from different blocks of trials. However, subjects were unlikely to treat each location-feature combination as a unitized event, as locations were repeated within movies, and features often contained overlapping components from the prototypes, making the combinations difficult to recognize as trial-unique or oddballs (31). Subjects were instructed to utilize the 2-s interval after each feature to memorize and rehearse it. In early data collection blocks, an orange progress bar was shown after each feature, during which subjects maintained fixation within 1.5° from the edge of the sphere in 0.333 s, remained extended for 0.667 s, and then retracted in 0.333 s (total feature duration, 1.333 s). A 2-s delay showing only the sphere was presented after each feature, during which subjects maintained fixation but were permitted to blink to prevent dry eyes.

Features for each movie were chosen randomly from a pool of 470 unique features. This large pool ensured that most features were not repeated in a day (~400 features per day), minimizing interactions between movies from different blocks of trials. However, subjects were unlikely to treat each location-feature combination as a unitized event, as locations were repeated within movies, and features often contained overlapping components from the prototypes, making the combinations difficult to recognize as trial-unique or oddballs (31). Subjects were instructed to utilize the 2-s interval after each feature to memorize and rehearse it. In early data collection blocks, an orange progress bar was shown 3° below the FP. The progress bar grew horizontally at a constant speed throughout the movie (0.3°/s) and indicated the time of each feature. In later blocks of data collection, the progress bar was eliminated for all subjects and they were asked to mentally keep track of elapsed time. Removing the progress bar ensured that subjects could not perform the task by visually associating the length of the bar with simultaneously observed features. Subjects were informed of the change before the progress bar was removed and, because of their strong knowledge of the task structure, typically within one block performed as accurately on these blocks as those with the progress bar. After the movie, subjects moved from the encoding phase to the retrieval phase, where they were tested on their memory of the movie.

Subjects initiated each retrieval trial by fixating on the FP. After a short delay (0.2–0.5 s, truncated exponential with mean = 0.3 s), a blue and a red target (equiluminant, 18.5 cd/m²) appeared to the right and left of the FP (0.5° radius, 8° eccentricity). The location of the two targets swapped randomly across trials. After a second short delay (0.2–0.5 s, truncated exponential with mean = 0.3 s), a portion of the progress bar appeared on the screen cuing subjects to a particular time in the movie. Subjects were instructed to recall the feature shown during the cued time. The length of the progress bar always matched the midpoint time of an extended feature from the movie, meaning there were only one to six possible bar lengths for three- to six-feature movies. Using discrete lengths allowed subjects to only keep track of the order of features and not the absolute times they were shown. Note that even after the progress bar was removed from the movie during the encoding period in later sessions it was still required to cue a time during retrieval. However, well-trained subjects could easily identify the one to six discrete lengths of the bar that referred to features 1–6 from the movie. After a variable delay (0.5–1.5 s, truncated exponential with mean = 0.83 s), an image was shown for 0.3–1.2 s (truncated exponential with mean = 0.6 s) at the center of the screen. Subjects decided whether this image matched the cued time in the movie. Finally, the FP,
stimulus, and time cue disappeared, indicating that subjects should report their choice with a saccadic eye movement to one of the two targets.

Subjects were instructed to identify “match” or “mismatch” trials during the retrieval phase based on the content they saw during the encoding phase. Choosing the blue target indicated that the image matched the cued time in the movie, and the red target indicated a mismatch. Subjects received distinct auditory feedbacks for correct and error responses. We limited the number of retrieval trials per movie block to minimize potential interference due to consecutive retrievals. For movies with three, four, and five or more features, we showed one, one or two, and two or three retrieval trials, respectively. To avoid learning from feedback of previous retrieval trials in a block, each cued time in the movie could be queried only once in the retrieval phase.

For a retrieval image to be a match, it should include the same feature at the same location as those shown in the movie at the cued time. A mismatch could arise for different reasons. Throughout this article we focus on four mismatch trial types: “location-mismatch,” where the retrieval image was a movie frame of a time-matching feature shown in a different location than during the movie (but was a location used in the movie); “feature-mismatch,” where the retrieval image was a movie frame of a mismatching feature taken from a different time in the movie but shown in a time-matching location; “time-mismatch,” where the retrieval image was a movie frame that appeared at a time other than the cued one; and “novel-feature,” where the retrieval image consisted of a feature not shown during the movie (but appearing at a time-matching location). The match and mismatch trials were balanced, each comprising 50% of retrieval trials. Location-, feature-, and time-mismatch trials were balanced too, each comprising 30% of mismatch trials (15% of all trials). Novel-feature trials comprised the remaining 10% of mismatch trials (5% of all trials). The match and mismatch trials were randomly intermixed throughout the session. We note that by using the cue bar to denote a specific time within the movie we aim to probe absolute time instead of relative time on time-mismatch trials, although we are unable to distinguish whether duration or serial order is being probed on these trials.

By the nature of the task, time-mismatch trials always showed the retrieval image of a feature in the original location it was shown in the movie, but with a cue indicating a different time slot from the movie. However, for feature-mismatch trials, since the retrieval image was selected as one of the other features from a different part of the movie, there was a ¼ chance (due to there being only 4 possible locations) that the location of the other feature would happen to be the same location as the original feature shown during the cued time slot. In that case, since the subject would see a feature from a different time of the movie, but in the same location where it was originally shown, the retrieval image was functionally identical to a time-mismatch trial. Therefore, in our analyses we grouped these trials with our time-mismatch trials.

To ensure that consecutive movie blocks did not lead to confusion and interference, the retrieval trials of a block were separated from the movie of the next block with a 3-s interval, during which the screen turned blue to demarcate the end of the previous block. Subjects were instructed to rest and prepare for the next block.

Subjects were trained until they achieved a 75% accuracy criterion (typically for 3–4 days). Data collection began after training. Subjects typically performed the task for 1 h each day, during which they completed six to eight sessions (5–8 min per session), each including 15 blocks. Each subject contributed 2,390–4,877 trials to the data set for the TRANSFER task (median 3,624 trials; 14–28 h of data collection per subject). Additionally, 5 of the 10 subjects contributed 1,216–2,097 trials for the 1-back variant of the TRANSFER task explained below (median 1,466 trials; 10–18 h of data collection per subject). Combined, there were 38,735 trials for the TRANSFER task and a separate 7,416 trials for the 1-back TRANSFER task.

