Adaptive memory distortions are predicted by feature representations in parietal cortex

Yufei Zhao (赵雨菲)[1], Avi J. H. Chanales[2], & Brice A. Kuhl[1]

Department of Psychology, University of Oregon[1], 97401

Department of Psychology, New York University[2], 10016

[1]

Abbreviated title: Memory distortions in parietal cortex

Corresponding author email address: Yufei Zhao (yzhao17@uoregon.edu) or Brice Kuhl (bkuhl@uoregon.edu)

Pages: 35

Figures: 4

Tables: 1

Number of words: Abstract (152), Introduction (696), discussion (1448)

Acknowledgements: This work was supported by NIH grant NINDS R01-NS107727 and NSF CAREER Award BCS-1752921 to B.A.K.

[1] The authors declare no competing financial interests.

ABSTRACT

Similarity between memories is a primary cause of interference and forgetting. Exaggerating subtle differences between memories is therefore a potential mechanism for reducing interference. Here, we report a human fMRI study (n = 29, 19 female) that tested whether behavioral and neural expressions of memories are adaptively distorted to reduce interference. Participants learned and repeatedly retrieved object images, some of which were identical except for subtle color differences. Behavioral measures of color memory revealed exaggeration of differences between similar objects. Importantly, greater memory exaggeration was associated with lower memory interference. fMRI pattern analyses revealed that color information in parietal cortex was stronger during memory recall when color information was critical for discriminating competing memories. Moreover, greater representational distance between competing memories in parietal cortex predicted greater color memory exaggeration and lower memory interference. Together, these findings reveal that competition between memories induces adaptive, feature-specific distortions in parietal representations and corresponding behavioral expressions.

Keywords: episodic memory, interference, repulsion, fMRI, pattern similarity

Significance Statement (120 words)

Similarity between memories is a primary cause of interference and forgetting. Here, we show that when remembering highly similar objects, subtle differences in the features of these objects are exaggerated in memory in order to reduce interference. These memory distortions are reflected in, and predicted by, overlap of activity patterns in lateral parietal cortex. These findings provide unique insight into how memory interference is resolved and specifically implicate lateral parietal cortex in representing feature-specific memory distortions.

INTRODUCTION

Given the vast number of memories that humans store, overlap between memories is inevitable. For example, one may have taken multiple vacations to the same town or parked in the same garage on many occasions. There is a long history of behavioral studies in psychology documenting the many contexts in which this type of overlap leads to memory interference and forgetting (Anderson & Spellman, 1995; Barnes & Underwood, 1959; Mensink & Raaijmakers, 1988; Osgood, 1949; Wixted, 2004). As a result, a primary focus of theoretical models of memory has been to specify the computational mechanisms by which interference is resolved (Colgin, Moser, & Moser, 2008; O'Reilly & McClelland, 1994; Treves & Rolls, 1994). These models have largely focused on how memories are encoded so that the content of memories is protected against interference. An alternative perspective, however, is that instead of protecting memories from interference, there is adaptive value in allowing the content of memories to be *shaped* by interference (Hulbert & Norman, 2015; Kim, Norman, & Turk-Browne, 2017). Specifically, to the extent that overlap across memories is the root cause of interference, then distorting memories to reduce this overlap is a potentially effective remedy.

Evidence from recent neuroimaging studies hints at the idea that memory representations are distorted as an adaptive response to interference. Namely, several studies have found that when similar events are encoded into memory, this triggers a targeted exaggeration of differences in patterns of activity in the hippocampus (Ballard, Wagner, & McClure, 2019; Chanales, Oza, Favila, & Kuhl, 2017; Dimsdale-Zucker, Ritchey, Ekstrom, Yonelinas, & Ranganath, 2018; Favila, Chanales, & Kuhl, 2016; Hulbert & Norman, 2015; Kim et al., 2017; Schapiro, Kustner, & Turk-Browne, 2012; Schlichting, Mumford, & Preston, 2015). The key observation in these studies is that similar memories 'move apart' from each other in representational space, suggesting a form of memory repulsion. Yet, a critical limitation of these studies is that the feature dimensions along which memories move are underspecified. That is, do changes in neural representations correspond to changes in the information content of memories? On the one hand, neural activity pattern may become separated without any changes to underlying memories. Alternatively, changes in neural activity patterns may reflect adaptive changes in memory content. For example, if two vacations

to the same city were associated with different weather conditions, then weather-related information may be a salient component of corresponding memories and weather-related *differences* between those vacations may be exaggerated to improve memory discriminability (e.g., "That was the year it was *really cold*," vs. "That was the year it was *really hot*").

While it has proven difficult to translate hippocampal activity patterns to explicit feature dimensions (LaRocque et al., 2013; Liang, Wagner, & Preston, 2013), feature dimensions are far more accessible in (or decodable from) neocortical regions involved in memory retrieval. In particular, there is rapidly growing evidence that lateral parietal cortex carries detailed information about the content of retrieved memories (Chen et al., 2017; Long, Lee, & Kuhl, 2016; Xiao et al., 2017) and amplifies behaviorally-relevant information (Favila, Samide, Sweigart, & Kuhl, 2018; Kuhl, Johnson, & Chun, 2013). Moreover, recent studies have shown that memory representations in parietal cortex can be decomposed into separable feature dimensions (Bone, Ahmad, & Buchsbaum, 2020; Favila et al., 2018; Lee, Samide, Richter, & Kuhl, 2019). Thus, lateral parietal cortex may provide a unique window into how memory representations are shaped by interference.

Here, we tested whether interference between highly similar memories triggers adaptive distortions in parietal memory representations and corresponding behavioral expressions of memories. Our motivating theoretical perspective was that subtle differences between similar memories are prioritized and exaggerated to reduce the potential for interference. To test these ideas, we modified a recent behavioral paradigm that demonstrated adaptive biases in long-term memory for objects (Chanales, Tremblay-McGaw, & Kuhl, in-press). We predicted that competition between memories for similar objects would trigger a memory-based exaggeration of subtle differences between those objects, and that greater exaggeration would be associated with lower memory interference. Using pattern-based fMRI analyses, we tested whether memory representations in lateral parietal cortex (a) preferentially express features that are critical for discriminating similar objects and (b) predict feature-specific distortions in behavioral expressions of memory.

Materials and Methods

Participants

Thirty-two (21 female; mean age = 23.5 years) right-handed, native English speakers from the University of Oregon community participated in the experiment. Three participants were excluded from analysis (two due to falling asleep inside the scanner, one due to technical error), resulting in a final set of 29 participants (19 female; mean age = 23.7 years) included in data analysis. Participants were screened for motion during the scanned recall tasks, but no participants exceeded the exclusion criteria (mean framewise displacement > 0.25) for any of the runs. The sample size was comparable to similar fMRI studies in the field. All participants had normal or corrected-to-normal vision. Informed consent was obtained in accordance with the University of Oregon Institutional Review Board.

Overview of Experimental Paradigm

We modified a paradigm from a recent behavioral study that was used to demonstrate adaptive biases in long-term memory for object colors (Chanales, Tremblay-McGaw, & Kuhl, in-press). In the prior (and current) study, participants learned associations between faces and object images. Critically, the objects contained 'pairmates' for which the object images were identical except for their color (e.g., a blue backpack and a purple backpack), and successful learning required discriminating between these pairmates. In the current study, we used a two-day procedure in which participants received extensive behavioral training on face-object associations on Day 1 and then returned on Day 2 for additional behavioral training, followed by an fMRI session, and finally a behavioral color memory test (Fig. 1). A critical feature of our design is that we held color similarity between pairmates constant (24 degrees apart), but we included a competitive and non-competitive condition (Fig. 1b). In the competitive condition, pairmate images corresponded to the *same object category* (e.g., two beanbags of slightly different colors). In the non-competitive condition, pairmates corresponded to *distinct object categories* (e.g., a pillow and a ball of slightly different colors). Thus, in both conditions the pairmates were 24 degrees apart in color space; but, for the competitive

condition, color was the only feature dimension on which the pairmates differed. In contrast, for the non-competitive condition, object category also differed between pairmates. Thus, although color distance between pairmates was matched across conditions, color information was more important in the competitive condition. For the fMRI session, participants were shown faces, one at a time, with the only instruction being to retrieve corresponding objects as vividly as possible. An important feature of our procedure is that participants were not explicitly instructed to retrieve color information during the fMRI scans, nor had color memory been tested at any point prior to scanning. Rather, we only tested color memory after participants exited the scanner.

