Hippocampal transformations occur along dimensions of memory interference

Soroush Mirjalili[1,2,*], Guo Wanjia[1,2,3,5], Dominik Grätz[1,5], Eric Wang[1,2,4], Ulrich Mayr[1], Brice A. Kuhl[1,2,*]

[1] Department of Psychology, University of Oregon, 1451 Onyx St., Eugene, OR 97403, USA
[2] Institute of Neuroscience, University of Oregon, 1585 E 13th Ave., Eugene, OR 97403, USA
[3] Department of Psychology, Princeton University, South Drive, Princeton, NJ 08540, USA
[4] Department of Psychology, University of Southern California, Seeley G. Mudd Building, 3620 McClintock Ave, Los Angeles, CA 90089, USA
[5] These authors contributed equally: Guo Wanjia, Dominik Grätz

[*]Correspondence: soroush@uoregon.edu (S.M.), bkuhl@uoregon.edu (B.A.K.)

ABSTRACT

The role of the hippocampus in resolving memory interference has been greatly elucidated by considering the relationship between the similarity of visual stimuli (input) and corresponding similarity of hippocampal activity patterns (output). However, these input-output functions can take surprisingly different forms. Here, we reconcile seemingly conflicting findings by considering the possibility that the hippocampus prioritizes different dimensions of visual similarity across different stages of learning. First, we generated a set of natural scene images from two visual categories and rigorously characterized visual similarity using a wide array of methods: neural networks, artificial intelligence models, and human perceptual and memory decisions. We then identified two orthogonal dimensions of visual similarity that each predicted memory interference, but that did so at distinct stages of learning. Using high-resolution fMRI, we then tested for dimension-specific input-output functions within the hippocampus. Within CA3 and dentate gyrus (CA3/DG), we show that dimensions of visual similarity were inverted (negative functions) at stages of learning when they contributed to memory interference. When the dimensions did not contribute to interference, functions were positive. These findings reveal that hippocampal representations of visually-similar stimuli are highly dynamic and critically depend on the dimensions of similarity that currently contribute to memory interference.

INTRODUCTION

A critical facet of the hippocampus' role in episodic memory is its capacity to minimize interference[1–6]. Because memory interference is driven by similarity between stimuli[7–11], theories of interference resolution in the hippocampus have focused on how stimulus similarity (input similarity) translates to the similarity of corresponding hippocampal representations (output similarity). While a general principle of interference resolution is that output similarity should be lower than input similarity[1,5], this principle can be satisfied in qualitatively distinct ways. Indeed, in some contexts, increases in stimulus similarity yield positive, but sub-linear increases in hippocampal similarity[1,5,12]. In other contexts, increases in stimulus similarity yield no change in hippocampal similarity[13,14]. In yet other contexts, higher stimulus similarity has been associated with *lower* hippocampal similarity[15–17]. A major challenge for theories of the hippocampus is to explain this variability in the shape of input output functions.

One reason why relationships between stimulus similarity and hippocampal similarity may be so variable is because similarity, itself, can be defined using multiple metrics[18–23] that potentially capture multiple underlying *dimensions* of similarity[18–21,23–25]. This raises the obvious question: which dimensions 'matter' to the hippocampus? If a primary function of the hippocampus is to minimize interference, then a simple, but potentially clarifying perspective is that the hippocampus will target dimensions of similarity that drive interference. That said, an underappreciated point is that interference may be a *moving target*: the specific dimensions of similarity that drive interference may shift with learning[26]. While speculative, this may explain why relationships between stimulus similarity and hippocampal similarity sometimes change with experience[16,17,27,28]. However, formally testing this idea requires (a) identifying distinct dimensions of stimulus similarity, (b) characterizing how these dimensions influence memory interference across stages of learning, and (c) testing whether dimension-specific influences on behavior explain the shape of input-output functions in the hippocampus.

Here, we first developed a set of natural scene images from two visual categories and obtained 10 different metrics of visual similarity. These metrics of visual similarity were generated from new human behavioral data and various computational methods. From these similarity metrics, we extracted two orthogonal dimensions of similarity that strongly predicted memory interference errors. Strikingly, we found that the relative influence of these dimensions on behavior sharply changed with experience (learning). Whereas one dimension drove interference during early stages of learning, the other drove interference during later stages of learning. Finally, using high-resolution fMRI, we show that dimension-specific input-output functions in CA3 and dentate gyrus (CA3/DG) assumed qualitatively different shapes across stages of learning. Specifically, CA3/DG *inverted* each dimension of visual similarity (negative input-output functions) when it contributed to memory interference but preserved each dimension (positive functions) when it was not actively contributing to interference. Thus, entirely opposite coding schemes of similarity dimensions in the hippocampus can be explained by the extent to which these dimensions were behaviorally relevant. These findings provide important insight into how the hippocampus resolves memory interference, demonstrating highly dynamic and adaptive processes that dramatically increase the representational distance between memories that are most at risk for interference.

RESULTS

Dimensions of visual stimulus similarity

We first generated a set of 48 visual stimuli (natural scenes) drawn from two categories: beaches and gazebos (**Fig. 1a**). We then selected 10 different similarity metrics that yielded (for each metric) a matrix (**Fig. 1b**) describing the similarity of each pair of scenes within each category (276 beach pairs + 276 gazebo pairs = 552 pairs in total). Two of these metrics were based on human behavioral ratings that we collected in separate experiments. In one experiment (*behavioral memory experiment*), we measured associative memory confusability (**Fig. 1c**, left). In the other experiment (*perceptual similarity experiment*), we measured subjective judgments of perceptual similarity (**Fig. 1c**, right). The other 8 metrics were based on neural network models [three different layers of VGG-16, AlexNet, and Contrastive Language-Image Pretraining (CLIP)], Natural Language Processing (NLP) applied to AI-generated text descriptions of images (**Extended Data Fig. 1**), a measure of structural image features [Structural Similarity Index (SSIM)], and a model of global image features (GIST). (See *Methods* for details of each measure). Rather than identifying a single measure or model that 'best' describes stimulus similarity, our goal was to extract *dimensions* of similarity from the *set* of measures/models.

Given that the behavioral memory experiment and the perceptual similarity experiment were based on human responses, it was important to establish that resulting similarity metrics were reliable. To test this, we computed split-half reliability. Specifically, we divided the pool of participants in half, calculated similarity matrices within each half of the participants, and then correlated these similarity matrices. For both experiments, the mean split-half reliability (r) exceeded 0.69, with reliability asymptotically increasing as a function of the total number of participants included (**Fig. 1d**).

While the 10 similarity matrices were generally positively correlated with each other, Pearson correlations ranged from -0.154 to 0.665 (**Fig. 1e**). To identify different dimensions of similarity, each of the 10 similarity matrices was vectorized and combined in a new matrix that contained the pairwise similarity for all 552 within-category pairs across all 10 similarity metrics (**Fig. 1f**). We then applied a principal component analysis (PCA) to this new matrix to identify orthogonal components (dimensions) of similarity across the 10 metrics (**Fig. 1g**). The first 2 principal components (PC1 and PC2) explained over 60% of the variance in the similarity matrix (41.3% and 19.1%, respectively; **Fig. 1h**). For PC1, the metrics with the highest positive loadings were VGG max pooling layers 1 and 3, SSIM, and AlexNet. For PC2, the metrics with the highest positive loadings were memory confusability (behavioral memory experiment), perceptual judgments (perceptual similarity experiment), NLP of text descriptions, and GIST (**Fig. 1i**). The loadings for all 10 PCs are in **Extended Data Table 1**.

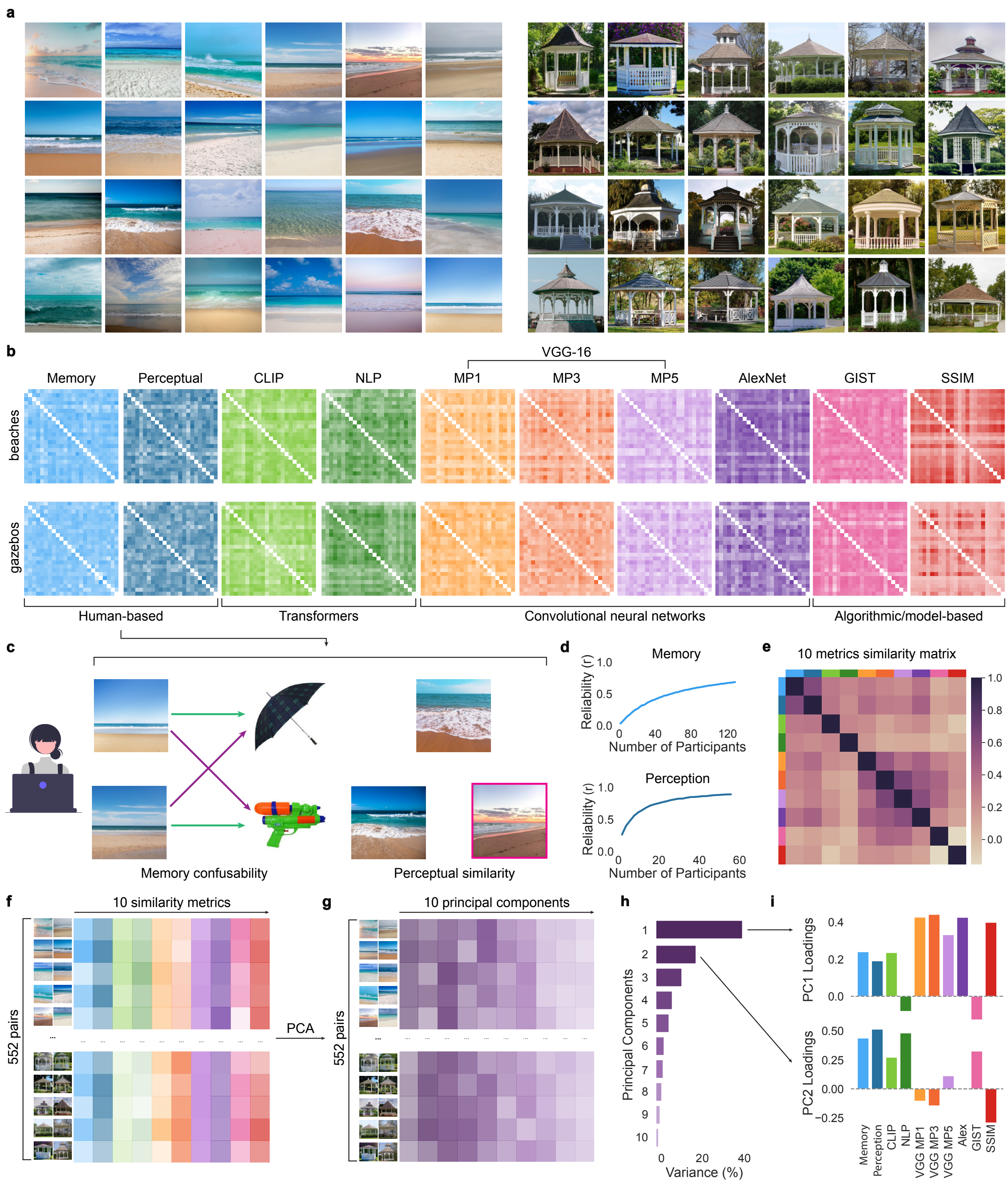


Fig. 1 | Identifying orthogonal dimensions of visual similarity. a, The stimulus set comprised 24 beaches and 24 gazebos (all shown here). **b**, 10 metrics of similarity were selected (see Methods). For each metric, a similarity matrix was generated for each visual category, with each cell representing the similarity between a pair of scenes. **c**, Left, the behavioral memory experiment involved a scene–object associative learning task that provided a measure of memory confusability. Green arrows reflect correct associations; purple arrows represent errors (interference) that were used to generate a confusion matrix. Right, the perceptual similarity experiment used an odd-one-out task, in which similarity between two scenes was defined by the probability that a third scene was chosen as 'odd'. **d**, Split-half reliability (collapsed across beaches and gazebos) as a function of number of participants included in the behavioral memory (top) and perceptual similarity (bottom) experiments. **e**, Matrix reflecting pairwise similarity (Pearson correlations) of the 10 similarity matrices. Higher values indicate greater alignment of the similarity matrices across metrics. **f**, All similarity data was combined into a single matrix with 552 row (each within-category pair of scenes) and 10 columns (each similarity metric). Each cell represents the similarity between a specific pair of scenes according to a specific metric. **g**, PCA was applied to the matrix shown in **f** to obtain orthogonal dimensions of similarity. Each cell represents the principal component (PC) score for a specific pair of scenes according to a specific PC (dimension). **h**, The variance explained by each PC. **i**, Contribution (loading) for each similarity metric for PC1 and PC2.