Stimuli

All features were parametrically defined as protrusions of a fixed number of vertices on the surface of a sphere. We started with 62 prototype features that were manually created to be easily discriminable. Next, to increase the number of features such that subjects could not easily memorize the entire set, we took each possible pair of these 62 prototype features and combined them such that the vector of vertices for each feature was exactly halfway between the pair (as explained in the next paragraph). This created 1,922 unique features. We used the 470 of these that displayed a smooth transition when one prototype was morphed into the other. We call these 470 features our main feature set.

Parametric morphing of features was implemented according to the following equation:

\[
V_i(\alpha) = (1 - \alpha)V_i + \alpha V_j
\]

where \(V_i\) and \(V_j\) are the vectors of vertices for any of our 470 pairs of prototypes \(i\) and \(j\), \(\alpha\) is the mixing coefficient that defines the morph level, and \(V_i(\alpha)\) is the vector of vertices of the intermediate morph between the two prototypes. For the main feature set, the morph level \(\alpha\) was set to 0.5.

However, by changing the morph level between 0 and 1, we could also parametrically morph one prototype into another (see Fig. 9). As we lower the value of the morph level from 0.5 toward 0 (or, symmetrically, raise it from 0.5 to 1), we create a morphed feature with increasing dissimilarity from one of the 470 main features (and increasing similarity to one of the 62 prototype features; see Fig. 9). We used such parametrically morphed features in a subset of task blocks as explained in Task Variations.

Task Variations

Paranormal variation of feature similarity.

To test how the similarity of features in the movies influenced behavior, we included one parametrically morphed feature in the movie in some blocks (67% of blocks for 6 subjects and 50% of blocks for the other 4 subjects). The feature was made by selecting one of the 470 main features and decreasing the value of the morph level from 0.5, as explained in Stimuli (Eq. 1). Both the parametrically morphed feature and its partner in the main set, which we refer to as the morph pair, were included in the movie.
Therefore, for mixing coefficients close to 0.5 the movie contained two similar features, whereas for mixing coefficients close to 0 the morph pair were visually distinct and easily discriminable (see Fig. 7). The value of the mixing coefficient in different blocks was chosen randomly from a truncated exponential distribution with the highest probability for $\alpha = 0.5 \pm \epsilon$ and the lowest probability for $\alpha = 0$ (mean $\alpha$, 0.33). $\alpha = 0.5$ was excluded to ensure an objectively correct answer on every trial. Because of this exponential distribution, many of the morph pairs were hard to distinguish (see Fig. 7). The order and time of the two features of the pair varied randomly in the movies across blocks. However, the locations of the pair were always the same in a movie. This location varied in different blocks.

1-Back TRANSFER task.
To test whether subjects' performance depended on the immediate working memory, we created a variant of the TRANSFER task in which retrieval trials in a block did not probe subjects' memory of the movie in the same block. Rather, they referred to the movie observed in the preceding block (1-back). The session began with subjects viewing a movie, then a second movie, followed by the retrieval trials referencing the first movie, then a third movie followed by the retrieval trials referencing the second movie, and so on (see Fig. 9). To perform this task, subjects had to commit the movie of the current block to memory and then retrieve the memory of the movie from the previous block, overcoming two sources of interference: the immediately observed movie and the retrieval trials that followed the relevant movie. Task parameters remained the same as the main task, with the exception that we limited the length of the movies to three or four features. To remind subjects that the retrieval trials were not related to the immediately observed movie in the current block, the color of the time cue of the retrieval trials was distinct from the color of the progress bar of the movie of the current block, but it matched that of the movie of the previous block. Five of ten subjects performed this task variation.

Data Analysis
We tested our hypotheses using a series of regression analyses. Regression coefficients were estimated with maximum likelihood methods. Standard errors of coefficients were estimated by inverting the Hessian matrix of the model log-likelihood function and calculating the square root of the diagonal elements of the inverted matrix (32). For maximum likelihood fitting, we derived the closed form of the Jacobian and Hessian matrices wherever possible. Where we could not directly derive the matrices, we estimated them numerically. Because single-subject data were consistent with each other (see Figs. 2 and 3), we pooled trials across subjects for all statistical tests. We also report single-subject results to provide a complete picture of the strength and diversity of results. Results were similar when using mixed-effects regression models because of the large number of trials collected per subject. Therefore, we report standard regression effects for simplicity and to use the same equations for both single-subject and population data.

To compare accuracy between time-mismatch and other trial types we used the following logistic regression:

$$\logit[P(\text{correct})] = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 G_4$$

where $\beta_0$ are regression coefficients and $G_i$ are indicator variables for different trial types. $G_1$ is for match trials, $G_2$ is for location-mismatch trials, $G_3$ is for feature-mismatch trials, and $G_4$ is for novel-feature trials. Equation 2 served two purposes. First, it enabled us to perform pairwise comparisons between the accuracy of each of the four trial types indicated by $G_1$–$G_4$ and the time-mismatch trials, whose accuracy is captured by $\beta_0$. Second, this equation was used to test whether the null hypothesis of identical accuracy for all five trial types adequately explains the data. For this purpose we use a likelihood ratio test (33) against a logistic regression that captures the null hypothesis that subjects' accuracy was the same for all trial types:

$$\logit[P(\text{correct})] = \beta_0$$

To compare accuracy between novel-feature and other trial types, we used a modified version of Eq. 2:

$$\logit[P(\text{correct})] = \beta_0 + \beta_{14} H_1 + \beta_{24} H_2 + \beta_{34} H_3 + \beta_{44} H_4$$

where $\beta_0$ represents accuracy for novel-feature trials and $H_i$ is for match trials, $H_2$ is for time-mismatch trials, $H_3$ is for
Subjects’ performance in the retrieval trials was influenced by both primacy and recency. Primacy caused features closer to the beginning of a movie to be remembered more accurately. Recency caused features closer to the end of the movie to be recalled more accurately. The combination of the two effects led to a v-shaped deflection in the plots that show retrieval accuracy as a function of cued time in the movie (see Figs. 2 and 6). We used the following logistic regression to quantify the strength of primacy and recency:

$$\logit[P(\text{correct})] = \begin{cases} \beta_0 + \beta_1(\beta_3 - T) & \text{for } T < \beta_3 \\ \beta_0 + \beta_2(T - \beta_3) & \text{for } T \geq \beta_3 \end{cases}$$

(5)

where $T$ is the cued time in the movie and $\beta_i$ are regression coefficients. $\beta_3$ defines the time in the movie when a phase transition from a primacy-dominant regime (first line of Eq. 5) to a recency-dominant regime (second line of Eq. 5) occurred. $\beta_1$ and $\beta_2$ quantify how quickly accuracy changed as a function of the cued time in the movie (the “slope”). We use these slopes to report the strength of primacy and recency, respectively, with significant positive slopes indicative of systematically higher accuracy for features shown at the beginning or end of the movie, respectively. To avoid local maxima, we repeated each fit 500 times from random starting points and chose the fit with the highest log-likelihood.