Stimuli

Participants learned associations between 24 object images and 24 images of white male faces. The 24 object images corresponded to 18 distinct object categories (e.g., beanbag, hat, umbrella, balloon) and 12 distinct color values. Thus, some of the 24 object images were from the same object category (e.g., two beanbags) or had the same color value. The object images were generated from an image set that allowed for each image's color to be rotated along a 360° color wheel (Brady, Konkle, Alvarez, & Oliva, 2013). To assign colors to each object, the 360° color wheel was divided into 15 evenly spaced color values (0°, 24°, 48°, etc.). These 15 values were arbitrarily chosen but were fixed across participants. For each participant, 6 consecutive color values were selected (randomly positioned among the set of 15 color values) for the competitive condition. For example, color values of 48°, 72°, 96°, 120°, 144°, and 168° might be selected for the competitive condition (Fig. 1b). Likewise, 6 consecutive color values were selected for the non-competitive condition. The 6 values for the non-competitive condition always 'started' 48° after the competitive color values 'ended.' For example, if the color values for the competitive condition spanned 48° to 168°, then the color values for the non-competitive condition would be 216°, 240°, 264°, 288°, 312°, 336° (Fig. 1b).

For both conditions, the 6 color values were clustered into 3 sets of consecutive color values: e.g., 48° and 72°, 96° and 120°, 144° and 168°. Each of these sets included a total of 4 object images (resulting in 12 object images for each condition). For the competitive condition, the four images in each set represented two color values (e.g., 48° and 72°) and two object categories (e.g., beanbag and jacket). For example, the set might include a 48° beanbag, a 72° beanbag, a 48° jacket and a 72° jacket (Fig. 1b). Object images within each set that were from the same object category (e.g., the 48° beanbag and the 72° beanbag) are referred to as 'pairmates.' For the non-competitive condition, the four images in each set represented two color values (e.g., 216° and 240°) and four distinct object categories (Fig. 1b). Although none of the object images in the non-competitive condition were from the same object category, the four images in each set were also divided into pairmates, with pairmates being images from distinct object categories and, as in the competitive condition, with color values 24° apart. For example, if a set in the non-competitive condition included a 216° lunchbox, a 216° pillow, a 240° hat, and a 240° ball, the 216° lunchbox and the 240° hat might be arbitrarily designated as one set of pairmates and the 216° pillow and the 240° ball as the other set of pairmates. These non-competitive pairmates functioned as a critical control condition for behavioral and fMRI analyses (see *fMRI Pattern Similarity Analyses*, below).

The mapping between the 24 object images and the 24 face images was randomly determined for each participant. All face and object images were 250 * 250 pixels.

Pre-scan face-object training

Participants completed the experiment on two consecutive days (Fig. 1a). On Day 1, participants learned 24 face-object associations across 14 training rounds. Each training round consisted of a study phase and an associative memory test phase. During study phases, participants were presented with the 24 face-object associations, one association at a time, in random order. Each trial started with a fixation cross presented in the center of the screen (1.5 s), followed by the face-object association (3.5 s). Faces were presented to the left of the objects. During the associative memory test phases, object images were

presented at the top of the screen with four face choices below. The four face choices always included the target face (i.e., the face associated with the presented object image), the pairmate's face (i.e., the face that was associated with the presented object's pairmate), and two foil faces (associated with non-pairmate objects). Participants were asked to select the face that was associated with the presented object. After responding, participants received feedbacks indicating whether or not they were correct and showing the correct face-object association for 1.5 s. Each trial in the associative memory test was self-paced up to a maximum of 8 s. On Day 2, participants completed 4 additional training rounds immediately prior to entering the fMRI scanner. The procedure was the same as on Day 1.

Scanned perception and cued recall tasks

During fMRI scanning, participants completed 6 consecutive rounds of a perception task and 6 consecutive rounds of a cued recall task (each round corresponded to a separate fMRI scan). The order of the perception and cued recall tasks was counterbalanced across participants. In the perception task, each trial presented one of the 24 object images in the center of the screen for 0.5 s followed by a fixation cross for 3.5 s. A black cross was embedded within the object images at a random location on 25% of trials and participants were instructed to make a button press whenever they detected a black cross. In each perception round, each object image was presented twice, in block randomized order. Participants were instructed to remain centrally-fixated, on a white fixation cross, throughout each perception run. Each perception round contained a 10 s null trial (fixation cross only) at the beginning and end of each scan and 12 null trials (4 s each) randomly distributed throughout the run. Here, we do not consider data from the perception task because (a) our primary hypotheses related to participants' *memories* for the object images and (b) subtle color differences between were more to detect in the scanner environment.

In the cued recall task, each trial started with one of the 24 face images presented at the center of the screen for 0.5 s, followed by a blank screen for 2.5 s, and then a question mark for 1 s. Participants were instructed to recall the object image that was associated with the presented face as vividly as possible and

to hold the image in mind throughout the trial. Participants were instructed to rate the vividness of their memories ('vivid' or 'not vivid') via a button box response when the question mark appeared. The question mark was followed by a fixation cross for 2 s before next trial began. Responses were recorded during the trial and during the 2 s fixation cross between trials. Together, the intertrial interval was 6 s. All face-object associations were tested twice in each retrieval round, in block randomized order. Each retrieval round contained a 10 s null trial (fixation cross only) at the beginning and end of each scan and 12 null trials (4 s each) randomly distributed throughout the run.

Post-scan behavioral tests

After participants completed the perception and cued recall tasks, they exited the scanner and completed five rounds of the color memory test. During the color memory test, each trial began with one of the 24 face images presented on the left side of the screen and the corresponding object image presented on the right of the screen. Importantly, the object image was initially in grey scale. Participants were instructed to move a cursor along a color wheel (Fig. 1a, c) to adjust the color of the object to the remembered color value. Participants clicked the mouse to record their response and then moved on to the next trial. Each face-object association was tested once per round and the task was self-paced. After completing the five color memory test rounds, participants completed two final rounds of the associative memory test—the same task they completed during the training rounds on Day 1 and just prior to fMRI scanning. The sole purpose of the post-scan associative memory test was to motivate participants to maintain their effort and memory accuracy throughout the fMRI session as the post-scan associative memory test was used to determine a monetary bonus for participants (a fact which participants were made aware of prior to the fMRI scan).

Measuring color memory bias

The post scan color memory test was used to measure participants' color memory for each object image. However, rather than focusing on the accuracy of recall, we were critically interested in recall bias. Bias

was measured in two ways. The first measure—*mean signed distance*—was computed by first averaging the responses across the 5 color memory test trials for each object image. The difference between the mean response and the actual color value for a given object image reflects the color memory distance for that object image. Critically, if the mean response was biased *away* from the color of the pairmate object (Fig. 1c), the distance measure was positively signed; if the mean response was biased *toward* the color of the pairmate object (Fig. 1c), the distance measure was negatively signed. By averaging the signed distance measure across the 12 object images within each condition, the mean signed distance was computed for each condition (competitive, non-competitive) and for each participant. The second measure—*percentage of away responses*—was computed by ignoring the distance between participants' responses and the actual color values and instead simply computing the percentage of responses that were biased away from the color of the pairmate object. It is important to note that this measure was computed at the trial level. Thus, for a given object image, if a participant recalled the object's color 'away from' the pairmate on 4 out of the 5 test trials for that object image, the percentage of away responses *for that object image* would be 80%. Although we did not expect (or observe) notable differences between the two measures (mean signed distance and percentage of away responses), the percentage of away responses addressed the concern that any observed effects for the mean signed distance measure were driven by a few extreme responses.

fMRI data acquisition

Imaging data were collected on a Siemens 3 T Skyra scanner at the Robert and Beverly Lewis Center for NeuroImaging at the University of Oregon. Functional data were acquired using a T2*-weighted multiband EPI sequence with whole-brain coverage (repetition time = 2 s, echo time = 36 ms, flip angle = 90°, multiband acceleration factor = 3, inplane acceleration factor = 2, 72 slices, 1.7 × 1.7 × 1.7 mm voxels) and a 32-channel head coil. Note that due to an *a priori* decision to focus on visual and parietal cortical areas, we used a high-resolution protocol that fully covered visual/parietal regions but only partially covered frontal cortex. Each perception scan (6 total) consisted of 130 total volumes. Each retrieval scan (6 total) consisted

of 190 total volumes. Oblique axial slices were aligned parallel to the plane defined by the anterior and posterior commissures. A whole-brain T1-weighted MPRAGE 3D anatomical volume (1 × 1 × 1 mm voxels) was also collected.

fMRI data preprocessing

fMRI data preprocessing was performed using fMRIPrep 1.3.1 (Esteban et al., 2019). The T1-weighted (T1w) image was corrected for intensity non-uniformity with N4BiasFieldCorrection (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh (ANTs 2.2.0) with OASIS30ANTs as the target template. Brain surfaces were reconstructed using recon-all from FreeSurfer 6.0.1 (Dale, Fischl, & Sereno, 1999). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0). For the functional data, susceptibility distortion corrections were estimated using 3dQwarp (Cox & Hyde, 1997). The BOLD reference was then co-registered to the T1w reference by bbregister (FreeSurfer) using boundary-based registration with nine degrees of freedom (Greve & Fischl, 2009). Head-motion parameters were estimated by mcflirt from FSL 5.0.9 (Jenkinson, Bannister, Brady, & Smith, 2002). Slice-time correction was done by 3dTshift from AFNI 20160207 (Cox & Hyde, 1997). Functional data were smoothed with a 1.7 mm FWHM Gaussian kernel and high pass filtered at 0.01Hz. Smoothing and filtering were done with the Nipype pipeline tool (Gorgolewski et al., 2011).