Establishing behaviorally relevant dimensions of similarity

We next sought to establish the dimensions of visual similarity that were related to behavioral expressions of memory interference. We addressed this by analyzing behavioral data from the *fMRI experiment*, which was entirely independent from the data reported above. The fMRI experiment was modeled after the behavioral memory experiment (**Fig. 1c**), using the same stimuli (24 beaches, 24 gazebos) and associative learning framework. The experiment included three phases (**Fig. 2a**): scene-object associative learning conducted prior to fMRI scanning (training phase), scene exposure during scanning (exposure phase), and an associative memory test conducted after scanning (post-test).

The training phase consisted of alternating study-test rounds (3 total). During study rounds, each scene was presented with a unique object. During test rounds, scenes were presented along with three objects and participants selected the associated object. Of the three objects, one was correct (target), one was associated with a scene from the same category as the current scene (competitor) and one was associated with a scene from the other category (non-competitor). Importantly, for each participant, one visual category was designated the 'high training' category, and the other was 'low training'. High training associations were presented more often, for longer durations, and with feedback during test rounds. During the exposure phase, participants were repeatedly shown all 48 scenes from the training phase, along with a handful of novel scenes. Participants were instructed to identify the novel scenes. In the post-test, each scene was presented along with all 24 objects that had been associated with that category (1 target, 23 competitors). Participants attempted to select the target.

The post-test data were used to identify dimensions of visual similarity that produced interference. In particular, the data were used to generate a pairwise memory confusability matrix for each visual category (**Fig. 2b**). To identify similarity dimensions that predicted memory interference errors, we ran mixed-effects models (a separate model for each PC) for which the dependent measure was the pairwise value from the post-test confusability matrix and the predictor included PC scores for one of the 10 dimensions of similarity (**Fig. 1g**). Significant effects of PC score (at Bonferroni corrected $\alpha = 0.005$) were observed for PC1 ($t_{(28699)} = 10.933$, $P < 0.001$), PC2 ($t_{(28699)} = 9.470$, $P < 0.001$), PC4 ($t_{(28699)} = -3.478$, $P < 0.001$), and PC6 ($t_{(28699)} = 2.823$, $P = 0.005$) (**Fig. 2c**). Given that PC1 and PC2 explained over 60% of the variance in the combined similarity matrix (**Fig. 1h**) and they each strongly predicted behavioral interference errors, subsequent analyses focused on these two components. For visualization, **Fig. 2d** shows the similarity structure of all scenes within each category, separately for PC1 and PC2. In this visualization, connections between pairs of images are equivalent to edges in a graph and individual images are equivalent to nodes. Pairwise similarity between images can therefore be described in terms of edge strength whereas an individual stimulus' aggregate similarity to all other images can be described in terms of degree centrality.

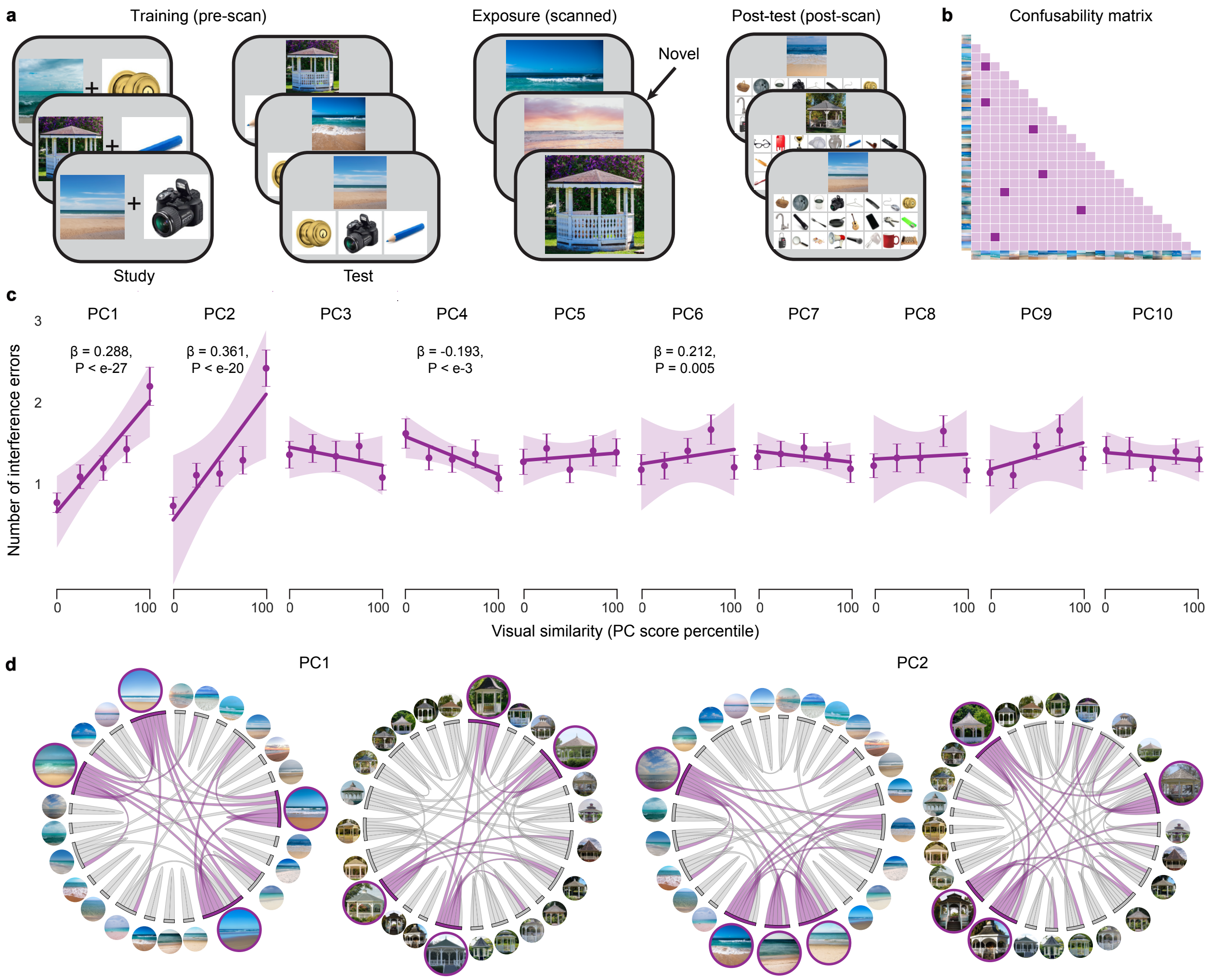


Fig. 2 | Establishing behaviorally relevant dimensions of similarity. a, Overview of the fMRI experiment: participants completed scene–object associative learning before scanning (training phase, left), a novelty detection task during scanning (exposure phase, middle), and an associative memory test after scanning (post-test; right). **b**, Post-test errors yielded a confusability matrix for each participant, indicating which pairs of scenes were confused. **c**, Relationship between visual similarity (PC scores derived as shown in **Fig. 1**) and pairwise confusability in the post-test of the fMRI experiment. Statistical analyses (linear mixed-effects models) treated similarity as a continuous variable. Lines of best fit (solid lines) and S.E.M.s (shaded areas) are shown. Circle markers show the mean number of errors for 5 levels (bins) of similarity. Error bars indicate S.E.M.. These binned data are purely for visualization purposes. For similarity dimensions (PC scores) that significantly predicted the number of interference errors, the corresponding beta estimates and p-values from the mixed-effects models are shown. **d**, Circle graphs indicating the similarity between scenes from each category based on PC1 score (left) and PC2 scores (right). Edge thickness reflects the similarity between a given pair of scenes; only the most strongly connected pairs—those with similarity above the cut-off threshold (0.65–0.72 across the four graphs)—are shown for clarity. For each graph, the four scenes with the highest degree centrality (aggregate connection strength) are shown with larger node sizes and a purple border.

Experience-dependent changes in behavior

The training manipulation in the fMRI experiment (high vs. low training) was used to test for experience-dependent changes in behavior and the hippocampus. To confirm that the training manipulation was successful, we used mixed-effects models to assess behavioral performance during the training phase (prior to scanning), the exposure phase (during scanning), and the post-test (after scanning). During the training phase, there was a main effect of training level (high vs. low; $t_{(307)}$ = 12.313, P < 0.001; **Fig. 3a**, left) and a main effect of test round on memory accuracy ($t_{(307)}$ = 11.460, P < 0.001). These main effects indicate, respectively, that associative memory was (i) better with high vs. low training and (ii) improved across test rounds. When participants did make errors during the training phase, these tended to be 'interference errors': choosing an object that was associated with the same scene category as the target, as opposed to the opposite category ($t_{(615)}$ = 15.091, P < 0.001; **Fig. 3a**, right). For the exposure phase—which only measured scene recognition—discriminability (d') was significantly greater for high- vs. low-training scenes ($t_{(101)}$ = 3.887, P < 0.001; **Fig. 3b**). For the post-test, associative memory accuracy was again better for high- vs. low-training scenes ($t_{(101)}$ = 9.320, P < 0.001; **Fig. 3c**), confirming that the benefits of high training (in the training phase) persisted throughout the experiment.

We next tested whether the relative influence of each dimension of similarity on memory errors changed as a function of training. To test this, we focused on performance during the test rounds of the training phase (of the fMRI experiment). Specifically, we used mixed-effects models in which the dependent variable represented trial-by-trial interference errors (selecting the competitor; **Fig. 2a**). The key independent fixed effects were the degree centrality of the presented scene, separately for PC1 and PC2. We predicted that scenes with higher degree centrality (stronger 'connections' to other scenes; **Fig. 2d**) would be associated with more interference errors.

For PC2, there was a significant interaction between degree centrality and training ($t_{(7464)}$ = 2.408, P = 0.016; **Fig. 3d**); for PC1, the interaction was qualitatively opposite, but not significant ($t_{(7464)}$ = -1.110, P = 0.267; **Fig. 3d**). A linear contrast (F-test) confirmed a significant difference between these two effects ($F_{(1, 7464)}$ = 4.144, P = 0.042; **Fig. 3d**). Qualitatively, the influence of PC1 on memory errors was stronger with low training, whereas the influence of PC2 was stronger with high training. Follow-up tests confirmed that for the low-training condition there was a significant effect of PC1 ($t_{(3732)}$ = 2.044, P = 0.041), but not PC2 ($t_{(3732)}$ = 0.111, P = 0.912) whereas for the high-training condition there was a significant effect of PC2 ($t_{(3732)}$ = 2.596, P = 0.009), but not PC1 ($t_{(3732)}$ = -0.510, P = 0.610).