Because of the variable movie lengths across blocks, we align feature presentation times to either the beginning or end of the movie and fit Eq. 5 separately for each alignment. The reported primacy effects are based on $\beta_0$ of data aligned to the beginning of the movie, and the reported recency effects are based on $\beta_2$ of data aligned to the end of the movie. This was not necessary for Fig. 3B, since when only one movie length was used primacy ($\beta_1$) and recency ($\beta_2$) slope can be measured from the same fit. When fitting Eq. 5 to individual trial types, we once again corrected for familywise error rate for both primacy and recency slopes by Bonferroni-correcting the $P$ values using the number of comparisons ($N = 4$) done for each. We also report a one-way sign test across individual subjects, where we assess significant primacy or recency by fitting $\beta_i$ for each subject’s data with the null hypothesis that the subjects’ slopes will not be different from a distribution with a median of 0.

The discontinuity in Eq. 5 prohibited calculation of the standard errors when $\beta_3$ approached an integer multiple of the interfeature interval in the movies. To obtain an approximate standard error in such cases, we jittered the estimated value of $\beta_3$ by 2% to move the log-likelihood function to a continuous region where the calculation of standard errors was permissible (i.e., diagonals of the inverse Hessian were positive). In all cases we confirmed that the small jitter of $\beta_3$ minimally changed the overall log-likelihood of the model (maximum change <1%). We also confirmed that the jitter minimally influenced the other regression coefficients. To quantify the effects of jitter on other parameters, we kept $\beta_3$ fixed at the jittered value and fit the other coefficients. The jitter led to <1% change in other coefficients.

To compare the magnitude of primacy and recency between two different trial types, we used a modified version of Eq. 5:

$$\logit[P(\text{correct})] = \begin{cases} \beta_0 + \beta_1(\beta_3 - T) + \beta_4(\beta_3 - T)K + \beta_6K & \text{for } T < \beta_3 \\ \beta_0 + \beta_2(T - \beta_3) + \beta_5(T - \beta_3)K + \beta_6K & \text{for } T \geq \beta_3 \end{cases}$$

(6)

where $K$ is an indicator variable for trial type (0 for one type and 1 for another). $\beta_4$ and $\beta_5$ define the difference in primacy and recency slopes between the two trial types, respectively. $\beta_6$ accounts for the overall difference in accuracy between the two trial types. The null hypothesis was no difference in primacy or recency. Since we did three tests, $P$ values were Bonferroni corrected for family-wise error rate.

Movies in morph blocks contained a morph pair, which was intended to increase the task load on “what” memory by tasking subjects to differentiate two similar features (see Task Variations). We quantified the effect of morph feature similarity on the accuracy between different trial types using the following logistic regression:

$$\logit[P(\text{correct})] = \beta_0 + \beta_1GM + \beta_2GM + \beta_3GMP$$

(7)

where $G$ is an indicator variable that contrasts two retrieval trial types, for example location-mismatch versus feature-mismatch trials, $M$ is the morph level, ranging from 0.5 to 0 on morph blocks and 0 on nonmorph blocks, and $P$ is an indicator variable that is 1 when the cued time of the retrieval trial probes one of the morph pair in the movie and 0 otherwise. $\beta_{0-4}$ are regression coefficients, where $\beta_0$ accounts for the accuracy of one trial type and $\beta_1$ accounts for the difference of accuracy with the other trial type, $\beta_2$ accounts for the overall effect of the similarity of morph pair on the accuracy of the two trial types, and $\beta_3$ quantifies how much the difference in accuracy between the two retrieval trial types is influenced by the similarity of the morph pair. $\beta_4$ controls for trials that involve the morph pair, ensuring that $\beta_3$
isolates the effect for the retrieval trials that do not involve the morph pair. The null hypothesis of interest is no difference of accuracy between trial types due to the morph pair ($H_0: \beta_1 = 0$). Results were FWER corrected for the three tests used.

In morph blocks, the morph pair could be shown back to back in the movie or be separated by one or more distinct features. The temporal separation of the morph pair in the movie could change task load via both elapsed time and interference from intervening features. To test whether the number of intervening features (NIF) between the morph pair influenced subjects’ accuracy, we used the following logistic regression:

$$\logit[P(\text{correct})] = \beta_0 + \beta_N N$$

where $N$ is the number of intervening features. The null hypothesis was that the number of intervening features did not influence accuracy ($H_0: \beta_1 = 0$).

To test whether the NIF had a differential effect on the memory of location, feature, or time, we extended Eq. 8 to

$$\logit[P(\text{correct})] = \beta_0 + \beta_N N + \beta_G G + \beta_3 NG$$

where $G$ is the indicator variable for a pair of trial types. $\beta_3$ quantifies whether the accuracy difference of the pair of trial types is modulated by the NIF. The null hypothesis was no modulation ($H_0: \beta_3 = 0$).

## RESULTS

To study how strongly different components of episodes are associated in memory, we developed the TRANSFER task, in which subjects observed a movie and after a short delay recalled events that occurred at particular times of the movie (Fig. 1A). To ensure equal complexity and salience of the events throughout the movie, we created synthetic movies in which a series of three-dimensional (3-D) features sequentially appeared and disappeared on the surface of a central sphere (Fig. 1A, inset). Sequential presentation of features forced subjects to contrast the location, order, and shape of each feature from memory. The features have an equal number of vertices, appear in clearly discrete locations (four locations: 45°, 135°, 225°, or 315° around the circumference), are comfortably separated in time (2 s from end of one feature protrusion until the start of the next), and are easily distinguishable from each other (except in a manipulation condition explained below).