Response estimates were obtained for each trial (one regressor per trial, 4 s duration) in each cued recall run using the “least-squares separate” method (Mumford, Turner, Ashby, & Poldrack, 2012). With this method, each item was estimated in a separate GLM as a separate regressor while all remaining items were modeled together with another regressor. The six movement parameters and framewise displacement were included in each GLM as confound regressors. This resulted in *t* maps that were used for the pattern similarity analysis. Given that all analyses averaged data across multiple trials—mitigating the influence of any one trial—we did not perform any data exclusion for outliers at the trial level.

Regions of interest

fMRI analyses were conducted using a set of visual and parietal regions of interest (ROIs) that were identical to those used by Favila, Samide, Sweigart, & Kuhl (2018) to measure object and color representations during memory recall. While our primary focus was on the parietal ROIs, we anticipated that visual regions might also reflect feature-specific information during memory retrieval. For low level visual regions, we combined bilateral V1v and V1d as V1 and combined bilateral LO1 and LO2 as LO based on Wang, Mruczek, Arcaro, & Kastner (2014). For high level visual regions, we generated a VTC ROI by combining bilateral fusiform gyrus, collateral sulcus, and lateral occipitotemporal sulcus derived from the output of Freesurfer segmentation routines. For lateral parietal cortex, we referenced Yeo et al. (2011)'s 17-network resting state atlas. The parietal nodes from Network 12 and 13 (subcomponents of the frontoparietal control network) are referred to as dorsal lateral intraparietal sulcus (dLatIPS) and ventral lateral intraparietal sulcus (vLatIPS), respectively. For the parietal node of Network 5 (dorsal attention network), we separated it along the intraparietal sulcus to create a dorsal region we refer to as posterior intraparietal sulcus (pIPS) and a ventral region we refer to as ventral IPS (vIPS) (Sestieri et al., 2017) . The vertices in lateral occipital cortex were eliminated in these two regions. The parietal nodes of Networks 15–17 (subcomponents of the default mode network) were combined into a region we refer to as angular gyrus (AnG).

For post hoc analyses, we generated medial temporal and hippocampus subfield ROIs using ASHS (Yushkevich et al., 2015). We selected bilateral CA1, subiculum, entorhinal cortex, and parahippocampal cortex. We combined CA2, CA3 and dentate gyrus into a single ROI (CA23DG) and combined BA35 and BA36 into a perirhinal cortex ROI.

fMRI Pattern similarity analyses

Pattern similarity analyses were used to measure the similarity of fMRI activity patterns for various pairs of object images during the cued recall task. To calculate pattern similarity, we first computed the mean activity pattern for each of the 24 recalled objects by averaging *t* maps for odd runs and even runs separately. Pearson correlations were then computed between the mean *t* map of odd runs and even runs. All the correlations were *z*-transformed (Fisher's *z*) before subsequent analyses. All analyses were performed in the participant's native T1w space and were done separately for each ROI. Pattern similarity analyses focused on three specific correlations within each 'set' of 4 object images (see Fig. 1b and *Stimuli* for explanation of 'sets'): (1) 'Pairmate correlations' (see *Stimuli* for definition of pairmates), (2) 'Same-color correlations,' which refer to correlations between object images from different object categories but with identical color values (Fig. 1b), and (3) 'Baseline correlations,' which refer to object images from different object categories and different color values (24 degrees apart; Fig 1b). Again, it is important to emphasize that all pattern similarity analyses were performed within the sets of 4 object images and, critically, the same correlations were applied for the competitive and non-competitive conditions.

Neural representation of color information

To test whether representation of color information was stronger in the competitive condition than the non-competitive condition, we first obtained (for each condition, ROI, and participant) the mean 'Same-color correlation' and the mean 'Baseline correlation.' Both of these correlations reflect correlations between object images from different object categories (Fig. 1b), but the same-color correlation reflects images with identical color values whereas the baseline correlation reflects images with a 24° difference in color. Thus, the difference between these measures (same-color – baseline) isolates color-related similarity. Of critical interest was whether this color-related similarity was stronger in the competitive condition than the non-competitive condition. Critically, color similarity was *objectively identical* across conditions, but we predicted stronger color representation in the competitive condition owing to its greater diagnostic value in the competitive condition. It is important to note that the inclusion of a separate baseline correlation for each condition (competitive, non-competitive) controlled for potential global similarity differences between

conditions (i.e., that correlations among *all pairs* of object images might be higher in one condition vs. the other).

Neural similarity between pairmates

To test whether similarity between pairmates was stronger in the competitive condition than the non-competitive condition, we first obtained (for each condition, ROI, and participant) the mean 'Pairmate correlation' and the mean 'Baseline correlation.' For the competitive condition, pairmate correlations reflect object images from the same object category but with a 24° difference in color (Fig. 1b). For the non-competitive condition, pairmate correlations reflect object images from different object categories, again with a 24° difference in color (Fig. 1b). Thus, pairmate similarity was *objectively greater* in the competitive condition than the non-competitive condition. For both conditions, the baseline correlations reflect object images from different object categories and with a 24° difference in color. Thus, the difference between these measures (pairmate – baseline) was intended to isolate object-related similarity (specifically for the competitive condition). As with the color information analysis, the condition-specific baseline correlations controlled for potential global similarity differences between conditions.

Neural measures of pairmate similarity predict color memory repulsion

To test whether similarity between vIPS representations of pairmates during competitive recall predicted the degree to which there was repulsion of color memories (as measured in the post-scan color memory test), we first computed the mean signed color memory distance for the two objects in each set of pairmates. This yielded a single value representing the distance between a given set of pairmates, with greater distance reflecting greater repulsion. Next, for vIPS we computed *dissimilarity* between each set of pairmates, as defined by: 1 – the Pairmate correlation. (Note: for this analysis we used dissimilarity, as opposed to similarity, simply for ease of interpretation). Thus, for each participant and for each condition (competitive, non-competitive), this resulted in 6 values representing color memory distance between each

set of pairmates and 6 values representing vIPS dissimilarity between each set of pairmates. We then performed a Spearman correlation between these two measures. For each condition, one-sample *t*-tests were performed on the participants' *z*-transformed Spearman's r_s values to test whether the mean correlation between color memory distance and vIPS dissimilarity differed from 0. For comparison, similar analyses were also performed for other ROIs (Table 1).

To better visualize the relationship between color memory distance and vIPS dissimilarity, for each participant the 6 pairmates in the competitive condition were divided into three bins (2 pairmates per bin) based on vIPS pairmate dissimilarity (low, medium, high). We then computed the mean signed color memory distance (from the post-scan color memory test) and the mean associative memory accuracy (from the pre-scan associative memory test) for each of these bins. One-way ANOVA was used to test whether mean signed distance and/or mean associative memory accuracy varied as a function of vIPS dissimilarity bin. Finally, we performed a multilevel mediation analysis to test whether color memory mediated the relationship between vIPS pairmate dissimilarity and associative memory accuracy. This analysis was performed by obtaining, for each participant, the mean color memory distance, vIPS dissimilarity, and associative memory performance for each of the 6 pairmates in each condition. Mediation analyses included a random intercept for each participant, but random slopes were not included due to the small number of data points per condition/participant.

Statistical analysis

Statistical analyses were performed using R version 3.6.3. All *t*-tests were two-tailed, with $\alpha = 0.05$. All repeated measures ANOVAs were computed with the afex package using Type III sums of squares. Effect sizes for *t*-tests were estimated using the effsize package. Multilevel mediation analyses were computed using the mediation package. Multilevel models were built using the lme4 package. All error bars in the figures represent S.E.M.