We next sought to conceptually replicate the preceding findings using data from the training phase of the behavioral memory experiment. Because the behavioral memory experiment did not manipulate training level (high vs. low), we instead focused on changes across the three training rounds, with the rationale being that early training rounds correspond to 'low' training and later training rounds correspond to 'high' training. As in the fMRI experiment, for PC2, there was a significant interaction between degree centrality and training ($t_{(17560)}$ = 2.817, P = 0.005; **Fig. 3e**), and a qualitatively opposite effect for PC1 ($t_{(17560)}$ = -1.654, P = 0.098; **Fig. 3e**). Again, these effects were significantly different from each other ($F_{(1, 17560)}$ = 7.223, P = 0.007; **Fig. 3e**). With low training (here, round 1), there was a significant effect of PC1 ($t_{(5852)}$ = 2.059, P = 0.040), but not PC2 ($t_{(5852)}$ = 0.585, P = 0.559), and with high training (round 3), there was a significant effect of PC2 ($t_{(5852)}$ = 4.171, P < 0.001), but not PC1 ($t_{(5852)}$ = -0.102, P = 0.919). Thus, the pattern of statistical results was entirely consistent across the two experiments, despite minor differences in the paradigm and models. Collectively, these experiments demonstrate that both dimensions of similarity contributed to behavioral interference, but these influences occurred *sequentially* across training. Whereas PC1 drove errors at earlier stages of learning, PC2 drove errors at later stages of learning.

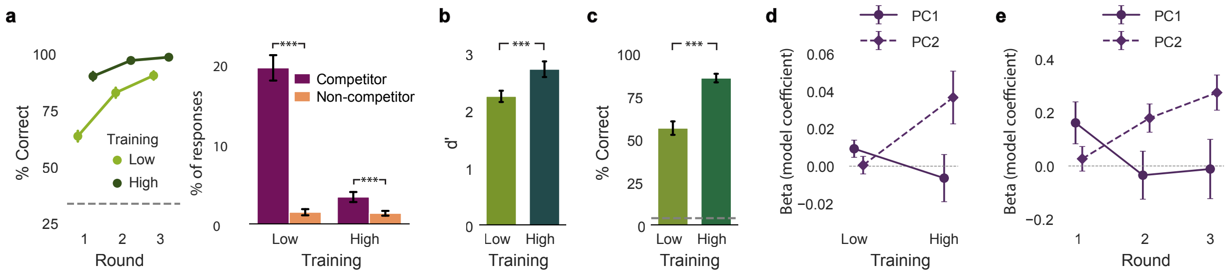


Fig. 3 | Influence of training on behavior. a, Performance during the training phase of the fMRI experiment. Left, associative memory accuracy (% correct responses) for low- vs. high-training scenes during each round. Circle markers represent mean accuracy and error bars indicate S.E.M. Right, percentage of responses corresponding to selection of the competitor vs. non-competitor, separately for low- vs. high-training scenes. **b**, Memory sensitivity (d') during the exposure phase of the fMRI experiment for low- vs. high-training scenes. **c**, Associative memory accuracy during the post-test of the fMRI experiment for low- vs. high-training scenes. In **a** and **c**, dashed lines represent chance performance. For all bar plots in **a-c**, bars represent the mean, and the error bars indicate S.E.M. Asterisks reflect significant paired samples t-tests (***$P < 0.001$). **d**, The influence of each similarity dimension (PC1, PC2) on interference errors during the training phase of the fMRI experiment, separately for low vs. high training categories. **e**, The influence of each similarity dimension (PC1, PC2) on interference errors during the training phase of the behavioral memory experiment, separately for each round of training (round 1 = low training, round 3 = high training). In plots **d** and **e**, the circle markers indicate beta estimates from linear mixed-effects models in which trial-by-trial degree centrality values (from PC1 and PC2) were used to predict the likelihood of an interference error (selecting the competitor). Error bars represent the corresponding S.E.M.s from the models.

Relationships between experience, similarity dimensions, and the hippocampus

We next turned to the central question of how visual similarity and experience influenced representational structure in the hippocampus. Based on prior studies in humans[15,27–32] and rodents[1,4,5,33], we had an *a priori* interest in hippocampal subfields CA3 and dentate gyrus (CA3/DG). For comparison, we also included subfield CA1 and two visual regions: early visual cortex (EVC) and parahippocampal place area (PPA). We anticipated that effects of visual similarity and experience would qualitatively differ between CA3/DG and the visual regions[16,17,27,28]. All fMRI analyses were based on measures of *relative* pattern similarity obtained during the exposure phase of the fMRI experiment. Specifically, for each participant and region, we computed the fMRI pattern similarity (Pearson correlation) for scenes from the same category (within-category similarity) and expressed this relative to similarity between scenes from different categories (across-category similarity) (**Fig. 4a**)[16,17,27,28]. We refer to the difference between these measures (within – across) as the *fMRI similarity score*.

As with our behavioral analyses, we first tested for overall effects of training—that is, whether fMRI similarity scores differed for high vs. low training (**Fig. 4b**). Linear mixed-effects models revealed that, within the visual regions, fMRI similarity scores decreased with training (EVC: $t_{(100)} = -3.234$, $P = 0.002$; PPA: $t_{(100)} = -1.807$, $P = 0.074$). With low training, similarity scores were positive (EVC: $t_{51} = 2.594$, $P = 0.012$; PPA: $t_{51} = 3.339$, $P = 0.002$) but did not differ from 0 with high training ($|t_{51}\text{'s}| < 1.688$, $P\text{'s} > 0.097$). In sharp contrast, similarity scores in CA3/DG significantly *increased* with training ($t_{(101)} = 2.186$, $P = 0.031$). Qualitatively, this reflected a transition from negative similarity scores (low training) to positive similarity scores (high training). However, similarity scores did not significantly differ from 0 at either training level ($|t_{51}\text{'s}| < 1.596$, $P\text{'s} > 0.116$). For CA1, there was no effect of training

($t_{(101)} = -0.283$, $P = 0.777$) and similarity scores did not differ from 0 at either training level ($|t_{51}$'s$| < 0.462$, P's > 0.646). Thus, training influenced representational structure within visual cortical ROIs and CA3/DG, but in entirely opposite ways. It is notable that CA3/DG representations were less similar (greater *differentiation*) in the low training condition despite the fact that, overall, participants experienced more interference in the low training condition (**Fig. 3a**).

We next tested whether similarity within CA3/DG was related to the two dimensions of stimulus similarity (PC1, PC2) and whether these relationships depended on experience (training). While counter-intuitive, prior studies have shown that hippocampal representations are sometimes *less similar* when visual stimuli are *more similar*[15–17,27,28,30,34–39]. This phenomenon—referred to as 'repulsion'—suggests that the hippocampus adaptively responds to memory interference by exaggerating differences between memories that are most similar (or most confusable). Thus, we predicted that the hippocampus (CA3/DG, in particular) would *invert* dimensions of similarity specifically when those dimensions were actively contributing to interference. Based on our behavioral findings (**Fig. 3d**), this translates to the low training condition for PC1 and the high training condition for PC2.

In contrast to the fMRI analyses described above, which computed the average similarity for all pairs of images within a scene category, here we were interested in fine-grained gradations in similarity within a category. Specifically, for each ROI, we ran linear mixed-effects models for which the dependent measure was fMRI pattern similarity for each of the 552 within-category image pairs (276 beach pairs + 276 gazebo pairs). Fixed effects included PC1 and PC2 scores (continuous predictors) and training level (low vs. high).

For CA3/DG, the interaction between PC score and training level was highly significant for both PC1 and PC2, but these interactions took opposite forms (PC1: $t_{(28697)} = 8.628$, $P < 0.001$; PC2: $t_{(28697)} = -6.287$, $P < 0.001$; **Fig. 4c**). Specifically, with low-training, PC1 scores were *negatively* related to CA3/DG similarity ($t_{(14348)} = -4.656$, $P < 0.001$) while PC2 scores were *positively* related to CA3/DG similarity ($t_{(14348)} = 1.981$, $P = 0.048$). With high training, these relationships *flipped*: (PC1: $t_{(14348)} = 3.910$, $P < 0.001$; PC2: $t_{(14348)} = -3.408$, $P < 0.001$). In other words, with low training, CA3/DG inverted similarity along PC1, but preserved similarity along PC2; with high training, CA3/DG inverted similarity along PC2, but preserved similarity along PC1. The fact that CA3/DG selectively inverted PC1 during low training, and PC2 during high training strikingly parallels our behavioral findings that PC1 drove interference during low training and PC2 drove interference during high training. In other words, CA3/DG selectively inverted each dimension of similarity when it was contributing to memory interference. These transformations were putatively adaptive because they dramatically—but transiently—increased the representational distance between the pairs of images that were most at risk for interference (**Fig. 4e**).

To further support the argument that repulsion within CA3/DG was a direct reaction to memory interference, we implemented another mixed-effects model that predicted similarity in CA3/DG, but instead of using PC1 and PC2 similarity, we used similarity values taken directly from the two behavioral similarity experiments: the behavioral memory experiment and the perceptual similarity experiment (**Fig. 1c**). Strikingly, across training levels, CA3/DG similarity was negatively related to similarity derived from the memory experiment ($t_{(28692)} = -2.157$, $P = 0.031$), with a qualitatively opposite effect for the perceptual experiment ($t_{(28692)} = 1.390$, $P = 0.165$; **Extended Data Fig. 3**). In other words, when fully controlling for perceptual similarity judgments, CA3/DG was sensitive to—and *inverted*—dimensions of similarity that contributed to memory interference. Interestingly, independent of any metrics or dimensions of similarity, we also observed that, with high training, CA3/DG inverted the representational structure that was present with low training (**Extended Data Fig. 4**), further demonstrating that CA3/DG transformed information across stages of learning.

Notably, the relationships between PC1/PC2 similarity and training that we observed in CA3/DG were not evident in CA1 (**Fig. 4d**) or in the visual regions (**Extended Data Fig. 2**). Additionally, training-related effects in CA3/DG were absent for other PCs (PCs 3-10; **Extended Data Table 2**).