In each block of the task, subjects first observed a movie with three to six features (encoding phase) and then were probed about features shown at particular times in the movie (retrieval phase). Between one and three retrieval trials occurred in the retrieval phase of each block depending on the length of the movie (see METHODS for details). In each retrieval trial, a progress bar cued a time in the movie that subjects had to recall, with the length of the bar in discrete increments of one to six that matched the times of feature extension in the movie. Then, a still image of a feature at a particular location on the sphere was shown. Subjects indicated with a saccadic eye movement to one of two targets whether the shape and location of the feature in the still image matched or mismatched the cued time in the movie.

We designed the mismatches to isolate only one WWW component for each mismatch trial type. These mismatch types included 1) a movie feature that does not match the cued time but is shown in the correct location that matches the cued time (feature-mismatch), 2) a movie feature matching the cued time but shown in a mismatched location (location-mismatch), 3) a movie feature in its correct location but from a different time in the movie (time-mismatch), and 4) a novel feature not shown in the movie (Fig. 1B). Therefore, each of the first three mismatch trial types only changed one component (feature, location, or time) from what was shown in the movie while the other two components were held constant. To perform well, subjects had to memorize episodes that encapsulated the location, shape, and temporal order of the features in the movie. The novel-feature trials tested subjects’ recognition memory, but the other three trial types tested the what, where, and when (WWW) components of the episode. This aspect of the TRANSFER task is unique from previous work in that each of the WWW components of episodes can be separately probed with the feature-mismatch, location-mismatch, and time-mismatch trials, respectively. Importantly, subjects are given no instruction to focus on any particular component of the movie episodes, and the three trial types are tested at equal proportions, making each component of the episodes equally important for achieving high performance.

If the WWW components of episodes are inseparable in memory, we would expect similar accuracies for different mismatch trial types. We tested this hypothesis with two series of analyses: by direct comparison of accuracy between trial types and by determining whether the accuracies for each trial type are equally influenced by three common mnemonic phenomena: primacy, recency, and interference.

### Different Components of Episodes Have Unequal Retrieval Accuracy

The retrieval accuracies are systematically different across trial types (Fig. 2A; average pairwise difference, 5.1 ± 3.5%). These differences are highly significant ($P < 10^{-8}$, likelihood ratio test, Eq. 2 vs. Eq. 3), remained if we removed novel-feature trials ($P = 2.4 \times 10^{-7}$, likelihood ratio test, Eq. 2 vs. Eq. 3), and are consistent across subjects (Fig. 2A). The lowest accuracy is for time-mismatch trials, which is significantly lower than match trials ($\beta_4 = 0.30 \pm 0.060, P = 1.3 \times 10^{-5};$ Eq. 2, FWER corrected), location-mismatch trials ($\beta_2 = 0.23 \pm 0.082, P = 0.024;$ Eq. 2, FWER corrected), feature-mismatch trials ($\beta_3 = 0.43 \pm 0.088, P = 3.3 \times 10^{-5};$ Eq. 2, FWER corrected), and novel-feature trials ($\beta_4 = 1.4 \pm 0.17, P < 10^{-8};$ Eq. 2, FWER corrected). The highest accuracy is for novel-feature trials, which are significantly higher than all other mismatch trials ($\beta_1 = -1.0 \pm 0.17, P < 10^{-8};$ $\beta_2 = -1.4 \pm 0.17, P < 10^{-8};$ $\beta_3 = -1.1 \pm 0.17, P < 10^{-5};$ $\beta_4 = -0.92 \pm 0.18, P = 8.8 \times 10^{-7};$ Eq. 4, all FWER corrected), as expected since a correct answer in these trials can be achieved from recognition memory (31) without recruitment of recall mechanisms.

These differences in accuracy are unlikely to arise from distinct perceptual reliabilities for the WWW components. Rather, memory inaccuracies are most apparent for blocks with longer movies. In blocks with only three
features in the movie (Fig. 2B), all mismatch trial types have accuracies >90% with no significant differences ($P = 0.54$, likelihood ratio test, Eq. 2 vs. Eq. 3; $P \geq 0.19$ for all pairwise comparisons, Eq. 2 without FWER correction), suggesting that subjects can reliably distinguish each WWW component.

What causes lower accuracy on time-mismatch trials in the full data set? We explore the underlying factors that have disproportionate effects on time-mismatch trials in the following sections, which include primacy, recency, and interference.

**Primacy and Recency Effects Differ across Components of Episodes**

When subjects are tasked to remember serially presented items, a common finding is that memory of items presented early (primacy) or late (recency) in the order are remembered better than items presented in the middle (34). We investigated the strength of primacy and recency in our data by plotting the accuracy on retrieval trials as a function of cued time in the movie.

Primacy and recency effects are both present in our experiment. Figure 3A shows changes of retrieval accuracy as a function of cued time from the beginning (Fig. 3A, left) or end (Fig. 3A, right) of the movies. Aggregated across all movie lengths and trial types, accuracy is systematically higher for the earliest ($\beta_1 = 0.21 \pm 0.014, P < 10^{-8}$; Eq. 5) and latest ($\beta_2 = 0.32 \pm 0.044, P < 10^{-8}$; Eq. 5) features in the movies.

These effects are not due to variable movie lengths across blocks as they are present for each movie length over three features. Primacy slope is significantly different from 0 for movie lengths of 4 or more (Eq. 5, movie length 3: $\beta_1 = 0.072 \pm 0.21; P = 0.36$; movie length 4: $\beta_1 = 0.21 \pm 0.084; P = 0.0065$; movie length 5: $\beta_1 = 0.26 \pm 0.043; P < 10^{-5}$; movie length 6: $\beta_1 = 0.28 \pm 0.029; P < 10^{-5}$). Recency slope is significantly different from 0 for movie lengths of 5 and 6 (Eq. 5, movie length 3: $\beta_2 = 0.58 \pm 2.3; P = 0.4$; movie length 4: $\beta_2 = 0.12 \pm 0.098; P = 0.11$; movie length 5: $\beta_2 = 0.29 \pm 0.077; P = 7.6 \times 10^{-5}$; movie length 6: $\beta_2 = 0.62 \pm 0.087; P < 10^{-5}$). Both effects are strongest for the longest movies, as expected (results from blocks with six-feature movies are depicted in Fig. 3B).