RESULTS

Associative Memory Performance

Participants completed three separate sessions that tested memory for object-face associations (14 rounds on Day 1; 4 rounds before scanning on Day 2; 2 rounds after scanning on Day 2; Fig. 1a). Participants showed improved accuracy across test rounds in the Day 1 session, from a mean of 56.9% (SD = 12.8%) on round 1 to a mean of 95.5% (SD = 4.8%) on round 14 (main effect of test round: $F_{5.56,\ 155.73} = 91.29$, $p < 0.0001$, $\eta^2 = 0.55$). Accuracy did not vary by test round for either of the Day 2 sessions (Day 2 pre-scan: $F_{2.77,\ 77.63} = 1.63$, $p = 0.194$, $\eta^2 = 0.01$; Day 2 post-scan: $F_{1,\ 28} = 0.14$, $p = 0.713$, $\eta^2 = 0.0009$). Critically, accuracy was lower in the competitive condition than the non-competitive condition for each of the sessions (Day 1: $F_{1,\ 28} = 115.89$, $p < 0.0001$, $\eta^2 = 0.29$; Day 2 pre-scan: $F_{1,\ 28} = 21.8{,}1$ $p < 0.0001$, $\eta^2 = 0.15$; Day 2 post-scan: $F_{1,\ 28} = 22.25$, $p < 0.0001$, $\eta^2 = 0.20$; Fig. 2a). For subsequent analyses, we focused on associative memory performance from the Day 2 pre-scan session (an *a priori* decision; see Methods). Notably, for the Day 2 pre-scan session, lower accuracy in the competitive condition (M = 93.2%, SD = 6.9%) than the non-competitive condition (M = 98.9%, SD = 2.1%) was driven by an increased rate of selecting faces that were associated with the pairmate image (competitive condition: M = 6.0%, SD = 6.6%; non-competitive condition: M = 0.2%, SD = 0.6%; t_{28}= 4.74, $p < 0.0001$, 95% CI = [0.03 0.08], Cohen's d = 1.16, paired t-test; Fig. 2a). The rate of other errors did not differ in the competitive vs. non-competitive conditions (competitive: M = 0.8%, SD = 1.4%; non-competitive: M = 0.98%, SD = 1.6%; t_{28} = -0.18, p = 0.861, 95% CI = [-0.01 0.01], Cohen's d = -0.04, paired t-test). Thus, as intended, the competitive condition specifically increased interference between pairmate images.

Color Memory Bias

Immediately after the fMRI session, participants completed a color memory test. Color memory was indexed in two ways: (1) using a continuous, signed measure of distance, in degrees, between the reported and

actual color; positive values indicate a bias *away from* the competing memory and negative values indicate a bias *toward* the competing memory, and (2) using a categorical measure of the percentage of responses that were biased away from the competing memory (see Methods for details of each measure). We refer to these two measures as the *signed distance* and *percentage of away responses*, respectively.

For the competitive condition, mean signed distance was significantly greater than 0 (5.09 ± 4.69, mean ± SD; t_{28} = 5.84, p = 0.000003, 95% CI = [3.30 6.87], Cohen's d = 1.08, one-sample t-test; Fig. 2b), indicating that participants' color memory was systematically biased away from the color of the pairmate. In contrast, for the non-competitive condition—where the only difference was that pairmates were not from the same object category—signed distance did not differ from 0 (-0.39 ± 7.08; t_{28} = -0.29, p = 0.771, 95% CI = [-3.08 2.31], Cohen's d = -0.05, one-sample t-test). Signed distance was significantly greater (i.e., a stronger bias away from the pairmate) in the competitive condition compared to the non-competitive condition (t_{28} = 2.90, p = 0.007, 95% CI = [1.61 9.34], Cohen's d = 0.92, paired t-test). These data clearly demonstrate that similarity between images triggered the color memory bias.

The pattern of data was identical when considering the percentage of away responses. Namely, the percentage of away responses was significantly greater than 50% for the competitive condition (61.4 ± 3.6%; t_{28} = 4.49, p = 0.0001, 95% CI = [56.2% 66.6%], Cohen's d = 0.83, one-sample t-test; Fig. 2c), but not for the non-competitive condition (46.5 ± 14%; t_{28} = -1.35, p = 0.189, 95% CI = [41.2% 51.8%], Cohen's d = -0.25, one-sample t-test). The difference between the two conditions was also significant (t_{28} = 3.58, p = 0.001, 95% CI = [0.06 0.23], Cohen's d = 1.08, paired t-test). While the percentage of away responses does not contain information about the magnitude of the bias in color memory, it rules out the possibility that the effects observed with the signed distance measure were driven by a minority of trials with very high bias.

Relationship between associative memory and color memory bias

A key component of our theoretical framework is that exaggerating the color distance (in memory) between similar objects plays an adaptive role in reducing memory interference. To test this idea, we correlated each

participant's associative memory performance (from the Day 2 pre-scan session) with their color memory performance. For the competitive condition, mean associative memory performance was positively correlated with mean signed distance (r = 0.50, t_{26} = 2.91, p = 0.007, 95% CI = [0.15 0.73], one outlier excluded for associative memory performance < 3 SD below mean; Fig. 2d), consistent with the idea that stronger color memory repulsion (i.e., a bias in color memory away from the pairmate) supports lower associative memory interference. For the non-competitive condition, this correlation was not significant (r = -0.31 t_{26} = -1.63, p = 0.114, 95% CI = [-0.61 0.08], one outlier excluded for signed distance > 3 SD above the mean). Thus, a bias in color memory away from the pairmate was not beneficial if the pairmate was not similar to (competitive with) the target. An identical pattern of data was observed when considering the percentage of away responses as an index of color memory. Namely, for the competitive condition there was a positive correlation between associative memory performance and the mean percentage of away responses (r = 0.42, t_{26} = 2.39, p = 0.025, 95% CI = [0.06 0.69], one outlier excluded for associative memory performance < 3 SD below mean) and no significant correlation for the non-competitive condition (r = -0.37, t_{27} = -2.05, p = 0.050, 95% CI = [-0.65 -0.002]).

Neural representation of color information during recall

The key design feature of the competitive condition was that color information was critical for discriminating between pairmates. Specifically, in the competitive condition the only difference between pairmates was a 24-degree color difference. This contrasts with the non-competitive condition where pairmates differed in color (again 24 degrees) *and* object category. Because color information was therefore more important in the competitive condition, we predicted that representation of color information during the scanned recall trials would be relatively stronger in the competitive condition than the non-competitive condition. Notably, participants' only instruction on the recall trials was to bring each stimulus to mind as vividly as possible (mean percentage of vivid responses = 95.42%, *SD* = 5.43%). Participants were not explicitly oriented to color information nor had participants' memory for color been tested in any way to that point in the experiment.

To test for representation of color information, we computed the mean correlation of activity patterns evoked during recall of non-pairmate stimuli that shared an identical color value (e.g., red bean bag and red jacket; 'same-color' comparison, see Fig. 1b) and subtracted from this value the mean correlation between non-pairmate stimuli that were 24 degrees apart in color space (e.g., red bean bag and brown jacket; 'baseline' comparison, see Fig. 1b). Thus, the difference between these two measures (same-color – baseline) provided an index of color information. We then compared this index across the competitive and non-competitive trials. Critically, in terms of physical properties of the stimuli, the comparison between the competitive and non-competitive trials was perfectly matched: there was no objectively greater similarity between the *stimuli included in this analysis* in the competitive condition compared to the non-competitive condition—there was only a difference in the importance of the information.

For this and subsequent fMRI analyses we used a set of visual and parietal regions of interest (ROIs) previously described in Favila et al. (2018) (see Methods; Fig. 3a). Critically, these ROIs were previously shown to contain color and object feature representations during a memory recall task very similar to the current study. The set of ROIs included three visual ROIs (V1, LO, VTC) and five lateral parietal ROIs (pIPS, dLatIPS, vLatIPS, AnG, vIPS).

An ANOVA with factors of condition (competitive, non-competitive) and ROI (all eight ROIs) revealed a significant main effect of condition, with relatively stronger color information in the competitive condition than the non-competitive condition ($F_{1, 28} = 5.03$, $p = 0.033$, $\eta^2 = 0.04$). Neither the main effect of ROI nor the condition x ROI interaction were significant (ROI: $F_{4.55, 127.36} = 0.12$, $p = 0.984$, $\eta^2 < 0.001$; condition x ROI: $F_{4.10, 114.92} = 0.78$, $p = 0.542$, $\eta^2 = 0.008$). Considering individual ROIs, only LO and vIPS exhibited significantly stronger color representation in the competitive than non-competitive condition (LO: $t_{28} = 2.27$, $p = 0.031$, 95% CI = [0.002 0.03], Cohen's $d = 0.69$; vIPS: $t_{28} = 2.67$, $p = 0.012$, 95% CI = [0.004 0.03], Cohen's $d = 0.63$; paired *t*-tests, uncorrected; Fig 3b). Thus, as predicted, the greater relevance of color information in the competitive condition resulted in stronger representation of color information during recall, despite the fact that participants had not been explicitly oriented to color information in any way by this point of the experiment (the critical behavioral test of color memory occurred after fMRI scanning).