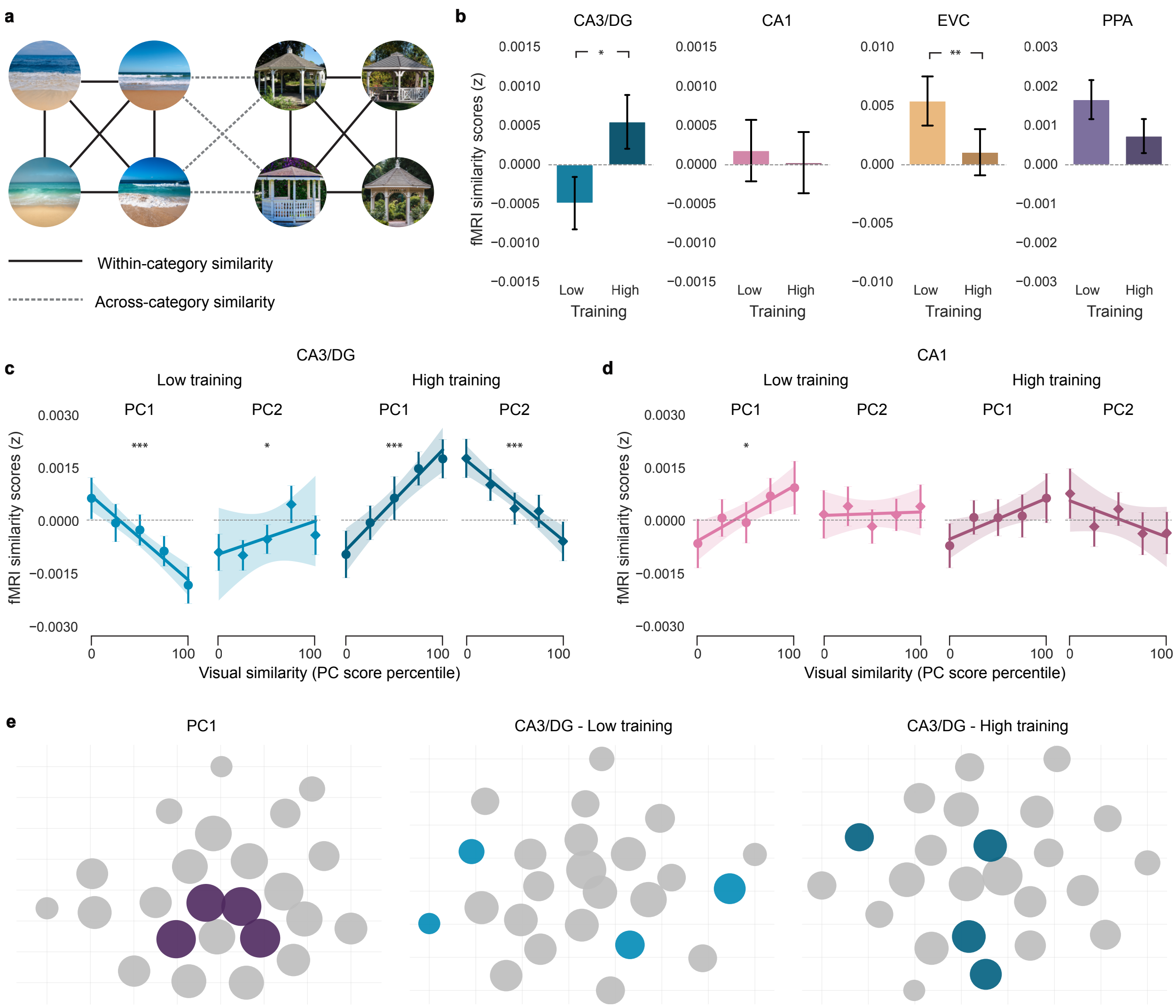


Fig. 4 | Relationships between similarity dimensions, experience, and fMRI pattern similarity. a, Schematic showing within- vs. across-category pattern similarity. **b**, Mean fMRI similarity scores by training level for CA3/DG, CA1, EVC, and PPA. Error bars reflect S.E.M. **c**,**d**, Relationships between visual similarity (PC1 and PC2 scores; circle vs. diamond markers, respectively) and similarity scores in CA3DG (**c**) and CA1 (**d**), separately for each training level (low, high). For **c,d**, statistical analyses were based on linear mixed effects models in which visual similarity (PC scores) were a continuous variable. Solid lines reflect lines of best fit and shaded area reflects S.E.M. Asterisks above plots indicate significant effects of visual similarity (PC score) on fMRI similarity scores (*P < 0.05, **P < 0.01, ***P < 0.001). Markers (circles, diamonds) reflect mean similarity scores at each of 5 levels (bins) of visual similarity. Error bars on markers reflect S.E.M. Binned data are purely for visualization purposes (as in **Fig. 2c**). **e**, 2-D visualization, using t-distributed Stochastic Neighbor Embedding (t-SNE), of the representational structure of the 24 scenes in the beach category, separately for visual similarity space (PC1, left), CA3/DG with low training (center) and CA3/DG with high training (right). Each circle reflects on scene. Circle size is proportional to a scene's degree centrality in the associated space. Distance between circles is proportional to similarity. The four scenes with the highest degree centrality in the visual similarity space are shown in purple (left) and shown again in CA3/DG (light blue in the center and dark blue on the right).

Discussion

Relationships between visual similarity and hippocampal similarity are at the core of theories of hippocampal function[1–5]. Yet, these relationships can take qualitatively different forms[5,12,14–17,27,28,30,31,34–41]. Here, we show that the shape of input-output functions critically depends on the current relevance of individual dimensions of visual similarity to behavioral expressions of memory interference. We first identified two dimensions of visual similarity that robustly contributed to memory interference. We then show that these dimensions of similarity influenced behavior *sequentially*, across stages of learning. Finally, in CA3 and dentate gyrus (CA3/DG), we reveal a striking parallel to these behavioral findings: across learning, CA3/DG inverted each dimension of similarity precisely when it contributed to memory interference. These findings provide important insight into when and why the hippocampus transforms dimensions of visual similarity. Specifically, hippocampal input-output functions dramatically change with experience, and these changes reflect a dynamic prioritization of dimensions of similarity that are most relevant to behavior.

The major innovation in our approach is that we rigorously characterized different dimensions of visual similarity. We did this by aggregating 10 similarity metrics and using PCA to extract orthogonal components (dimensions). Thus, rather than operationalizing dimensions based on an assumed 1-to-1 mapping (metric X = dimension Y), our approach allowed dimensions to be *distributed* across metrics (**Fig. 1i**). Importantly, the dimensions we identified were behaviorally relevant—participants were much more likely to confuse pairs of images that had high PC1 or PC2 scores (**Fig. 2c**). While many prior behavioral studies also show that similarity drives memory interference[7–11] and that interference occurs along specific visual dimensions[42], our findings reveal an important caveat: different dimensions of similarity drive interference at different stages of learning. Indeed, we show—and replicate—that during early stages of learning, there was an influence from PC1, *but not PC2*; in late stages of learning, there was an influence from PC2, *but not PC1* (**Fig. 3d-e**).

The fact that dimensions of similarity drove interference in serial (not parallel), is reminiscent of—and consistent with—findings from the perceptual discrimination literature which show that, when categorizing multidimensional stimuli, humans favor unidimensional strategies[43–47]. Here, PC1 putatively drove interference errors first (followed by PC2) because, by definition, PC1 explained more variance in the similarity matrices (41.3%) than PC2 (19.1%)[46]. The fact that PC1 had no influence during later stages of learning indicates that interference on this dimension was successfully resolved. While caution is warranted when explicitly labeling or describing PCA components, the observed loadings for PC1 and PC2 (**Fig. 1i**) suggest that PC1 reflected relatively lower-level/perceptual information whereas PC2 reflected higher-level/semantic information. This interpretation is potentially compatible with other evidence showing that, with experience, memories transition from perceptual to semantic information[48–51]. That said, we do not predict that the relative prioritization of PC1 vs. PC2 observed here will necessarily generalize across other stimulus sets and procedures. Rather, we believe the current findings reflect a generalizable *principle*: that interference is resolved ‘one dimension at a time’. For example, by changing task demands or the composition of the stimulus set, it may be possible to manipulate which dimensions of similarity are relevant at each stage of learning. However, the key prediction would be that even if dimensions are re-prioritized, interference would still be resolved one dimension at a time.

By considering different dimensions of visual similarity and different levels of experience, we show that relationships between visual similarity and hippocampal similarity (input-output functions) can take remarkably different forms, ranging from strongly positive to strongly negative (**Fig. 4c**; **Extended Data Table 2**). These findings add to recent evidence that relationships between visual similarity and hippocampal similarity are experience-dependent[17,27,28]. However, our behavioral findings provide a compelling framework for understanding these experience-dependent changes. Namely, for PC1 and PC2, CA3/DG either preserved or inverted each dimension according to whether it *actively* contributed to interference. Thus, CA3/DG had access to diverse visual input[52], but *selectively* transformed this input when it was adaptive for behavior[53–55].

The fact that we specifically implicate CA3/DG in resolving memory interference (as opposed to CA1) is consistent with prior human and rodent studies[15,31,36,37,56–58]. In particular, CA3/DG has been implicated in *orthogonalizing* similar memories (pattern separation)[1,5,14]. That said, standard accounts of pattern separation do not predict or explain negative input-output functions[1–5]. In fact, with perfect orthogonalization, input-output functions would be flat, not negative (for additional consideration of this point, see[27,59]). Thus, while reinforcing the relevance of CA3/DG to memory interference, our findings also challenge existing models of how CA3/DG resolves interference.

An important aspect of our design and analyses is that we densely sampled a narrow range at the high end of similarity space. Namely, our models included 276 pairwise comparisons within each visual category. In contrast, the vast majority of prior empirical evidence for negative input-output functions in the hippocampus comes from studies that have only sparsely sampled similarity space[15–17,27,28,30,34,35,37,39,60] (high vs. low similarity; *cf*.[34]). More broadly, the input-output functions we observed in CA3/DG (positive and negative) indicate a striking degree of sensitivity to gradations within a relatively narrow similarity space. In fact, CA3/DG was more sensitive to within-category differences (e.g., differences between beaches) than across-category differences (differences between beaches and gazebos) (i.e., within-category similarity did not systematically differ from across-category similarity; **Fig 4b**).

Although we primarily focus on input-output functions within each visual category, we also observed a surprising main effect of training in CA3/DG: lower similarity scores (greater differentiation) in the low training condition compared to the high training condition (**Fig. 4b**). This effect strongly contrasted with EVC, which showed a robust *decrease* in similarity scores with training. While speculative, these opposing effects are potentially interrelated: CA3/DG differentiation may be most likely to occur when input from visual cortical areas is un-differentiated[16,17,27,28]. From this perspective, if visual cortical areas differentiate stimuli (through experience-dependent attention to diagnostic features[61–63]), then there is less need for additional differentiation by CA3/DG[27,38].

In summary, we show that, in behavior and the hippocampus, memory interference is resolved one dimension at a time. When a particular dimension of visual similarity is actively contributing to memory interference, CA3 and dentate gyrus dramatically—but transiently—invert that dimension. These findings can be interpreted through a simple framework in which the hippocampus selectively exaggerates the representational distance between memories that are most at risk for interference.

METHODS

Participants

Human subject data were collected for three experiments: the behavioral memory experiment, the perceptual similarity experiment, and the fMRI experiment. Across the three experiments, a total of 655 participants were enrolled. All participants' enrollment followed procedures approved by the University of Oregon Institutional Review Board. For the behavioral memory experiment and the fMRI experiment, all participants were right-handed native-English speakers. For the fMRI experiment, all participants had normal or corrected-to-normal vision, with no self-reported psychiatric or neurological disease.

For the behavioral memory experiment, data were collected from 463 participants (200 female, mean age = 28.02, range = 18 – 35 years) using the online platform Prolific. Informed consent was obtained, online, for each participant prior to the experiment. To ensure that the memory confusability matrix obtained from this experiment was highly reliable, participants with low performance were excluded (see 'Experimental procedure – Behavioral memory experiment'). The final analysis included data from 122 participants. Participants in the behavioral memory experiment received monetary compensation.