However, the strength of primacy and recency effects varied considerably for different memory components (Fig. 4). Whereas time-mismatch trials are strongly influenced by both effects (primacy: $\beta_1 = 0.22 \pm 0.024, P < 10^{-8}$, recency: $\beta_2 = 0.31 \pm 0.07, P = 1.9 \times 10^{-5}$; FWER-corrected Eq. 5), primacy and recency are weak or virtually absent in feature-mismatch (primacy: $\beta_1 = -0.094 \pm 0.61, P = 1.0$, recency: $\beta_2 = 0.36 \pm 0.17, P = 0.064$; FWER-corrected Eq. 5), location-mismatch (primacy: $\beta_1 = 0.22 \pm 0.34, P = 1.0$, recency: $\beta_2 = 0.075 \pm 0.036, P = 0.076$; FWER-corrected Eq. 5), and novel-feature (primacy: $\beta_1 = 0.17 \pm 0.16, P = 0.6$, recency: $\beta_2 = 0.026 \pm 0.11, P < 1.0$; FWER-corrected Eq. 5) trials.

Also, the primacy and recency effects are strongly present in the data from all subjects: 10/10 subjects showed better accuracy for the first cued feature compared with the second ($P = 0.002$, exact binomial test) and for the last feature compared with the second from last ($P = 0.002$, exact binomial test; Fig. 5, A and B). Primacy and recency slopes are also significant when fit to individual subject behavior (primacy: $P = 0.011$; recency: $P = 0.0010$; 1-way sign test, Eq. 5). Primacy for time-mismatch trials is consistent across individual subjects, with 9/10 showing positive primacy slopes (Fig. 5C; $\beta_1 > 0; P = 0.011$, 1-way sign test, Eq. 5). Recency for time-mismatch trials is not as consistent, with only 7/10 subjects showing positive slopes (Fig. 5D; $\beta_2 > 0; P = 0.17$, 1-way sign test, Eq. 5).

To quantitatively compare the susceptibility of different memory components to primacy and recency, we focused on blocks with the longest (6 feature) movies, where primacy and recency are strongest (Fig. 6). The slope of the reduction of accuracy for intermediate features compared to earlier

---

**Figure 4.** Primacy and recency are most pronounced for time-mismatch trials. Conventions are as in Fig. 3A. A: time-mismatch trials. B: feature-mismatch trials. C: location-mismatch trials. D: novel-feature trials.
features are significantly more positive on time-mismatch trials than location-mismatch trials (stronger temporal primacy; $\beta_4 = 0.43 \pm 0.12; P = 0.00084$; Eq. 6, FWER corrected). The slope is also steeper for time-mismatch than feature-mismatch trials as well, although this does not reach statistical significance ($\beta_4 = 0.23 \pm 0.15; P = 0.36$; Eq. 6, FWER corrected). Also, the slope for the increase of accuracy for the features that are shown later in the movie compared with the intermediate features is significantly steeper on time-mismatch trials for both comparisons (stronger temporal recency; compared with location-mismatch trials: $\beta_5 = 0.33 \pm 0.12; P = 0.016$; compared with feature-mismatch trials: $\beta_5 = 0.34 \pm 0.13; P = 0.033$; Eq. 6, FWER corrected). In contrast, the strength of primacy or recency is comparable for location-mismatch and feature-mismatch trials (Fig. 6C; primacy: $\beta_5 = -0.094 \pm 0.092; P = 0.93$; recency: $\beta_5 = 0.17 \pm 0.41; P = 1.0$; Eq. 6, FWER corrected). Overall, our results suggest that the “when” component of episodes is most influenced by primacy and recency.

**Increasing Task Load Differentially Affects the Accuracy of Episode Components**

To further probe the separability of the WWW components of episodes, we asked if we increased the load on one of the components, would each of them be impacted equally? To test this, we used a variant of our task in which the movie included two features with parametrically variable similarity. One movie feature was a mixture of another feature in the movie with a third feature that was not shown in the encoding phase. By changing the mixing coefficient (Eq. 1), we could make the feature pair (the “morph pair”) similar and hard to discriminate, distinct and easy to discriminate, or anything in between (Fig. 7, A and B, METHODS). We note that the increase in task load involves remembering the first feature of the morph pair such that it can be (mentally) compared to the second morph feature when it is presented later in the movie. This load on memory is of course lowest when the pair is shown back to back and becomes progressively higher the longer the subject must hold the memory of the first feature. We also note that there is no additional demand if the participants find the morph pair easy to discriminate, although the mixing coefficient that defines the similarity of the morph pair was randomly drawn from an exponential distribution favoring more similar pairs (METHODS). The percentage of blocks in a session that showed a morph pair in the encoding phase was 66.7% for six subjects and 50.0% for four subjects. Because of the presence of other features in the movie, only a fraction of retrieval trials (~50%) probed the morph pair in those blocks. However, subjects could not know while watching a movie whether the morph pair would

---

**Figure 5.** Primacy and recency plots for individual subjects. Retrieval accuracies are plotted as a function of the time of the cued feature in the movie. Dots show individual subject accuracies and are jittered horizontally for clarity. Black lines are the average for each cued time, dark gray bars are 95% confidence intervals, and light gray bars are standard deviations. Data are combined across blocks with different movie lengths. A: aligned to beginning of the movie, combined across all trial types. B: aligned to end of the movie, combined across all trial types. C and D: same as A and B, but for time-mismatch trials only.

**Figure 6.** Quantitative comparison of the susceptibility of different memory components to primacy and recency. Data are from blocks with 6-feature movies ($N = 4,968$ trials, which represents 31.7% of all trials). In each panel, black represents one trial type and gray represents another trial type (see keys). Lines are fits to Eq. 6. Primacy and recency slopes are larger for time-mismatch trials but comparable for feature- and location-mismatch trials. Plotting conventions are as in Fig. 3B. A: time mismatch vs. location mismatch. B: time mismatch vs. feature mismatch. C: location mismatch vs. feature mismatch.
(or would not) be probed in the retrieval trials. The best strategy is therefore to discriminate and memorize the morph pair as well as other features in the movie. We hypothesized that blocks with two similar features in the movie imposed a higher task load, which could interfere with the storage and subsequent recall of all features.