Post-hoc analyses of medial temporal and hippocampal ROIs (see Methods) did not reveal stronger color representation in the competitive than non-competitive condition for any of the ROIs ($|t|$'s < 1.66, p's > 0.109).

Neural similarity between pairmates during recall

We next tested whether neural similarity between pairmate stimuli was greater in the competitive than non-competitive condition. In terms of physical stimulus properties, pairmates were, of course, more similar in the competitive condition (e.g., two bean bags 24 degrees apart in color space) than in the non-competitive condition (e.g., a pillow and a ball 24 degrees apart in color space). Thus, based on stimulus properties alone, fMRI pattern similarity between pairmates should be greater in the competitive condition than the non-competitive condition. To measure pairmate similarity we computed the mean correlation between pairmate stimuli ('pairmate' comparison, see Fig. 1b) and subtracted from this value the mean correlation between non-pairmate stimuli that were also 24 degrees apart in color space ('baseline' comparison, see Fig. 1b). The difference between these two values (pairmate – baseline) yielded an index of pairmate similarity which was then compared across the competitive and non-competitive conditions.

Although pairmate similarity was numerically greater in the competitive than non-competitive condition across each of the eight ROIs, an ANOVA with factors of ROI and condition did not reveal a significant main effect of condition ($F_{1, 28}$ = 2.30, p = 0.140, η^2 = 0.016). The main effect of ROI and the condition x ROI interaction were also not significant (ROI: $F_{4.57, 127.90}$ = 0.68, p = 0.626, η^2 = 0.006; condition x ROI: $F_{3.82, 106.85}$ = 0.58, p = 0.670, η^2 = 0.006). However, there was a significant effect of condition, corrected for multiple comparisons (Bonferroni corrected), in vIPS, with greater pattern similarity in the competitive than non-competitive conditions (t_{28} = 3.12, p = 0.004, 95% CI = [0.005 0.02], Cohen's d = 0.70, paired t-test; Fig. 3c). Notably, as described above (Fig. 3b), vIPS also exhibited significantly stronger color representation in the competitive than non-competitive condition. Moreover, vIPS also exhibited significant object and color representations during a recall task in a prior study (Favila et al., 2018). Thus, across two

534 independent studies, we have consistently observed feature representations in this ROI during memory
535 recall.
536 Post-hoc analyses of medial temporal and hippocampal ROIs (see Methods) did not reveal greater
537 pairmate similarity in the competitive than non-competitive condition for any of the ROIs (|*t*|'s < 1.42, *p*'s >
538 0.168).
539
540 ***Neural measures of pairmate similarity predict color memory bias***
541
542 Results from the preceding analysis revealed greater similarity in vIPS representations of pairmates in the
543 competitive condition than the non-competitive condition. While this measure of neural similarity reflects
544 the greater physical similarity between pairmates in the competitive condition than the non-competitive
545 condition, the key finding from our behavioral results is that there is an adaptive benefit to *reducing* similarity
546 (in memory) between pairmates in the competitive condition. This raises the question of whether similarity
547 between vIPS representations of pairmates during competitive recall predicted the degree to which there
548 was repulsion of color memories (as measured in the post-scan color memory test). To test this, for each
549 condition (competitive, non-competitive) we correlated fMRI measures of pairmate *dissimilarity* (1 – pattern
550 similarity) with behavioral measures of mean signed color memory distance. This analysis was performed
551 within participant (i.e., at the level of individual pairmates). Given that each condition only corresponded to
552 6 pairmates per participant, Spearman rank correlation was used in order to reduce the influence of any
553 one data point. Correlation coefficients were then *z*-transformed, yielding a single *z*-transformed value for
554 each condition and participant.
555 For the competitive condition, the mean correlation between pairmate dissimilarity in vIPS during
556 recall and mean signed color memory distance was significantly positive (vIPS: $t_{28} = 3.75$, $p = 0.0008$, 95%
557 CI = [0.34 1.14], Cohen's $d = 0.70$, one-sample *t*-test; Fig. 4a). In other words, the more dissimilar vIPS
558 activity patterns were when recalling pairmates, the greater the color memory repulsion effect for those
559 pairmates. There was no correlation between pairmate dissimilarity in vIPS and signed color memory
560 distance for the non-competitive condition ($t_{28} = 0.78$, $p = 0.443$, 95% CI = [-0.22 0.49], Cohen's $d = 0.14$;

Fig. 4a) and the difference between the competitive and non-competitive conditions was significant (t_{28} = 2.39, p = 0.024, 95% CI = [0.09 1.12], Cohen's d = 0.61, paired t-test). Significant positive relationships were also observed when pairmate dissimilarity was measured from pIPS, VTC, and vLatIPS—again, only for the competitive condition (see table 1).

As a complementary analysis—and to better visualize the results in vIPS—we binned pairmates, for each participant, based on vIPS dissimilarity (competitive condition only). We generated three bins per participant: low, medium, and high pairmate dissimilarity. We then computed the mean signed color memory distance for each of these bins. A one-way ANOVA revealed a significant main effect of pairmate dissimilarity in vIPS on mean signed color memory distance (Fig. 4b; $F_{1.75, 48.90}$ = 4.95, p = 0.014, η^2 = 0.062), with greater dissimilarity between vIPS representations associated with greater distance in remembered color values (i.e., greater repulsion). We also computed mean accuracy on the associative memory test for these same vIPS dissimilarity bins in order to more directly test whether vIPS dissimilarity was associated with lower interference. Indeed, we again found a significant main effect of bin ($F_{1.78, 49.87}$ = 4.52, p = 0.019, η^2 = 0.068), with behavioral accuracy increasing as a function of pairmate dissimilarity in vIPS. Finally, a mediation analysis performed at the level of individual pairmates (see Methods) revealed that the relationship between vIPS dissimilarity and associative memory accuracy was significantly mediated by signed color memory distance (β = 0.12, CI = [0.02 0.23], p = 0.016, 1000 bootstrapped samples), consistent with the interpretation that vIPS dissimilarity reflected the degree of color memory repulsion, which in turn was associated with better associative memory accuracy (lower interference).

DISCUSSION

Here, we show that competition between similar memories triggers biases in their neural representations and corresponding behavioral expressions. Specifically, we demonstrate that subtle, diagnostic differences between events were exaggerated in long-term memory and that this exaggeration reduced interference. Critically, these behavioral expressions of memory distortion were predicted by adaptive, feature-specific changes to memory representations in parietal cortex.

Our behavioral paradigm was designed to isolate the effect that competition had on color memory. Specifically, the competitive and non-competitive conditions had perfectly matched structures, with equivalent color distances between pairmates in both conditions (Fig. 1b). The only difference was that pairmates in the competitive condition were from the same object category. As intended, this increased the number of interference-related errors, particularly during early stages of learning (Fig. 2a). The increase in interference-related errors is consistent with a long history of behavioral studies of memory interference (Anderson & Spellman, 1995; Mensink & Raaijmakers, 1988; Wixted, 2004). Our critical question, however, was whether competition distorted memory for object *features* that were otherwise successfully remembered. Results from the color memory post-test revealed a robust bias in color memory in the competitive condition—that is, participants exaggerated the distance between pairmates—but no systematic bias in the non-competitive condition. We refer to the bias in the competitive condition as a repulsion effect in order to emphasize that the bias was triggered by the representational proximity of competing memories (Bae & Luck, 2017; Chanales et al., 2017, in-press; Golomb, 2015), just as spatial proximity of like-poled magnets triggers magnetic repulsion.

It is important to emphasize that the repulsion effect is distinct from—in fact, opposite to—an interference effect. That is, interference-related errors should lead participants to occasionally recall the color of the competing object—an error that would produce a bias in color memory *toward* the pairmate (Fig. 1c, d). Here, we did not test color memory until the very end of the experiment, so as to avoid explicitly orienting participants to color information prior to (or during) the fMRI session, but our speculation is that the repulsion effect only emerged after extensive practice and as interference errors subsided (Chanales et al., in-press). In this sense, the repulsion effect can be thought of as an aftereffect of initial memory interference. Although repulsion reflects a form of memory error, our findings indicate that it is an adaptive error: participants who exhibited a stronger repulsion effect also exhibited fewer interference-related errors (Fig. 2d). To the extent that objective similarity between stimuli is a root cause of memory interference (Osgood, 1949), then exaggerating the difference between stimuli in memory is a potentially powerful means for reducing interference (Chanales et al., in-press; Favila et al., 2016; Hulbert & Norman, 2015).