For the perceptual similarity experiment, data were collected from 141 undergraduate students from the University of Oregon Human Subjects pool (95 female, mean age = 19.04, range = 18 – 27 years) that completed the experiment online through the Sona system. Informed consent was obtained, online, for each participant prior to the experiment. There were two versions of the perceptual similarity experiment: one that only included images of beaches, and another with images of gazebos. Of the 141 participants, 71 participated in the beach experiment and 70 participated in the gazebo experiment. To ensure that the perceptual similarity measure from this experiment was reliable, participants with an insufficient amount of data and/or low performance were excluded (see 'Experimental procedure – Perceptual similarity experiment'). The final analysis included data from 113 participants—57 for the beach experiment and 56 for the gazebo experiment. Participants in the perceptual similarity experiment were compensated with study credit.

For the fMRI experiment, 53 participants (42 female, mean age = 21.23, range = 18 – 31 years) completed the experiment. Written informed consent was collected for each participant prior to the experiment. One participant's data was excluded due to a technical error in the MRI acquisition. All analyses of fMRI data are based on 52 participants. Participants in the fMRI experiment received monetary compensation.

Stimuli

All three experiments used the same set of 24 images of beaches and 24 images of gazebos (**Fig. 1a**). For the fMRI experiment, a separate set of 24 gazebos and 24 beaches were used as lures for the exposure phase. All scene images were color photographs, drawn from Google Images. The behavioral memory experiment and the fMRI experiment also used 48 images of everyday objects (color photographs, drawn from Google Images) shown against a white background. All scene images and objects were square, available in the public domain, and will be shared by the time of publication at https://osf.io/gy23h/ as part of the manuscript's dataset. For the fMRI and behavioral memory experiments, the 48 scenes were randomly paired with the 48 objects, separately for each participant, to form 48 scene-object associations. For the fMRI experiment, one scene category (e.g., gazebos) was assigned as the 'high-training' category, and the other scene category (e.g., beaches) was assigned as the 'low-training' category. The high- and low-training category assignment alternated across participants.

Experimental procedure

Behavioral memory experiment

The behavioral memory experiment, conducted online (using PsychoPy's online platform, Pavlovia), involved learning scene-object associations. It consisted of a training phase followed by a post-test. Across all phases, stimuli were presented on a gray background on a computer screen. The experiment was implemented in PsychoPy2024.1.5 and lasted for about 35 minutes. The training phase consisted of 3 study-test rounds. In each study/test round, all 48 scene-object associations were studied/tested once, in random order. The assignment of objects to scenes was random for each participant. For each trial of the study rounds, a scene and its associated object were displayed next to each other on the screen for 1500 ms (scene on left, object on right), with a white fixation cross in between. Study trials were separated by a 1000 ms inter-trial interval during which a fixation cross was displayed. For each trial of the test rounds, a scene (probe) was displayed at the top of the screen, with 3 objects presented underneath. One of the objects was the target (the object that was paired with the probe during the study rounds). One object was the competitor (an object that was paired with a scene from the same category as the probe). One object was the non-competitor (an object that was paired with a scene from the opposite category as the probe). In each test round, each object appeared once as the target, once as the competitor, and once as the non-competitor. For each scene, the competitor and non-competitor objects were randomly and independently selected in each round. Participants had 10 s to select one of the objects using the keyboard. After making a response, or after 10 s elapsed, the trial ended. Test trials were separated by a 1000 ms inter-trial interval during which a fixation cross was displayed.

The post-test consisted of three rounds, during which each of the 48 studied scenes was tested once per round, in random order. On each post-test trial, a scene (probe) was displayed at the top of the screen, with 24 objects displayed below (3 rows x 8 columns). One of the 24 objects (target) had been paired with the probe. The other 23 objects (competitors) had been paired with a scene from the same category as the probe. Participants were instructed to use their trackpad or mouse to select the object that had been paired with the scene. Participants had 30 seconds to respond on each trial. After responding, or after 30 s elapsed, the trial ended. Post-test trials were separated by a 500 ms inter-trial interval during which a fixation cross was displayed. Post-test data were used to generate a confusability matrix (see 'Metrics of Similarity,' below). The rationale for testing each scene three times during the post-test was to obtain a more reliable confusability matrix.

Because the behavioral memory experiment was used to establish the normative confusability of scene pairs, it was important that the data were internally reliable. Thus, separately for each scene category, we excluded participants whose mean accuracy across the three post-test rounds was below 28%. While this threshold was well above chance (4.17%), we selected this threshold to optimize split-half reliability of the memory confusability measure.

Perceptual similarity experiment

The perceptual similarity experiment, conducted online (using PsychoPy's online platform, Pavlovia), involved making subjective judgments about the relative similarity of scene images. It was modeled after the triplet odd-one-out task described in[20]. Stimuli were presented on a gray background on a computer screen. The experiment was implemented in PsychoPy2024.1.4 and each session lasted approximately 15 minutes. In short, participants completed a single task in which three scenes were shown, and participants select the 'odd one out' (the scene that was least similar to the other two scenes). Each participant in the experiment made perceptual similarity judgments for one of the two visual categories (beach experiment or gazebo experiment). For each experiment (beach or gazebo), participants were allowed to enroll in a second session (for the same category) to be completed within three days of the first session for additional credit. The second session presented triplets that were not shown in the first session. A minimum two-hour break between sessions was enforced. For each category, because there were 24 scenes, there were 2,024 possible triplets. For each participant, the sample of triplets was randomly and independently chosen. On each trial, one of the 2,024 triplets was randomly selected (without replacement within participants). The corresponding three images were presented equidistantly on a virtual circle. On each trial, the image array was randomly rotated by 1–360 degrees to prevent predictable locations and discourage habitual responses. This prevented participants from repeatedly selecting the same screen location without evaluating the presented options. Participants selected the odd image by clicking on it with the trackpad or mouse. There was no time limit for responses. Trials were separated by a 500 ms inter-trial interval during which a fixation cross was displayed. Each session lasted 15 minutes with a self-terminated break after every 50 trials. On average, each participant completed a total of 339 trials per session (SD = 106.0) in the beach experiment and 314 trials per session (SD = 73.4) in the gazebo experiment. 5% of trials were 'catch trials' in which two of the scenes were identical. The catch trials were used to identify participants responding randomly. Participants were not informed about the presence of catch trials. Participants were excluded if their accuracy on catch trials (selecting the objectively different image) was lower than 80%. Data from the perceptual similarity experiment were used to generate a perceptual similarity matrix (see 'Metrics of Similarity,' below). The 80% threshold for exclusion was chosen to optimize split-half reliability of the perceptual similarity matrix.

fMRI experiment

The fMRI experiment, which was entirely in-person, was similar to the behavioral memory experiment, with the two biggest differences being (1) during the training phase, one category received 'high training' and the other category received 'low training,' and (2) the training phase and the post-test were separated by an exposure phase. The training phase was conducted in a preparation room, using a laptop, prior to entering the MRI scanner. Inside the scanner, participants completed eight rounds (functional scans) of the exposure phase. Participants then exited the scanner and completed the post-test, using a laptop, in the preparation room. Across all phases, stimuli were presented on a gray background, either directly on a laptop screen or projected from the back of the MRI scanner. The experiment was implemented in PsychoPy2022.1.1 and lasted for about 2 hours and 30 minutes, with about 1 hour and 15 minutes inside the MRI scanner.

The training phase consisted of the 3 study-test rounds. In each study/test round, all 48 scene object associations were studied/tested once, in random order. For each trial of the study rounds, a scene and its associated object were displayed next to each other on the screen (scene on left, object on right), with a white fixation cross in between. The display time for each study trial varied by training condition (high vs. low) and study/test round. For the high-training category, the display times were: 3000 ms (round 1), 2500 ms (round 2), and 2000 ms (round 3). For the low-training category, the display times were: 2500 ms (round 1), 2000 ms (round 2), and 2000 ms (round 3). Study trials were separated by a 500 ms inter-trial-interval during which a fixation cross was displayed. For each trial of the test rounds, a scene (probe) was displayed at the top of the screen, with 3 objects presented underneath. As in the behavioral memory experiment, one of the objects was the target, one was the competitor, and one was the non-competitor. In each test round, each object appeared once as the target, once as the competitor, and once as the non-competitor. For each scene, the competitor and non-competitor objects were randomly and independently selected in each round. Participants had 10 s to select one of the objects using keyboard. For the high-training category, immediately after a response was made, or after 10 s elapsed, participants received feedback: the two incorrect objects were removed and only the correct object remained on the screen (1000 ms). Feedback was not provided for the low-training category. Test trials were separated by a 1000 ms inter-trial interval during which a fixation cross was displayed.

The exposure phase was conducted during fMRI scanning. In each of the 8 rounds (scan runs), participants were shown each of the 48 previously-studied scenes (24 beaches, 24 gazebos) along with 6 new 'lure' scenes (3 beaches, 3 gazebos). The 54 scenes were presented one at a time, in random order. The lures never repeated across scan runs. Thus, the exposure phase contained a total of 6 (lures) * 8 (runs) = 48 lures. Participants' only task during the exposure phase was to make a button press using their index finger when they detected a new scene—i.e., one not presented during the study phase. On each trial, a scene was displayed for 2000 ms, followed by 4000 ms inter-trial interval during which a white fixation cross was displayed. Participants were instructed to respond (if relevant) by the end of the fixation period.

The post-test consisted of a single round in which each of the 48 studied scenes was tested once, in random order. On each post-test trial, a scene (probe) was displayed at the top of the screen, with 24 objects displayed below (3 rows x 8 columns). As in the behavioral memory experiment, one of the 24 objects was the target, and the other 23 objects were competitors. Participants were instructed to use the trackpad to select the object that had been paired with the scene. Participants had 99 seconds to respond on each trial. After responding, or after 99 s elapsed, the trial ended. Post-test trials were separated by a 500 ms inter-trial interval during which a fixation cross was displayed.

Metrics of similarity

Memory confusability

Responses from the post-test of the behavioral memory experiment were used to generate a normative memory confusability matrix. After excluding participants based on performance thresholds (see 'Experimental procedure – Behavioral memory experiment'), a matrix was generated for each participant that reflected the pairwise confusability for each pair of scenes within a category. For example, in a post-test round, when presented with beach 3 during the post-test, if participants selected the object that had been paired with beach 7 (see **Fig. 1c**, left), this would be considered a 'confusion' between beach 3 and beach 7 (coded as '1' in the confusability matrix for that round, see **Fig. 2b**). For each participant, the confusion matrices over the 3 post-test rounds were averaged, constituting the confusion matrix for that participant. For example, if beach 3 was confused with beach 7 in 2 out of the 3 post-test rounds, that element of the confusability matrix would be 0.667. Participant-level matrices were then averaged. Because the resulting distribution of confusability values was skewed (many pairs were rarely confused, resulting in a left-skew), a log transformation was applied, yielding a distribution of confusability values that was approximately normal.

Perceptual similarity

Responses from the perceptual similarity experiment were used to generate a normative perceptual similarity matrix. Similarity for a given pair of scenes (e.g., beach 2, beach 7 similarity) was defined as the proportion of trials, when beach 2 and beach 7 were presented together, on which the other scene (the third scene) was selected as the 'odd one out'. In other words, the proportion of trials when the 'other' scene was selected reflects the similarity of a given pair of scenes relative to other possible pairs of scenes. For each participant, a similarity matrix was generated where each cell represented the proportion of 'other' responses. After performance-based exclusion of participants (see the details in 'Experimental procedure – Perceptual similarity experiment'), similarity matrices were averaged across participants. Because the resulting distribution of similarity values was skewed (many pairs had low/zero similarity, resulting in a left-skew), a log transformation was applied, yielding a distribution of similarity values that was approximately normal.