Does the presence of the morph pair in the movie interfere with accuracy on retrieval trials that do not specifically probe the morph pair? And does the increased task load compromise accuracy on some trial types more than others? The answer to both questions is yes. Our analyses focus on retrieval trials that targeted features other than those in the morph pair, because accuracy for retrieval trials featuring the morph pair approaches chance (50%) as the features become increasingly similar. We observed a reduction of accuracy in the pairwise comparison between time-mismatch and location-mismatch trials of nonmorph features that significantly depended on the similarity of the morph pair ($\beta_3 = -0.91 \pm 0.3$; $P = 0.0078$; Eq. 7, FWER corrected). Furthermore, we found a trend toward morph pair-dependent lower accuracy on time-mismatch than feature-mismatch trials ($\beta_3 = -0.82 \pm 0.35$; $P = 0.063$; Eq. 7, FWER corrected). However, there is no evidence of a morph pair-dependent difference between location-mismatch and feature-mismatch trials ($\beta_3 = -0.098 \pm 0.39$; $P = 1.0$; Eq. 7, FWER corrected). Overall, the “when” component of memory is most affected by the increased load caused by the morph pair.

Our hypothesis about the effect of task load on retrieval accuracy also predicts that longer gaps between the morph pair in movies of the same length would cause lower recall accuracy, as a larger number of intervening features (NIF) makes distinction of the morph pair more difficult and thereby increases memory interference for all features. Figure 8A shows the retrieval accuracy of different trial types as a function of the NIF. Once again, the figure excludes trials that specifically probe the morph pair, since their accuracy approaches 50% as the features in the morph pair become more similar, and focuses on the remaining features of the movie. Since movies have between three and six features, the possible NIF ranged between 0 (when the morph pair occurred back to back) and 4 (6-feature movies with the morph pair shown at the beginning and end). There is a significant decline in retrieval accuracy with increased NIF for time-mismatch (Fig. 8A, left; $\beta_3 = -0.25 \pm 0.046$; $P = 8.4 \times 10^{-6}$; Eq. 8, FWER corrected) and location-mismatch (Fig. 8A, center; $\beta_3 = -0.19 \pm 0.058$; $P = 0.0025$; Eq. 8, FWER corrected) trials. Feature-mismatch trials do not show a significant decline in retrieval accuracy with increased NIF for time-mismatch (Fig. 8A, right; $\beta_3 = -0.026 \pm 0.077$; $P = 1.0$; Eq. 8, FWER corrected). To test whether changes in retrieval accuracy with the NIF are stronger for certain trial types, we performed pairwise comparisons between the slopes in Fig. 8A. The drop in accuracy is significantly steeper for time-mismatch trials than feature-mismatch trials ($\beta_3 = 0.26 \pm 0.099$; $P = 0.028$; Eq. 9, FWER corrected). Time-mismatch trials have a steeper decrease in accuracy for location-mismatch trials but do not reach statistical significance ($\beta_3 = 0.13 \pm 0.081$; $P = 0.30$; Eq. 9, FWER corrected), nor do location-mismatch versus feature-mismatch trials ($\beta_3 = 0.12 \pm 0.11$; $P = 0.72$; Eq. 9, FWER corrected).

These results indicate that accuracy on time-mismatch and location-mismatch trials, which test memory for “when” and “where,” is affected by the number of intervening features between the morph pair, with “when” the most affected. Surprisingly, feature-mismatch trials, which test memory for “what,” are minimally affected. The significant difference between these WWW trial types provides additional evidence that encoding and retrieval of different components of episodes are separable.

The largest reduction in accuracy occurs on time-mismatch trials with a large NIF, suggesting that keeping track of time information is more vulnerable than location.
of feature order is compromised by the difficulty of memorizing morph pairs that are similar and also shown far apart. To further test this hypothesis, we compared accuracy on time-mismatch trials for two types of blocks: four-feature movies with a back-to-back morph pair (A–B–B′–C sequence, where B and B′ are the morph pair and A and C are distinct, nonmorph features) and three-feature movies with no morph pair (A–B–C sequence). Our expectation is that time-mismatch trials for features A and C will have similar accuracies on these blocks, as keeping track of the back-to-back morph pair (NIF = 0) will impose minimal additional load. Indeed, accuracy for four-feature movies with NIF = 0 was 90% (Fig. 8B), while accuracy for three-feature movies without a morph pair was comparable at 92% (Fig. 2B; $P = 0.64$, chi-square test of proportions). These results suggest that the reduction of accuracy for larger NIFs is not a simple effect of perceptual discrimination. Rather, it is due to added memory demands caused by the separation of similar-looking morph pairs by the intervening features.

**Similar Performance in a Task Variation Designed to Interfere with Working Memory**

The relatively short gap (<1 min) between encoding and retrieval in our main task may raise a concern that subjects performed the task by relying solely on working memory. Although recent studies reduce the likelihood of this concern by showing that regions involved in episodic memory also engage in tasks similar to ours (35–37), we designed a control experiment that directly tested the role of working memory in shaping the results explained in previous sections. This experiment, which disrupted working memory, was a “1-back” version of our TRANSFER task, in which subjects responded to the retrieval trials based on the movie shown in the block before (1-back from) the current block (Fig. 9A; METHODS). Therefore, subjects are tasked to hold two movies in mind simultaneously, and responses to retrieval trials are separated from the relevant movie by the retrieval trials from the previous block and the movie of the current block. Five of our ten original subjects performed this task variant.

Despite the high-intensity interference of working memory, subjects have only a small reduction of accuracy for the substantial increase in the complexity of the 1-back task compared with the original TRANSFER task (overall accuracy in nonmorph blocks, 83.7 ± 5.5% on the 1-back task vs. 91.0 ± 4.7% for the original task). Critically, the main effects that demonstrated separability of the three memory components on the original task are also present in the 1-back task. Subjects showed unequal accuracies across trial types, which reached statistical significance (P = 0.0016, likelihood ratio test, Eq. 2 vs. Eq. 3) despite a much smaller trial count (<20%) in the 1-back task. The average pairwise accuracy differences are the same as the original task (6.2 ± 4.6% vs. 5.1 ± 3.5%). Furthermore, the accuracies of different trial types followed a similar pattern as on the original task (Fig. 9B): time-mismatch trials showed the lowest accuracy (80.5 ± 7.1%), whereas novel-feature trials showed the highest accuracy (94.6 ± 7.4%). The trial counts in this task variant are too small to test the effect of task load. However, the primacy (Fig. 9C; $\beta_1 = -0.14 ± 0.086$; $P = 0.054$, Eq. 5) and recency (Fig. 9C; $\beta_1 = -0.3 ± 0.15$; $P = 0.027$, Eq. 5) effects are present and strong. Notably, as with the original TRANSFER task, the strongest serial order effects occurred for time-mismatch trials, which showed significant primacy (Fig. 9D; $\beta_1 = -0.27 ± 0.072$; $P = 0.0001$, Eq. 5) and a trend toward recency (Fig. 9D; $\beta_1 = -0.36 ± 0.25$; $P = 0.077$, Eq. 5).