Our fMRI analyses, which measured neural activity patterns as participants recalled object images, provided a unique means for covertly probing the qualities of participants' memories. These analyses revealed two forms of adaptive memory representations in parietal cortex. First, despite the fact that participants were not instructed to think about or report objects' colors during these recall trials, we observed stronger color information—across the full set of visual and parietal ROIs, and in vIPS specifically—during competitive than non-competitive recall trials. The stronger representation of color information during competitive trials can be viewed as an adaptive response to competition in that color information was the only (or diagnostic) feature dimension for discriminating pairmates in the competitive condition.

Second, although pairmate similarity in vIPS was stronger during competitive than non-competitive recall trials (indicating that vIPS was sensitive to object similarity; Fig. 3c), we found that greater *dissimilarity* between vIPS pairmate representations during competitive recall trials was associated with greater color memory repulsion and less memory interference. In other words, minimizing the overlap of neural representations of pairmates was an adaptive response to competition. This relationship was observed within participants, at the level of individual pairmates, but it is important to emphasize that these measures were temporally offset: vIPS pattern similarity was measured during recall trials in the scanner (with the only instruction being to recall objects as vividly as possible) whereas behavioral expressions of color memory were only tested after scanning was completed. This again makes the point that color information—in this case the subtle *difference* in pairmate colors—was a salient component of activity patterns in vIPS during competitive recall.

Importantly, when our two main fMRI findings are taken together, they indicate that an adaptive response to competition involved an *increase* in similarity between stimuli that shared a diagnostic feature value (i.e., objects of the same color) but a *decrease* in similarity between stimuli that had subtly different values for a diagnostic feature (i.e., pairmates, which had slightly different colors). This indicates that avoiding memory interference does not necessarily require a global reduction in similarity to all other memories (LaRocque et al., 2013), but instead may be accomplished by more *targeted changes* in representational structure that emphasize relevant similarities as well as important differences between events that are stored in memory. Critically, this idea is distinct from—if not fundamentally incompatible

with—the traditional, and dominant view that interference is avoided through the orthogonalization of memory representations (Colgin et al., 2008; Yassa & Stark, 2011). Specifically, whereas orthogonalization emphasizes an initial encoding of new memories as independent from existing memories, our findings instead emphasize that the representation of a given memory is *highly dependent* on representations of other memories (Hulbert & Norman, 2015).

Our fMRI findings also add to a growing body of evidence that implicates parietal cortex in actively representing content during memory retrieval (Kuhl & Chun, 2014; Lee & Kuhl, 2016; Lee et al., 2019; Rugg & King, 2018; Sestieri, Shulman, & Corbetta, 2017). Of most direct relevance, in a recent study we found that vIPS (a ventral subregion of parietal cortex) actively represents color and object category information during memory recall (Favila et al., 2018). However, this prior study focused on decoding the *objective properties* of recalled stimuli and did not test whether competition influenced or distorted these representations, nor did it establish a link between vIPS representations and behavioral expressions of memory. The current findings provide unique evidence that representations within this same vIPS subregion reflect subtle distortions in how events are remembered that are *dissociable from the objective properties of the event*. More generally, our findings highlight the behavioral relevance and detailed nature of memory representations in parietal cortex.

While our findings provide strong evidence that representations in parietal cortex reflect the influence that competition had on memory representations, it is not necessarily the case that parietal cortex was the source of this influence. Rather, competition between memories is thought to induce targeted plasticity in the hippocampus (Norman, Newman, & Detre, 2007; Ritvo, Turk-Browne, & Norman, 2019). In fact, hippocampal representations have been shown to specifically exaggerate differences between highly similar stimuli (Ballard et al., 2019; Chanales et al., 2017; Dimsdale-Zucker et al., 2018; Favila et al., 2016; Hulbert & Norman, 2015; Schapiro et al., 2012; Schlichting et al., 2015). However, these exaggerations in hippocampal activity patterns have generally been observed during memory encoding or perception (Ballard et al., 2019; Chanales et al., 2017; Dimsdale-Zucker et al., 2018; Favila et al., 2016; Hulbert & Norman, 2015; Schapiro et al., 2012; Schlichting et al., 2015), as opposed to memory recall, and they have

not been translated to explicit feature spaces. Indeed, attempts to translate hippocampal activity patterns to explicit feature dimensions or categories have tended to be unsuccessful (LaRocque et al., 2013; Liang et al., 2013). In post hoc analyses, we did not find any evidence that competition influenced feature representations in the hippocampus or medial temporal lobe ROIs. That said, one notable aspect of our study is that each object was retrieved from memory many times before fMRI scanning began. Given that repeated retrieval has specifically been shown to hasten the transfer of representations to parietal cortex (Brodt et al., 2018, 2016), this raises the question of whether the observed findings in parietal cortex were dependent on repeated retrieval. For example, it is possible that competition induces exaggerated representations that are initially expressed in the hippocampus but ultimately transformed, via retrieval, into stable representations in parietal cortex (Favila, Lee, & Kuhl, 2020). While the current study cannot address this question, it represents an interesting avenue for future research.

In summary, our findings provide unique evidence that memory-based representations in parietal cortex exhibit adaptive, feature-specific changes in response to competition and that these changes in parietal representations predict distortions in behavioral expressions of memory. More generally, our findings provide unique evidence in support of the perspective that memory distortions are an adaptive component of the memory system (Schacter, Guerin, & St. Jacques, 2011).

REFERENCES

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*. US: American Psychological Association. https://doi.org/10.1037/0033-295X.102.1.68

Bae, G. Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, and Psychophysics*, *79*(8), 2376–2395. https://doi.org/10.3758/s13414-017-1404-8

Ballard, I. C., Wagner, A. D., & McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications*, *10*(1), 1073. https://doi.org/10.1038/s41467-019-08998-1

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*(2), 97–105. https://doi.org/10.1037/h0047507

Bone, M. B., Ahmad, F., & Buchsbaum, B. R. (2020). Feature-specific neural reactivation during episodic memory. *Nature Communications*, *11*(1), 1945. https://doi.org/10.1038/s41467-020-15763-2

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791–808. https://doi.org/10.1037/a0029649

Brodt, S., Gais, S., Beck, J., Erb, M., Scheffler, K., & Schönauer, M. (2018). Fast track to the neocortex: A memory engram in the posterior parietal cortex. *Science*, *362*(6418), 1045 LP – 1048. https://doi.org/10.1126/science.aau2528

Brodt, S., Pöhlchen, D., Flanagin, V. L., Glasauer, S., Gais, S., & Schönauer, M. (2016). Rapid and independent memory formation in the parietal cortex. *Proceedings of the National Academy of Sciences*, *113*(46), 13251 LP – 13256. https://doi.org/10.1073/pnas.1605719113

Chanales, A. J. H., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Current Biology*, *27*(15), 2307-2317.e5. https://doi.org/10.1016/j.cub.2017.06.057

Chanales, A. J. H., Tremblay-McGaw, A. G., & Kuhl, B. A. (in-press). Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological Science*.

https://doi.org/10.1101/2020.01.14.900381

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1). https://doi.org/10.1038/nn.4450

Colgin, L. L., Moser, E. I., & Moser, M.-B. (2008). Understanding memory through hippocampal remapping. *Trends in Neurosciences*, *31*(9), 469–477. https://doi.org/https://doi.org/10.1016/j.tins.2008.06.008

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4-5), 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/https://doi.org/10.1006/nimg.1998.0395

Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P., & Ranganath, C. (2018). CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02752-1

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*(1), 11066. https://doi.org/10.1038/ncomms11066

Favila, S. E., Lee, H., & Kuhl, B. A. (2020). Transforming the Concept of Memory Reactivation. *Trends in Neurosciences*. https://doi.org/https://doi.org/10.1016/j.tins.2020.09.006

Favila, S. E., Samide, R., Sweigart, S. C., & Kuhl, B. A. (2018). Parietal representations of stimulus features are amplified during memory retrieval and flexibly aligned with top-down goals. *Journal of Neuroscience*, *38*(36), 0564–18. https://doi.org/10.1523/JNEUROSCI.0564-18.2018

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/https://doi.org/10.1016/S1053-8119(09)70884-5

Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception, & Psychophysics*, *77*(8), 2562–2569. https://doi.org/10.3758/s13414-015-0951-0

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python . *Frontiers in Neuroinformatics* . Retrieved from https://www.frontiersin.org/article/10.3389/fninf.2011.00013

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.06.060

Hulbert, J. C., & Norman, K. A. (2015). Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cerebral Cortex*, *25*(10), 3994–4008. https://doi.org/10.1093/cercor/bhu284

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/https://doi.org/10.1006/nimg.2002.1132

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural differentiation of incorrectly predicted memories. *Journal of Neuroscience*, *37*(8), 2022–2031.