CLIP

To quantify the similarity between scenes using CLIP[64], the scene images were passed through the image encoder of the CLIP model (ViT-B/32 variant), a transformer-based deep learning architecture pretrained on a large corpus of paired image-text data. The model and preprocessing pipeline were accessed using the open-source CLIP library (https://github.com/openai/CLIP) implemented in PyTorch. Each image was preprocessed according to the model's specifications, encoded into a high-dimensional feature embedding, and normalized to unit length. Cosine similarity was then computed between all pairs of image embeddings for each scene category. CLIP captures visual and semantic relationships between images due to its joint image–text training.

NLP of text

To compute the similarity between scenes' text descriptions using natural language processing (NLP), descriptive captions were first generated for each scene using OpenAI's ChatGPT (GPT-4-turbo; gpt-4o variant), a multimodal large language model accessed through a ChatGPT Plus subscription. The prompt for each scene image was the following: 'Please describe the visual content of this scene in one clear sentence, using no more than 20 words'. The resulting descriptions for each scene are shown in **Extended Data Fig. 1**. To compute pairwise similarity between the resulting text descriptions, embeddings were generated using the all-mpnet-base-v2 model from the Sentence-Transformers library. This model is based on MPNET, a deep learning architecture that integrates transformer-based attention with permutation-based pretraining to capture rich semantic information in text[65]. Sentence embeddings were computed for each description, and cosine similarity was calculated between each pair of scenes within each category. This approach primarily captured semantic similarity between scenes.

VGG-16

To quantify the similarity between scenes according to VGG-16[66], we first extracted feature representations from the max pooling layers of a pretrained VGG-16 model using PyTorch (version 2.6.0). Input scene images were resized to 224×224 pixels and normalized according to VGG-16's requirements. We passed each scene image through the convolutional feature extractor and collected activation maps from all five max pooling layers, though subsequent analyses focused specifically on layers 1, 3, and 5. Layers 2 and 4 were not considered given their high similarity to layers 1 and 3 respectively. Feature maps were flattened into one-dimensional vectors, and pairwise similarity between scenes was computed as the Pearson correlation between their corresponding feature vectors. This procedure yielded a layer-specific similarity matrix for each selected max pooling layer, reflecting progressively more abstract visual similarity as depth increased within the network hierarchy.

AlexNet

To measure the similarity between scenes according to AlexNet[67], we used a pretrained AlexNet model using PyTorch (version 2.6.0). Each scene image was resized to 224×224 pixels and normalized using ImageNet preprocessing parameters before being passed through the network. 4096-dimensional feature vectors were extracted from the penultimate fully connected layer (fc7) using a forward hook. Pairwise similarity between scenes was computed using cosine similarity. Overall, this approach captures image-based similarity between scenes derived from the visual representations learned by the network.

GIST

GIST descriptor similarities between all pairs of scene images were computed using MATLAB R2023b's LMgist implementation. Each scene image was converted to grayscale and resized to 256×256 pixels. GIST descriptors were calculated using 4 spatial scales with 8 orientations per scale, distributed across a 4×4 grid with a pre-filtering coefficient of 4. Similarity between scene image pairs was quantified as the cosine similarity between their GIST descriptor vectors. GIST similarity reflects the coarse structural and layout information of each scene, capturing its overall gist while ignoring fine-grained features[68].

SSIM

Structural similarity indices (SSIM) were computed between all scene image pairs using MATLAB R2023b's ssim function. Scene Images were converted to grayscale prior to analysis. The SSIM algorithm compared luminance, contrast, and structure between local image patches using an 11×11 Gaussian window, producing similarity values ranging from 0 (no similarity) to 1 (identical images) for each scene pair. SSIM quantifies how similar images are in terms of local luminance, contrast, and structure, emphasizing detailed perceptual features rather than global layout[69].

MRI acquisition

All images were obtained on a Siemens 3T Skyra MRI system in the Lewis Center for Neuroimaging at the University of Oregon. Functional data were obtained using a T2*-weighted echo-planar imaging (EPI) sequence with partial brain coverage that prioritized full coverage of the hippocampus as well as visual cortex (repetition time = 2000 ms, echo time = 34 ms, flip angle = 90°, 66 slices, 1.7 × 1.7 × 1.7 mm voxels). Slices for functional scans were oriented parallel to the long axis of the hippocampus. A total of 8 functional scans were conducted. Each functional scan consisted of 168 volumes and included 6 s of lead-in time at the beginning and 6 s of lead-out time at the end of each scan. Anatomical scans included a whole-brain high-resolution T1- weighted magnetization prepared rapid acquisition gradient-echo anatomical volume (1 95 × 1 × 1 mm voxels) as well as a high-resolution (coronal direction) T2-weighted scan (0.43 × 0.43 × 1.8 mm voxels) to facilitate segmentation of hippocampal subfields.

Anatomical data preprocessing

Preprocessing was conducted with fMRIPrep 23.2.0[70] (RRID:SCR_016216), which is based on Nipype 1.8.6[71] (RRID:SCR_002502). The T1-weighted (T1w) image was corrected for intensity nonuniformity (INU) with N4BiasFieldCorrection[72] (ANTs 2.5.0[73], RRID: SCR_004757) and used as the T1w reference throughout the workflow. The T1w reference was skull-stripped with the antsBrainExtraction.sh workflow (ANTs) in Nipype, using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), gray-matter (GM), and white-matter (WM) was conducted on the brain-extracted T1w using FAST[74] (FSL; RRID:SCR_002823). The T2-weighted image was used to enhance pial surface refinement. Brain surfaces were reconstructed using recon-all (FreeSurfer 7.3.2, RRID:SCR_001847). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was conducted through nonlinear registration with antsRegistration (ANTs 2.5.0), using brain-extracted versions of T1w reference as well as the T1w template. ICBM 152 Nonlinear Asymmetrical template version 2009c[75] was used for spatial normalization (RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing

For each of the 8 functional (BOLD) scans per participant, preprocessing began with the generation of a reference volume for head motion correction. Head-motion parameters relative to this BOLD reference (six translation and rotation parameters plus the corresponding transformation matrices) were estimated. A fieldmap was collected at the scanner to correct for susceptibility-induced distortions. In addition, short EPI reference images with opposite phase-encoding directions were acquired; these were used in FSL's topup to estimate distortions. If the first acquisition of the EPI references was of insufficient quality, it was repeated, resulting in two or more images. The raw fieldmap and EPI references were combined to produce a participant-specific distortion map ('estimated fieldmap'). The estimated fieldmap was then rigidly aligned and applied to the reference EPI to correct distortions. The BOLD reference was next co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration[76]. Co-registration was configured with six degrees of freedom. Multiple time-series of potentially confounding variables were computed based on the preprocessed BOLD: framewise displacement (FD), DVARS, and three region-wise global signals[77]. Furthermore, a set of physiological regressors were extracted for component-based noise correction (CompCor)[78]. Principal components were estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128 s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Frames exceeding a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers and excluded from the Generalized linear models (GLMs). The BOLD time-series were resampled onto fsaverage6. Gridded (volumetric) resampling was performed using nitransforms, configured with cubic B-spline interpolation.

Eight brain masks were generated using fMRIPrep for each of the eight functional scans per participant and the intersection of all eight masks was used as the final brain mask. The processed BOLD time-series were scaled to a mean of 100, with values clipped between 0 and 200. For the eight functional scans, all first-level GLMs were performed in participants' native space with FSL using a Double-Gamma HRF with temporal derivatives, implemented in Python3.10. GLMs were calculated with AFNI's 3dREMLfit[79] using the Least Squares-Separate method[80]: for each scan, a separate GLM was calculated for each of the 48 scene images (24 beaches and 24 gazebos). Thus, for each GLM, there was one regressor of interest (representing a single scene image per scan). All other trials (including lure scene images), FD, xyz translation, xyz rotation, aCompCor 00-05, and csf were represented with nuisance regressors. This model resulted in 48 beta-maps per scan (one map per scene image) which were later converted to t-stats maps that represented the pattern of activity elicited by each scene in each functional scan.

Regions of interest

A region of interest (ROI) for early visual cortex (EVC) was created from the probabilistic maps of Visual Topography[81] in MNI space with a 0.5 threshold. This ROI was transformed into each participant's native space using inverse T1w-to-MNI nonlinear transformation. For each participant, an ROI for the parahippocampal place area (PPA) was created by first using an automated metanalyses in Neurosynth with the key term 'place'. Then, clusters were generated using voxels with a z-score > 2 based on the Neurosynth associative tests. Because these clusters were created through an automated meta-analysis and were not anatomically exclusive to PPA, we visually inspected the results and manually chose the two largest clusters that were spatially consistent with PPA. One cluster was in the left hemisphere (voxel size = 163), and the other cluster was in the right hemisphere (voxel size = 247). These clusters were combined into a single PPA mask. This mask was then transformed into each participant's native space using the inverse T1w-to-MNI transformation. To create hippocampal ROIs, we used the Automatic Segmentation of Hippocampal Subfields (ASHS)[82] toolbox with the upenn2020 (10.1016/j.media.2022.102683) atlas to create subfield ROIs in each participant's hippocampus, including CA3/dentate gyrus (which consisted of CA2, CA3, and dentate gyrus) and CA1. Each participant's subfield segmentations were manually inspected to ensure the accuracy of the segmentation protocol. These ROIs were then manually edited, for each participant, to only include the hippocampal body. This was done by manually identifying the most anterior and posterior slices of the hippocampal body based on each participant's T2-weighted anatomical scan. Next, each subfield ROI was transformed into each participant's native space using the T2-to-T1w transformation, calculated with FLIRT (FSL) with six degrees of freedom, implemented with Nipype. All ROIs were again visually inspected after the transformation to native space to ensure the ROIs were anatomically accurate.

fMRI pattern similarity scores

fMRI pattern similarity was defined as the Pearson correlation between t-stats maps from a given ROI. All correlations were restricted to t-maps from different scan runs; that is, correlations were never computed between t-maps (trials) from the same scan run. Correlations were performed for all pairwise comparisons of runs and then averaged. For example, to compute the pattern similarity between beach 4 and beach 2 within a given ROI, 56 Pearson correlations were computed, reflecting all possible combinations across different scan runs (e.g., beach 4–run 1 vs. beach 2–run 2, beach 4–run 3 vs. beach 2–run 7, etc.). The 56 correlation coefficients were then Fisher z-transformed and averaged to yield a single similarity value for that scene pair.

For each ROI, pattern similarity between a pair of scenes from the same category (within-category similarity, e.g., beach 4 vs. beach 2) was always expressed relative to the mean pattern similarity between all scenes from opposite categories (across-category similarity). The difference between these measures (within-category minus across-category) is referred to as the *similarity score*. Thus, if a given pair of images from the same category (or a set of images from the same category) has a positive similarity score, this indicates that the brain region (ROI) represented these scenes as more similar than scenes from opposite categories[17,27]. The rationale for using across-category similarity as a baseline is that it facilitates comparisons of representational structure across ROIs. Namely, raw correlations values are difficult to interpret and vary substantially across ROIs. In contrast, *relative* similarity (within- *versus* across-category) corrects for these differences.