These similar patterns of accuracies between trial types on the original and 1-back TRANSFER tasks support the hypothesis that the differential mnemonic effects between trial types in the original TRANSFER task are not due to working
memory mechanisms. Instead, they are inherent to the encoding and retrieval of short-term episodes.

**DISCUSSION**

We developed a task to simultaneously investigate memory for the what, where, and when (WWW) components of episodes. The TRANSFER task probes different WWW components with randomly intermixed trials, each isolating one of the components without instructing subjects to focus on any particular WWW component during encoding or retrieval, thereby enabling unbiased comparison of the accuracy between the WWW components. We find that the WWW components of an episode do not form a coherent whole that is inseparably stored and retrieved. Instead, retrieval accuracies differ across components, with memory for “when” showing the lowest accuracy. Furthermore, the WWW components have different susceptibility to primacy, recency, and interference caused by task load—three common phenomena associated with memory. Trials probing “when” show the strongest primacy and recency effects and also are most interfered by increasing task load through a task manipulation that requires discrimination of similar features. These results suggest that memory processes maintain (at least partial) separability and that memory processes maintain (at least partial) separability and that memory processes maintain (at least partial) separability.

At first glance, it might seem that the lower accuracy on time- than location- and feature-mismatch trials is merely due to greater difficulty of perceiving “when” in the TRANSFER task. But perceptual difficulty does not explain our results. As mentioned in INTRODUCTION, if we took any of the WWW components and separately designed a task to probe only one of them (e.g., to probe when, show 2 features serially and ask which came first), accuracy would be at or near 100% (27). Compatible with this expectation, we do not observe noticeable differences in the accuracy of WWW components in blocks with three-feature movies (Fig. 2B); accuracy differences arose for longer movies and increased with movie length. Therefore, unequal accuracies are unlikely to come from differences in sensory discrimination during the movie.

Can the number of each WWW component explain differences in accuracy? Four hundred seventy features are shown in three to six time slots in four locations. But in any given movie, the same number of each WWW component have to be remembered (e.g., a 5-feature movie has 5 features in 5 locations in 5 time slots). If anything, since there are only four possible locations, location-mismatch trials are expected to show the lowest accuracy due to intrusions from the frequent reuse of locations both between and within trials (27). However, despite this potential for redundancy confusion, responses on time-mismatch trials are significantly less accurate than those on location-mismatch trials (Fig. 2). Therefore, difficulty of time-mismatch trials is due to the inherent challenge in keeping track of time or temporal order, with significant drops in accuracy from primacy and recency effects on time-mismatch versus other trial types (Figs. 4 and 6).
If the perception of when is not more difficult than the perception of what or where in the task, what explains the reduction in accuracy on time-mismatch trials? First, we believe the stronger impact of serial order effects on retrieval accuracy for the time of intermediate features (e.g., time slots 3 and 4 in a 6-feature movie; Fig. 3A), compared to their location and shape, provides one mechanism responsible for the overall reduction in the accuracy on time-mismatch trials. Furthermore, for blocks with the addition of the morph pair, interference is strongest when the pair is most separated in time, providing a second mechanism that affects the retrieval of when more than what or where (Fig. 8). Another possibility is that the brain partially sacrifices the encoding quality of when to maintain what and where information. This interpretation is compatible with the results on our morph blocks, where the subject is tasked with increased demand on the perception of what and therefore less accurately memorizes when.

We note, however, that we cannot distinguish absolute time from serial order (38), making it unknown which is responsible for the accuracy deficits on these trials. Furthermore, we cannot pinpoint at what steps in the memory process the differences in memory for when and what/where arise, as there are many steps between the viewing of stimuli and a subject’s responses, including encoding and the unfolding of retrieval (39). Future neurophysiological recordings with the TRANSFER task could help answer these questions.

The mechanisms responsible for lower accuracy on time-mismatch trials are unlikely to be particular to our task design, as an advantage of the TRANSFER task is that it equalizes rehearsal of different memory components, a key factor behind serial order effects. That is, if subjects rehearse all the observed movie features following each new feature in the encoding phase of our task, earlier features will be rehearsed more than later ones, thereby improving their retrieval accuracies (40, 41). However, such automatic rehearsals, if present, would equally apply to the what, where, and when aspects of each feature and therefore are unlikely to underlie different strengths of primacy for different memory components in our task (Figs. 4 and 6). Even if subjects employed a phonemic coding strategy, like attempting to name the abstract features and rehearse the names in order, they are equally rehearsing both what and when, and yet we find that these components have significantly different accuracies (Fig. 2) and primacy slopes (Figs. 4 and 6). In any case, behavior is consistent across subjects, with 10/10 showing both primacy and recency, and coefficients of slopes fit to each subject separately significantly greater than chance. Future work could attempt to remove primacy by eliminating the rehearsal period (the 2-s interfeature interval) or actively preventing rehearsal (40) to test whether time-mismatch trials still show lower accuracy. Similar arguments can be made for recency effects, which can be removed via a distractor (42).

It may be tempting to associate the removal of the progress bar during the encoding phase, which is done within a few sessions of the subjects having qualified by meeting our accuracy threshold, with making when more difficult. On the contrary, subjects showed no changes in accuracy after the removal of the progress bar (METHODS). To be clear, the progress bar is always shown during the retrieval phase in order to reference a time slot from the movie. However, the same length of the bar always represents the same time slot of the movie, e.g., 2/6th of the bar length always represents the second time slot, and 4/6th of the bar length always represents the fourth time slot, regardless of the movie being four or six features long, thereby making the removal of the progress bar trivial to all subjects, who had 3–4 h of training on the task structure before data collection began (METHODS). The accuracy before and after the removal of the progress bar also serves to control for the possibility that subjects merely remember the feature, location, and progress bar as a single unitized “snapshot.” If subjects are memorizing such snapshots to circumvent having to keep track of time, we would expect a drop in accuracy when the progress bar is removed during encoding. Instead, subjects show no difference in accuracy within one session (METHODS) and report the change being easy to overcome.