Kuhl, B. A., & Chun, M. M. (2014). Successful Remembering Elicits Event-Specific Activity Patterns in Lateral Parietal Cortex. *Journal of Neuroscience*, *34*(23), 8051–8060. https://doi.org/10.1523/jneurosci.4328-13.2014

Kuhl, B. A., Johnson, M. K., & Chun, M. M. (2013). Dissociable Neural Mechanisms for Goal-Directed Versus Incidental Memory Reactivation. *Journal of Neuroscience*, *33*(41), 16099–16109. https://doi.org/10.1523/jneurosci.0207-13.2013

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global

Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *Journal of Neuroscience*, *33*(13), 5466–5474. https://doi.org/10.1523/JNEUROSCI.4293-12.2013

Lee, H., & Kuhl, B. A. (2016). Reconstructing Perceived and Retrieved Faces from Activity Patterns in Lateral Parietal Cortex. *Journal of Neuroscience*, *36*(22), 6069–6082. https://doi.org/10.1523/JNEUROSCI.4286-15.2016

Lee, H., Samide, R., Richter, F. R., & Kuhl, B. A. (2019). Decomposing Parietal Memory Reactivation to Predict Consequences of Remembering. *Cerebral Cortex*, *29*(8), 3305–3318. https://doi.org/10.1093/cercor/bhy200

Liang, J. C., Wagner, A. D., & Preston, A. R. (2013). Content Representation in the Human Medial Temporal Lobe. *Cerebral Cortex*, *23*(1), 80–96. https://doi.org/10.1093/cercor/bhr379

Long, N. M., Lee, H., & Kuhl, B. A. (2016). Hippocampal Mismatch Signals Are Modulated by the Strength of Neural Predictions and Their Similarity to Outcomes. *Journal of Neuroscience*, *36*(50), 12677–12687.

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/https://doi.org/10.1016/j.neuroimage.2011.08.076

Norman, K. A., Newman, E. L., & Detre, G. (2007). A Neural Network Model of Retrieval-Induced Forgetting. *Psychological Review*, *114*(4), 887–953. https://doi.org/10.1037/0033-295X.114.4.887

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

Osgood, C. E. (1949). The similarity paradox in human learning: a resolution. *Psychological Review*. US: American Psychological Association. https://doi.org/10.1037/h0057488

Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends in Cognitive Sciences*, *23*(9), 726–742. https://doi.org/10.1016/j.tics.2019.06.007

Rugg, M. D., & King, D. R. (2018). Ventral lateral parietal cortex and episodic memory retrieval. *Cortex*, *107*, 238–250. https://doi.org/https://doi.org/10.1016/j.cortex.2017.07.012

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467–474. https://doi.org/10.1016/j.tics.2011.08.004

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*(1), 8151. https://doi.org/10.1038/ncomms9151

Sestieri, C., Shulman, G. L., & Corbetta, M. (2017). The contribution of the human posterior parietal cortex to episodic memory. *Nature Reviews Neuroscience*, *18*(3), 183–192. https://doi.org/10.1038/nrn.2017.6

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374–391. https://doi.org/10.1002/hipo.450040319

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2014). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wixted, J. T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, *55*(1), 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Xiao, X., Dong, Q., Gao, J., Men, W., Poldrack, R. A., & Xue, G. (2017). Transformed Neural Pattern Reinstatement during Episodic Memory Retrieval. *The Journal of Neuroscience*, *37*(11), 2986–2998. https://doi.org/10.1523/JNEUROSCI.2324-16.2017

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/https://doi.org/10.1016/j.tins.2011.06.006

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., … Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*(1), 258–287. https://doi.org/10.1002/hbm.22627

Table 1. Summary of key statistical analyses. Color representation analyses refer to paired-samples *t*-tests comparing color similarity effects (see Methods) for the competitive vs. non-competitive conditions. Pairmate similarity analyses refer to paired-samples *t*-tests comparing pairmate similarity effects (see Methods) for the competitive vs. non-competitive conditions. The relation to mean signed distance refers to one-sample *t*-tests comparing *z*-transformed correlations between fMRI pairmate dissimilarity and mean signed color memory distance to a test statistic of 0 (no relationship). Results from individual visual and parietal ROIs are presented in separate rows. Note: * $p < .05$, uncorrected; ** $p < .05$, Bonferroni corrected; *** $p < .01$, Bonferroni corrected.

ROI	Color representation		Pairmate similarity		Relation to mean signed distance			
					Competitive		Non-competitive	
	t_{28}	*p*	t_{28}	*p*	t_{28}	*p*	t_{28}	*p*
V1	1.22	0.232	0.89	0.382	0.82	0.417	-0.34	0.734
LO	**2.27**	**0.031***	1.71	0.098	1.34	0.190	-0.75	0.458
VTC	1.16	0.257	0.45	0.653	**2.13**	**0.042***	0.59	0.558
pIPS	1.85	0.075	0.84	0.409	**3.08**	**0.005****	1.08	0.289
dLatIPS	1.68	0.104	0.73	0.472	1.50	0.145	0.65	0.520
vLatIPS	1.69	0.101	0.52	0.609	**2.92**	**0.007****	-1.89	0.069
AnG	0.57	0.573	0.36	0.720	0.75	0.462	-0.72	0.475
vIPS	**2.67**	**0.012***	**3.12**	**0.004****	**3.75**	**0.0008*****	0.78	0.443

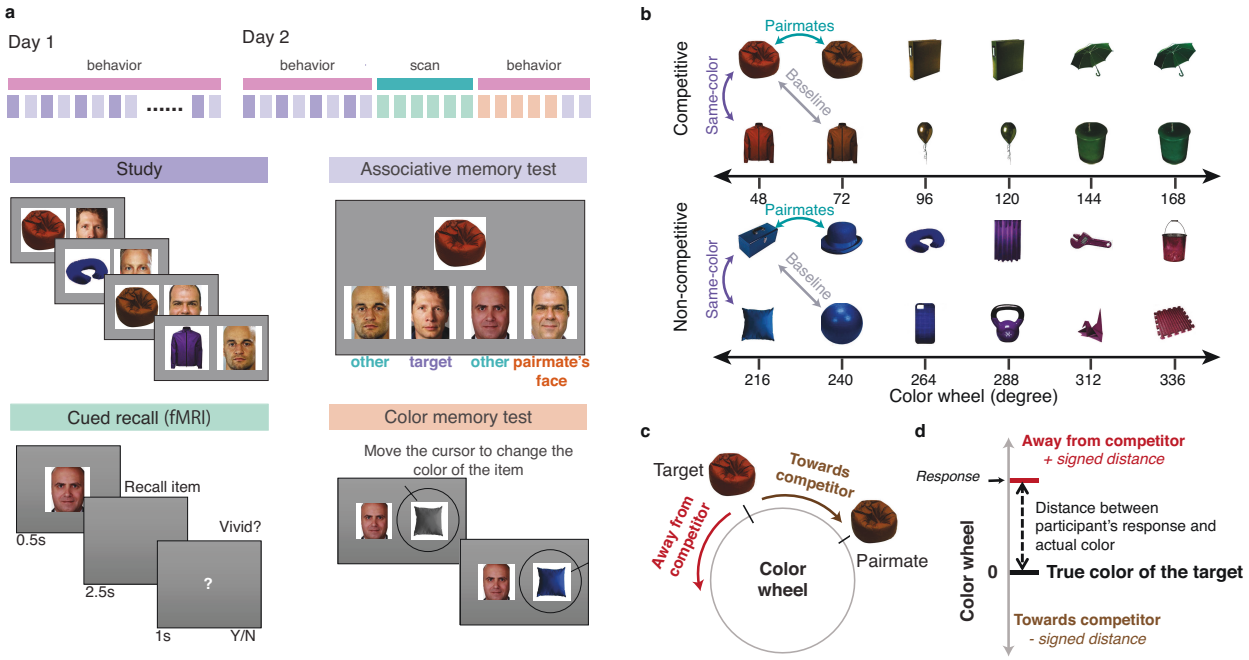

Fig. 1. Experimental Design and Procedure. **a**, Overview of paradigm. On Day 1, participants completed 14 Study and Associative Memory Test rounds. During Study, participants were shown object-face pairs and during Associative Memory Test, participants were shown an object and selected the corresponding face from a set of four choices. The set of four choices included the target face along with the face associated with the object's pairmate. On Day 2, participants completed four additional Study and Associative Memory Test rounds before entering the fMRI scanner. During scanning, participants completed a Cued Recall task during which face images were shown and participants recalled the corresponding image and indicated, by button press, the vividness of their recall. After exiting the scanner, participants completed a Color Memory Test during which a face image was shown alongside a greyscale version of the corresponding object. Participants used a continuous color wheel to indicate their memory for the object's color. Finally participants completed 2 more Associative Memory Test rounds. **b**, Sample structure of object stimuli. For both the competitive and non-competitive conditions, pairmate stimuli were 24 degrees apart in color space. For the competitive condition, pairmates were from the same object category; for the non-competitive condition, pairmates were from distinct categories. For both conditions, some objects had identical colors (Same-color). fMRI pattern similarity for Pairmate and Same-color comparisons were compared against a Baseline comparison of stimuli that were from different object categories and 24 degrees apart in color space. **c**,**d**, Responses on the color memory test were used to categorize memory for each object's color as being biased toward or away from the color of the competing object (**c**) and to measure the signed distance, in degrees, between participants' responses and the true color of the target (**d**).