Statistical analyses and reproducibility

Split-half reliability for human-based similarity metrics

For human-based measures of memory confusability and perceptual similarity, we conducted reliability analyses. Specifically, split-half reliability was computed separately for each category (beaches, gazebos) as a function of the number of participants included (participant count). For each participant count, 50 random permutations were generated by selecting that number of participants and randomly dividing them into two equal groups. For each participant group, confusability/similarity matrices were averaged across participants. Reliability was defined as the Pearson's correlation (r) between the matrices for each group. To compute the overall reliability across all 50 permutations, correlation values were first converted to Fisher's z, then averaged, and then converted back to r (for interpretability).

Mixed-effects models

Mixed-effects (LME) models were used to evaluate the extent to which different similarity dimensions, training level, and the interaction between training level and each similarity dimension predicted behavior and/or fMRI similarity scores. The fixed effects (predictors) included in each model are mentioned in the sections below. Critically, in every model that involved similarity dimensions, random intercepts and slopes for similarity dimensions were included for each participant to account for individual differences. Interactions between the identity of the high training category (beaches vs. gazebos) and other fixed effects were included only when they significantly improved model fit or revealed interpretable effects, to avoid overfitting and obscuring lower-order effects. All models were implemented using fitlme or fitglme in MATLAB R2023b, with model comparisons guided by likelihood ratio tests and inspection of residuals to ensure model assumptions were met. Degrees of freedom were computed using the Satterthwaite approximation.

PC scores, memory confusability, perceptual similarity, and normative low training fMRI similarity scores were treated as continuous measures in mixed-effects models. For visualization purposes only, data were also binned (see **Fig 2c**, **Fig. 4c-d**, **Extended Data Fig. 2-4**). For the linear mixed-effects models corresponding to **Fig. 4c,d** and **Extended Data Fig. 2**, the models included both PC1 and PC2 similarity as fixed effects. The lines of fit shown in these figures therefore reflect the effect of one PC while controlling for the other PC. To ensure that the binned visualizations accurately reflect the model, we regressed out PC1 from PC2 (and vice versa) within each scene category before binning the PC scores. In other words, the binned data reflect the effect of one PC while controlling for the other PC. Again, this step (regressing one PC out from the other) was purely for the sake of accurate visualization for the binned data. The same procedure was applied to memory confusability and perceptual similarity in **Extended Data Fig. 3** (i.e., memory confusability was regressed out from perceptual similarity, and vice versa).

Establishing behaviorally relevant dimensions of similarity

To identify stimulus similarity dimensions that predicted memory interference errors (**Fig. 2c**), we tested whether PC scores (derived from the 10 metrics of similarity) predicted pairwise confusability in the post-test of the fMRI experiment. Specifically, we ran mixed-effects models for which pairwise confusability was the dependent measure and scores on one of the 10 PCs (**Fig. 1g**) were included as a predictor (separate models for each PC). For each participant, the model included 552 rows corresponding to the 552 pairs of scene images (276 beach pairs and 276 gazebo pairs). If a pair of scenes was confused on the post-test, pairwise confusability for that row was coded as '1';

otherwise, confusability was '0' (see **Fig. 2b**). Additional fixed effects included training level (high vs. low), and the designated high-training category for each participant (beaches vs. gazebos).

Experience-dependent changes in behavior

To validate the effectiveness of the training manipulation in the fMRI experiment, we analyzed behavioral performance across the training, exposure, and post-test phases using mixed-effects models (**Fig. 3a-c**). Note: these models were based on mean responses per participant/condition (i.e., they did not include trial-level data). For the training phase, two separate models were run. For the first model (**Fig. 3a**, left), the dependent measure was accuracy (% correct) in the test rounds. Fixed effects included training level (high vs. low), training round (round 1 to round 3), and the participant's designated high-training category (beaches vs. gazebos). This model was used to test for main effects of training round and training level. The second model (**Fig. 3a**, right) focused on errors. For this model, the dependent measure was the frequency of responses and fixed effects included error type (competitor vs. non-competitor), training level (high vs. low), training round (round 1 to round 3), and the participant's designated high-training category (beaches vs. gazebos). This model was used to test whether interference errors (selecting the competitor) were more common than non-interference errors (selecting the non-competitor). For the exposure phase (**Fig. 3b**) and post-test phase (**Fig. 3c**), the dependent measures in the mixed-effects models were d-prime and accuracy (% correct), respectively and fixed effects included training level (high vs. low) and the participant's assigned high-training category (beaches vs. gazebos).

To measure the influence of different dimensions of similarity (PC1 vs. PC2) on interference errors during the training phases of the fMRI experiment and the behavioral memory experiment (**Fig. 3d-e**), we used logistic mixed-effects models that included binary trial-by-trial measures of performance. Specifically, the dependent variable in the models represented whether, on each trial, participants made an interference error (selecting the competitor, coded as '1') or correctly selected the target object (coded as '0') (**Fig. 2a, 3a**, right). While rare, trials on which participants selected the non-competitor were excluded from the model. The key fixed effects were the degree centrality of the presented scene (the probe), separately for PC1 and PC2. The rationale for these models was that, in general, higher degree centrality (greater aggregate similarity to other scenes from the same category) would be associated with a higher rate of interference errors. For the fMRI experiment (**Fig. 3d**), other fixed effects included in the models were the training level (high vs. low), training round (round 1 to round 3), and the designated high-training category for the participant (beaches vs. gazebos). For the behavioral memory experiment (**Fig. 3e**), other factors included in the models were the training round (round 1 to round 3), and the category of the current probe (beach vs. gazebo).

To formally compare interaction effects with training across PC1 and PC2, we performed a linear contrast test (F-test) on the associated fixed effects of our model to evaluate the significance of coefficient differences (i.e., H_0: $\beta_{(PC1 \times training)} - \beta_{(PC2 \times training)} = 0$ for the fMRI experiment, H_0: $\beta_{(PC1 \times round)} - \beta_{(PC2 \times round)} = 0$ for the behavioral memory experiment).

Relationships between experience, similarity dimensions, and the hippocampus

To test the overall effects of training on fMRI similarity scores for each ROI (**Fig. 4b**), we used linear mixed-effects models where the dependent measure was the mean fMRI similarity score, and fixed effects included training level (high vs. low) and the participant's designated high-training category (beaches vs. gazebos). Note: these models were based on mean responses per participant/condition (i.e., they did not include trial-level data). In addition, to test whether an ROI's average similarity score was above zero, two-tailed one-sample t-tests (contrast value = zero) were conducted.

To test whether fMRI similarity scores within each ROI were related to dimensions of visual similarity (PC1, PC2 for **Fig. 4c-d** and **Extended Data Fig. 2**; all 10 PCs for **Extended Data Table 2**) and training level (low vs. high), we ran a separate set of linear mixed-effects models. In these models, the dependent measure was the fMRI similarity score for every pairwise combination of scenes from the same category (baselined against the mean across-category similarity). Specifically, for each participant, the model included 552 fMRI similarity values, reflecting every pairwise combination of beaches (276 total) and gazebos (276 total). Fixed effects included PC scores (continuous predictors), training level (low vs. high), and the high-training scene category for each participant (beaches vs. gazebos).

To further support the argument that repulsion within CA3/DG was a direct reaction to memory interference, we implemented another mixed-effects model that predicted similarity scores in CA3/DG, but instead of using PC1 and PC2 scores as fixed effects, we used similarity values taken directly from the two behavioral similarity experiments (**Fig. 1c**): the behavioral memory experiment and the perceptual similarity experiment (**Extended Data Fig. 3**). In these models, the dependent measure was CA3/DG similarity scores (552 values per participant) and fixed effects included perceptual similarity and memory confusability values (continuous predictors), training level (low vs. high), and the high-training scene category for each participant (beaches vs. gazebos).

Finally, for each ROI and each scene category, we tested the extent to which the representational structure after high training preserved vs. transformed the representational structure after low training (independent of any similarity metric or PC scores) (**Extended Data Fig. 4**). Because the assignment of each scene category to each training condition was counterbalanced *across participants*, this analysis required across-participant alignment. Specifically, for each participant, pairwise similarity scores in the high training condition were treated as the dependent measure whereas similarity scores in the low training condition *from other participants* were the critical predictor (fixed effect). For example, for a given participant, if the high training category was beaches, then the dependent measure for that participant would be the 276 pairwise similarity scores for beaches. To obtain the corresponding similarity scores for the low training condition, the mean similarity score (for each of the 276 pairs of beaches) was computed from all participants (N = 26) for which the low training category was beaches. Thus, participant-specific and category-specific representational structure in the high training condition was considered as a function of *normative representational structure* in the low training condition. We refer to this normative measure as the 'normative low-training fMRI similarity scores'. For the mixed-effects models, the high-training category (beaches vs gazebos) was also included as a fixed effect.

Statistical significance

A threshold of $p < .05$ was used to determine statistical significance. All t-tests were two-tailed. To correct for multiple comparisons when testing effects across all 10 PCs (**Fig. 2c, Extended Data Table 2**), we applied a Bonferroni-adjusted significance threshold of $\alpha = 0.005$.

Data availability

All data and materials will be available by the time of publication at https://osf.io/gy23h/.

Code availability

Analysis codes used for the analysis will be available by the time of publication at https://osf.io/gy23h/.

Acknowledgements This work was supported by National Institutes of Health grant NIH-NINDS 2R01NS089729 to B.A.K and Ruth L. Kirschstein NRSA Individual Predoctoral Fellowship (from the National Institutes of Health) grant F31NS126016 to W.G.

Author contributions Conceptualization: S.M., W.G., D.G., U.M., B.A.K. Methodology: S.M., W.G., D.G., U.M., E.W., B.A.K. Investigation: S.M., W.G., D.G., B.A.K. Visualization: S.M., B.A.K. Funding acquisition: W.G., B.A.K. Project administration: U.M., B.A.K. Supervision: U.M., B.A.K. Writing, first draft: S.M., B.A.K. Review and editing: S.M., W.G., D.G., B.A.K.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Soroush Mirjalili or Brice A. Kuhl

References

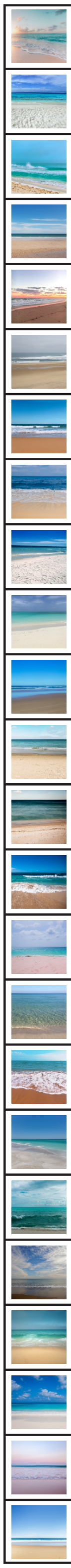
1. Colgin, L. L., Moser, E. I. & Moser, M.-B. Understanding memory through hippocampal remapping. *Trends Neurosci.* **31**, 469–477 (2008).
2. O'Reilly, R. C. & McClelland, J. L. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* **4**, 661–682 (1994).
3. O'Reilly, R. C. & Rudy, J. W. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychol. Rev.* **108**, 311–345 (2001).
4. Rolls, E. The mechanisms for pattern completion and pattern separation in the hippocampus. *Front. Syst. Neurosci.* **7**, (2013).
5. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).
6. Ritvo, V. J., Nguyen, A., Turk-Browne, N. B. & Norman, K. A. A neural network model of differentiation and integration of competing memories. *eLife* **12**, RP88608 (2024).
7. Anderson, M. C. & Spellman, B. A. On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychol. Rev.* **102**, 68–100 (1995).
8. Barnes, J. M. & Underwood, B. J. 'Fate' of first-list associations in transfer theory. *J. Exp. Psychol.* **58**, 97–105 (1959).
9. Mensink, G.-J. & Raaijmakers, J. G. A model for interference and forgetting. *Psychol. Rev.* **95**, 434–455 (1988).
10. Osgood, C. E. The similarity paradox in human learning: a resolution. *Psychol. Rev.* **56**, 132–143 (1949).
11. Wixted, J. T. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* **55**, 235–269 (2004).
12. Knierim, J. J. & Neunuebel, J. P. Tracking the flow of hippocampal computation: Pattern separation, pattern completion, and attractor dynamics. *Neurobiol. Learn. Mem.* **129**, 38–49 (2016).
13. Liang, J. C., Wagner, A. D. & Preston, A. R. Content Representation in the Human Medial Temporal Lobe. *Cereb. Cortex* **23**, 80–96 (2013).
14. LaRocque, K. F. *et al.* Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *J. Neurosci.* **33**, 5466–5474 (2013).
15. Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P. & Ranganath, C. CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nat. Commun.* **9**, 294 (2018).
16. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Curr. Biol.* **27**, 2307-2317.e5 (2017).
17. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat. Commun.* **7**, 11066 (2016).
18. Mahner, F. P., Muttenthaler, L., Güçlü, U. & Hebart, M. N. Dimensions underlying the representational alignment of deep neural networks with humans. *Nat. Mach. Intell.* **7**, 848–859 (2025).