Serial order effects have been assessed via a perturbation model (43) in which noise causes confusion between an encoded event and its adjoining events with a probability, \( \theta \). For example, the third feature of a movie in our task could be remembered either as the second feature or as the fourth feature with a probability of \( \theta /2 \), with multiple perturbations possible before recall. Values of \( \theta \) near 0.05 have typically been found using letter or word recall tasks (44). The strength of primacy and recency in our time-mismatch trials is compatible with \( \theta \) values of 0.094 ± 0.024 across subjects (mean ± SE), suggesting that serial order effects in our task are more potent than in previous tasks (43, 44). This might be expected, considering that phonemic coding strategies (13) could be less easily utilized with our abstract features.

The abstract visual features in the TRANSFER task and instructions that gave subjects no indication of task goals are integral characteristics of our design. Previous work has had difficulty isolating memory for item and order information (19, 45, 46), largely owing to the use of sequences of consonants or words that introduced phonemic coding confounds (13). Later studies used abstract shapes to avoid such confounds. However, many of these tasks relied on instructions that focused subjects’ attention to one or the other memory component (14–18, 21), making it difficult to parse whether differences in memory effects are specific to the instructions or inherent to the organization of memory. We overcame this challenge by using abstract shapes and only instructing subjects to identify trials as “match” or “mismatch” during the retrieval phase without directing attention to any particular component. The tight control of our stimuli, which can be shaped to create any desired level of diversity and similarity in the what, where, and when domains, also has advantages over tasks that use more lifelike movie segments (25, 47, 48). These experiments do not afford such easy experimental control and require participants to answer targeted questions to probe individual WWW components of memory (e.g., “which scene happened first?”). However, we note that our stimuli are only one of many possibilities that satisfy our criteria, and it would be fruitful to test the generalization of our findings to other stimulus sets in future studies.

Our finding that time-mismatch trials are most affected by the presence of two similar features (the morph pair) in the movie (Fig. 8) could reflect a fundamental limitation of
memory. Since we manipulated the feature similarity, which bears on the what component of episodes, why do we see stronger effects on the when and where components? Considering the present evidence, we suggest that the stronger effect of the morph pair on when and to a lesser degree where is related to separability of these components. In blocks without the morph pair, time-mismatch trials are significantly less accurate than location-mismatch trials, which themselves are less accurate than feature-mismatch trials (Fig. 2). If the memory system has a limited mnemonic reservoir (36, 49, 50), the added load caused by the morph pair would impact the hardest trials more strongly.

Although we cannot establish the engagement of structures responsible for remembering episodes without neurophysiological recordings, we believe the setup of our task promotes the use of episodic memory. At first glance our stimuli resemble those used in studies of perception or working memory, but our task deliberately deviates from standard designs in these fields to engage episodic mechanisms. We consider four lines of evidence to support this claim. First, each of the movies includes multiple features, locations, and times, and the difficulty of remembering these combinatorial cues in a three-to-six-feature movie is likely to exceed the low capacity of working memory. Typically, young, healthy subjects can hold three or four simple objects in visual working memory (51, 52). And as suggested from studies of hippocampal lesion patients (36, 53) and single-unit recordings in human patients (35), the episodic mechanisms of the medial temporal lobe are recruited when tasks exceed working memory capacity.

Second, the features are presented sequentially, so that later features interfere with the working memory of the earlier features. Although strategies such as rehearsal can be used to overcome such interference, rehearsal itself may be dependent on episodic memory (54).

Third, in a 1-back version of the task that substantially increases the task load and amplifies interference, we find similar results as the normal version of the task (Fig. 9). The recall trials of a block in the 1-back task are separated from their corresponding movie by the recall trials of the preceding block and the movie of the subsequent block. It is exceedingly unlikely that short-term or working memory can survive all of these manipulations and interferences. This point is corroborated by our debriefing of the participants in our task and by our own impression as subjects in the pilot studies, as successful performance in the task typically depended on active construction of an episode with a corresponding mental narrative (e.g., I saw a featherlike feature growing at the top, followed by a hammerlike feature growing at the bottom of the sphere, and so on). Consequently, we do not think that purely perceptual or working memory mechanisms are sufficient for our task or match the mental processes that subjects adopted.

Finally, although the “where” component of episodic memory typically refers to an egocentric sense of location, allocentric viewing of space, as we operationalize here with location-mismatch trials, is likely to be tracked by medial temporal lobe structures in the service of episodic memory (55). Reinforcing this link, recent descriptions of saccadic temporal lobe structures in the service of episodic location-mismatch trials, is likely to be tracked by medial centric viewing of space, as we operationalize here with ory typically refers to an egocentric sense of location, allo-

**“I meet you. I remember you… Why not you in this city and in this night, so like other cities and other nights you can hardly tell the difference?”** Elle’s soliloquy in Marguerite Duras’s *Hiroshima Mon Amour* laments on the fungibility of our episodic memories, where it is often difficult for us to separate events in time when they share similar components. In our work, when humans are tasked to remember what, where, and when components of short-term episodes, we show evidence for the difficulties she describes. In particular, we find that temporal information is the least accurate and most susceptible to interference by serial order and task load, especially when the recalled episode is surrounded by a sequence of similar episodes. These differences in accuracy between the what, where, and when components of episodes in our experiments suggest a degree of independence between memory components, which implies that the apparent coherence of recalled events is largely a product of a reconstruction process (28, 29). Future work utilizing neural recordings will offer a window into the mechanisms behind each memory component (57) and how the brain reconstructs them into the seamless yet overlapping episodes we experience on an everyday basis.

**DATA AND CODE AVAILABILITY**

Full data are available upon request. Analysis code (in MATLAB) and preprocessed data to reproduce figures are available upon request.

**ACKNOWLEDGMENTS**

We thank Gouki Okazawa, Michael Waskom, Saleh Esteki, and Michael Kahana for helpful discussions.

**GRANTS**

This research was supported by the National Institutes of Health (RO1-MH109180), Simons Collaboration on the Global Brain (542997), and a Pew Scholarship in the Biomedical Sciences.

**DISCLOSURES**

No conflicts of interest, financial or otherwise, are declared by the authors.

**AUTHOR CONTRIBUTIONS**

J.J.S. and R.K. conceived and designed research; J.J.S. performed experiments; J.J.S. analyzed data; J.J.S. and R.K. interpreted results of experiments; J.J.S. prepared figures; J.J.S. drafted manuscript; J.J.S. and R.K. edited and revised manuscript; J.J.S. and R.K. approved final version of manuscript.

**REFERENCES**


Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of