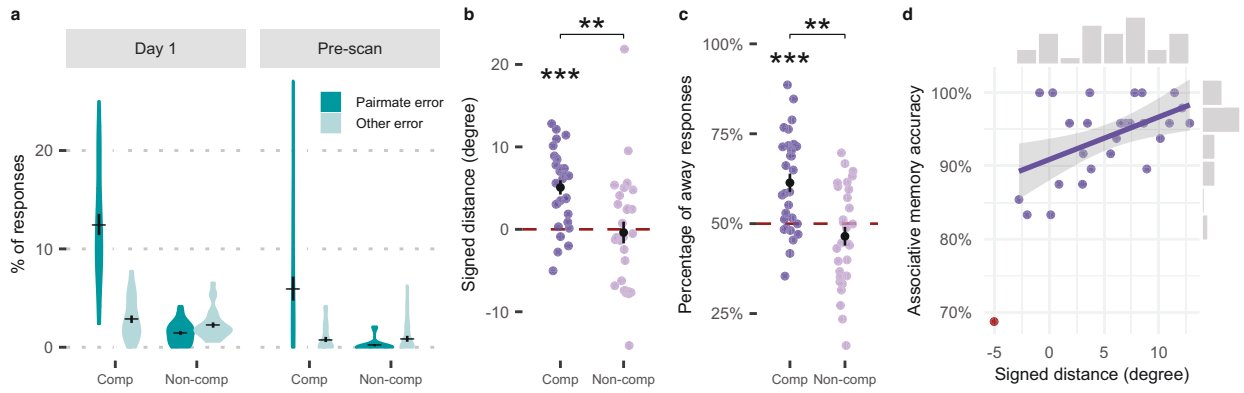

Fig. 2. Behavioral results. a, Associative memory performance across the experiment. The overall error rate (pairmate error + other error) was higher in the competitive condition than the non-competitive condition for each of the associative memory test sessions (Day 1, Day 2 pre-scan, Day 2 post-scan (not shown); all *p*'s < 0.0001). Subsequent analyses focused on associative memory performance from the Day 2 pre-scan session. For the Day 2 pre-scan session, participants were significantly more likely to select faces that were associated with the pairmate image (pairmate error) in the competitive condition (M = 6.0%, SD = 6.6%) compared to the non-competitive condition (M = 0.2%, SD = 0.6%; $p < 0.0001$), confirming that similarity between pairmates was a source of interference. **b**, Signed distance of responses in the color memory test. For the competitive condition, mean signed distance was significantly greater than 0 (p = 0.000003), reflecting a bias away from the color of the pairmate object (repulsion). Signed distance did not differ from 0 in the non-competitive condition ($p = 0.771$). The difference between the competitive and non-competitive conditions was also significant ($p = 0.007$). **c**, Percentage of away responses in the color memory test. The percentage of color memory responses 'away from' the color of the pairmate object was significantly greater than 50% for the competitive condition ($p = 0.0001$), but not for the non-competitive condition ($p = 0.189$). The difference between the competitive and non-competitive conditions was also significant ($p = 0.001$). **d**, Relationship between associative memory accuracy and mean signed color memory distance. For the competitive condition, participants with greater mean signed color memory distance (greater repulsion) exhibited better associative memory accuracy [r = 0.50, p = 0.007, one outlier (red dot) excluded for associative memory performance < 3 SD below mean]. Notes: colored dots reflect data from individual participants. Error bars reflect +/- S.E.M.; *** $p < 0.001$, ** $p < 0.01$

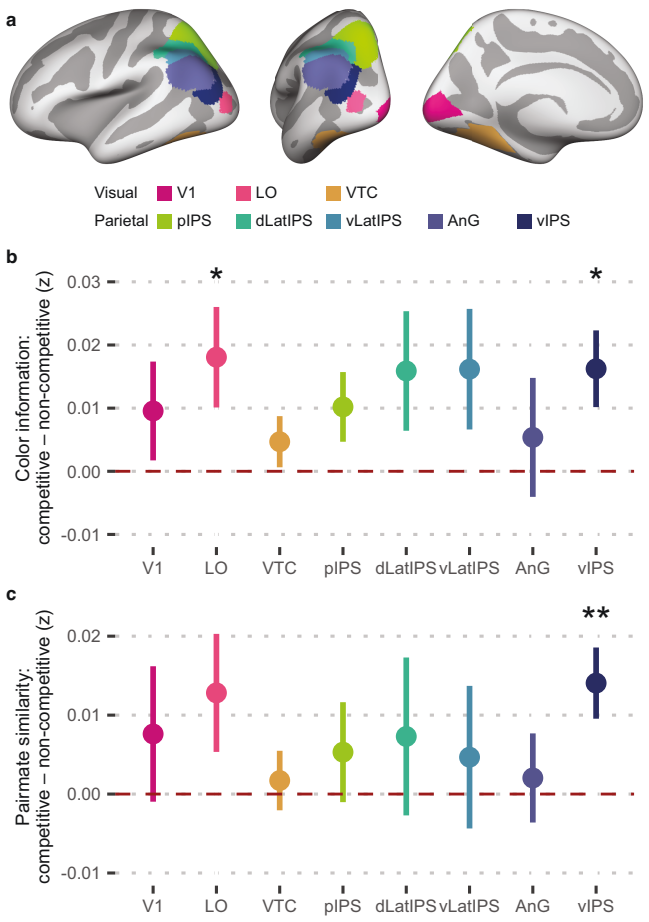

Fig. 3. Neural feature representations as a function of memory competition. a, Anatomical ROIs visualized on the Freesurfer average cortical surface. **b**, Color information as a function of memory competition. Color information was defined as the fMRI pattern similarity between pairs of same-color objects relative to pattern similarity between baseline pairs of objects (see **Fig. 1b**). Color information was significantly stronger in the competitive than non-competitive condition (i.e., values greater than 0) across the set of ROIs as a whole and in LO and vIPS individually (p's < .05). **c**, Pairmate similarity as a function of memory competition. Pairmate similarity was defined as the fMRI pattern similarity between pairmate objects relative to pattern similarity between baseline pairs of objects. Only vIPS showed significantly greater pairmate similarity in the competitive than non-competitive conditions ($p = 0.004$). Error bars reflect +/- S.E.M.; ** $p < 0.01$, * $p < 0.05$

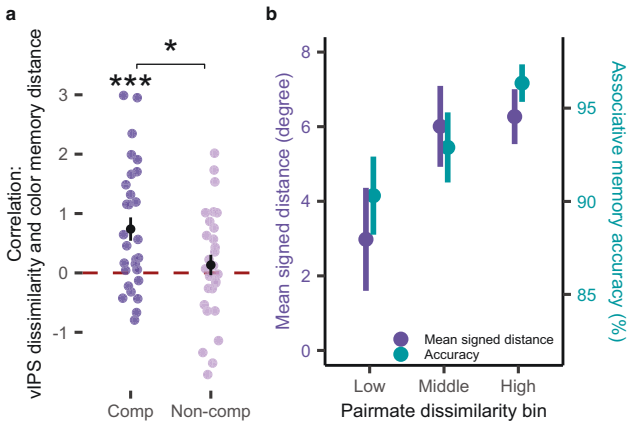

Fig. 4. Neural measures of pairmate (dis)similarity predict color memory bias in vIPS. a, Mean correlation between vIPS pairmate dissimilarity during recall and mean signed color memory distance. Correlations were performed within participant and correlation coefficients were *z*-transformed. For the competitive condition, the mean correlation was significantly positive ($p = 0.004$), indicating that greater pairmate dissimilarity in vIPS was associated with a stronger bias to remember pairmates' colors as *away from* each other. There was no correlation between vIPS pairmate dissimilarity and signed color memory distance for the non-competitive condition ($p = 0.566$). **b**, Relationship between vIPS pairmate dissimilarity (binned into low, medium, high groups) and mean signed color memory distance (purple) and associative memory accuracy (teal). Mean signed color memory distance and associative memory accuracy each significantly varied as a function of vIPS dissimilarity (p's < .05), with greater vIPS dissimilarity associated with greater mean signed color memory distance and higher associative memory accuracy. *** $p < 0.001$, * $p < 0.05$