19. Naspi, L. *et al.* Multiple dimensions of semantic and perceptual similarity contribute to mnemonic discrimination for pictures. *J. Exp. Psychol. Learn. Mem. Cogn.* **47**, 1903–1923 (2021).
20. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
21. King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J. & Baker, C. I. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage* **197**, 368–382 (2019).
22. Doerig, A. *et al.* High-level visual representations in the human brain are aligned with large language models. *Nat. Mach. Intell.* 1–15 (2025) doi:10.1038/s42256-025-01072-0.
23. Hout, M. C., Goldinger, S. D. & Brady, K. J. MM-MDS: A Multidimensional Scaling Database with Similarity Ratings for 240 Object Categories from the Massive Memory Picture Database. *PLOS ONE* **9**, e112644 (2014).
24. Kaniuth, P., Mahner, F. P., Perkuhn, J. & Hebart, M. N. A high-throughput approach for the efficient prediction of perceived similarity of natural objects. *eLife* **14**, (2025).
25. Groen, I. I. A., Silson, E. H. & Baker, C. I. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. B Biol. Sci.* https://doi.org/10.1098/rstb.2016.0102 (2017) doi:10.1098/rstb.2016.0102.
26. Walsh, C. R. & Rissman, J. Behavioral representational similarity analysis reveals how episodic learning is influenced by and reshapes semantic memory. *Nat. Commun.* **14**, 7548 (2023).
27. Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J. & Kuhl, B. A. Abrupt hippocampal remapping signals resolution of memory interference. *Nat. Commun.* **12**, 4816 (2021).
28. Wanjia, G., Han, S. & Kuhl, B. A. Repulsion of hippocampal representations driven by distinct internal beliefs. *Curr. Biol.* **35**, 2893-2902.e5 (2025).
29. Bakker, A., Kirwan, C. B., Miller, M. & Stark, C. E. L. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).
30. Kim, G., Norman, K. A. & Turk-Browne, N. B. Neural Differentiation of Incorrectly Predicted Memories. *J. Neurosci.* **37**, 2022–2031 (2017).
31. Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A. & Preston, A. R. Memory Reactivation during Learning Simultaneously Promotes Dentate Gyrus/$CA_{2,3}$ Pattern Differentiation and CA_1 Memory Integration. *J. Neurosci.* **41**, 726–738 (2021).
32. Yassa, M. A. *et al.* Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus* **21**, 968–979 (2011).
33. Knierim, J. J. The hippocampus. *Curr. Biol.* **25**, R1116–R1121 (2015).
34. Wammes, J., Norman, K. A. & Turk-Browne, N. Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *eLife* **11**, e68344 (2022).
35. Zheng, L., Gao, Z., McAvan, A. S., Isham, E. A. & Ekstrom, A. D. Partially overlapping spatial environments trigger reinstatement in hippocampus and schema representations in prefrontal cortex. *Nat. Commun.* **12**, 6231 (2021).
36. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* **153**, 40–56 (2018).

37. Schlichting, M. L., Mumford, J. A. & Preston, A. R. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* **6**, 8151 (2015).
38. Hulbert, J. C. & Norman, K. A. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cereb. Cortex* **25**, 3994–4008 (2015).
39. Ballard, I. C., Wagner, A. D. & McClure, S. M. Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* **10**, 1073 (2019).
40. Bein, O. & Davachi, L. Event Integration and Temporal Differentiation: How Hierarchical Knowledge Emerges in Hippocampal Subfields through Learning. *J. Neurosci.* **44**, e0627232023 (2024).
41. Morton, N. W., Sherrill, K. R. & Preston, A. R. Memory integration constructs maps of space, time, and concepts. *Curr. Opin. Behav. Sci.* **17**, 161–168 (2017).
42. Drascher, M. L. & Kuhl, B. A. Long-term memory interference is resolved via repulsion and precision along diagnostic memory dimensions. *Psychon. Bull. Rev.* **29**, 1898–1912 (2022).
43. Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U. & Waldron, E. M. A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* **105**, 442–481 (1998).
44. Ashby, F. G., Queller, S. & Berretty, P. M. On the dominance of unidimensional rules in unsupervised categorization. *Percept. Psychophys.* **61**, 1178–1199 (1999).
45. Egeth, H. E. Parallel versus serial processes in multidimensional stimulus discrimination. *Percept. Psychophys.* **1**, 245–252 (1966).
46. Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychol. Rev.* **101**, 53–79 (1994).
47. Pothos, E. M. & Close, J. One or two dimensions in spontaneous classification: A simplicity approach. *Cognition* **107**, 581–602 (2008).
48. Krenz, V., Alink, A., Sommer, T., Roozendaal, B. & Schwabe, L. Time-dependent memory transformation in hippocampus and neocortex is semantic in nature. *Nat. Commun.* **14**, 6037 (2023).
49. Lifanov, J., Linde-Domingo, J. & Wimber, M. Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nat. Commun.* **12**, 3177 (2021).
50. Winocur, G. & Moscovitch, M. Memory Transformation and Systems Consolidation. *J. Int. Neuropsychol. Soc.* **17**, 766–780 (2011).
51. Zeng, T., Tompary, A., Schapiro, A. C. & Thompson-Schill, S. L. Tracking the relation between gist and item memory over the course of long-term memory consolidation. *eLife* **10**, e65588 (2021).
52. Turk-Browne, N. B. The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis. *Vision Res.* **165**, 123–130 (2019).
53. Biane, J. S. *et al.* Representations of stimulus features in the ventral hippocampus. *Neuron* **0**, (2025).
54. Theves, S., Fernández, G. & Doeller, C. F. The Hippocampus Maps Concept Space, Not Feature Space. *J. Neurosci.* **40**, 7318–7325 (2020).

55. Komorowski, R. W. *et al.* Ventral Hippocampal Neurons Are Shaped by Experience to Represent Behaviorally Relevant Contexts. *J. Neurosci.* **33**, 8079–8087 (2013).
56. Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M. & Tanila, H. The Hippocampus, Memory, and Place Cells: Is It Spatial Memory or a Memory Space? *Neuron* **23**, 209–226 (1999).
57. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160049 (2017).
58. Schlichting, M. L. & Preston, A. R. Memory integration: neural mechanisms and implications for behavior. *Curr. Opin. Behav. Sci.* **1**, 1–8 (2015).
59. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends Cogn. Sci.* **23**, 726–742 (2019).
60. Schapiro, A. C., Kustner, L. V. & Turk-Browne, N. B. Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Curr. Biol.* **22**, 1622–1627 (2012).
61. Gibson, E. J. *Principles of Perceptual Learning and Development*. (Appleton-Century-Crofts, East Norwalk, CT, US, 1969).
62. Kellman, P. J. & Garrigan, P. Perceptual learning and human expertise. *Phys. Life Rev.* **6**, 53–84 (2009).
63. Petrov, A. A., Dosher, B. A. & Lu, Z.-L. The Dynamics of Perceptual Learning: An Incremental Reweighting Model. *Psychol. Rev.* **112**, 715–743 (2005).
64. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).
65. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. MPNet: Masked and Permuted Pre-training for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.2004.09297 (2020).
66. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2015).
67. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun ACM* **60**, 84–90 (2017).
68. Oliva, A. & Torralba, A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope.
69. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
70. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
71. Gorgolewski, K. *et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinformatics* **5**, 13 (2011).
72. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
73. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).

74. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
75. Fonov, V., Evans, A., McKinstry, R., Almli, C. & Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
76. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
77. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
78. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
79. Cox, R. W. & Hyde, J. S. Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 171–178 (1997).
80. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
81. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex N. Y. N 1991* **25**, 3911–3931 (2015).
82. Yushkevich, P. A. *et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.* **36**, 258–287 (2015).

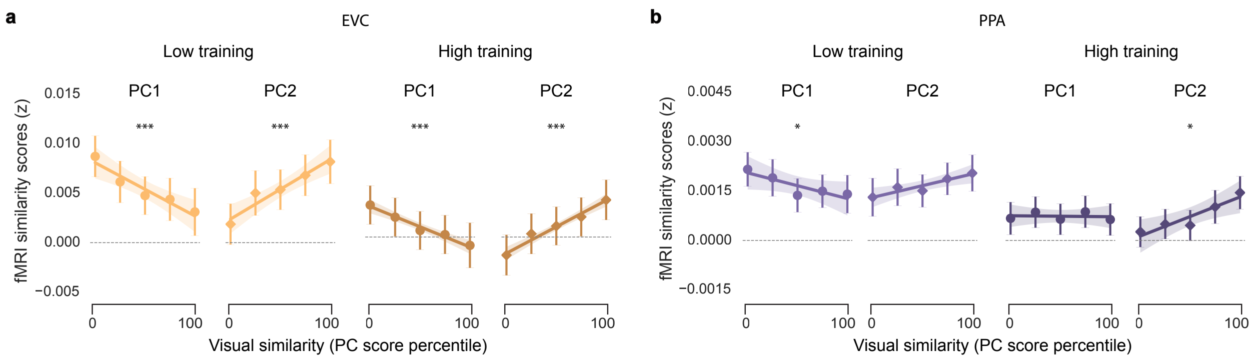


A calm turquoise ocean meets a sandy beach under a partly cloudy sky at sunset.	A small white gazebo with a wooden bench sits in a lush green garden under bright sunlight.
Bright white sand leads to crystal-clear turquoise waters under a vivid blue sky with light scattered clouds.	A weathered white gazebo stands on green grass, with vibrant purple flowering bushes in the background.
Gentle waves crash onto a sandy beach under a clear blue sky, leaving foamy trails along the shoreline.	An ornate white gazebo with a shingled roof and decorative trim stands in a garden surrounded by greenery and houses.
A peaceful beach stretches under a clear blue sky, with gentle waves meeting the sandy shore and a single bird flying.	A large white gazebo with a wooden shingled roof stands elevated among neatly trimmed bushes and lush green trees.
Waves crash onto a sandy beach under a colorful sunset sky filled with scattered clouds.	A classic white gazebo with a shingled roof stands surrounded by bushes and trees on a sunny day.
A wide, empty beach stretches to the ocean under a cloudy sky, with gentle waves rolling toward the shore.	A large, elegant white gazebo with a double-tiered roof stands surrounded by lush flowers and greenery under a cloudy sky.
Bright waves crash onto a smooth, golden beach under a vibrant blue sky with a few distant clouds.	A tall-roofed, white and red brick gazebo stands on a grassy area, surrounded by trees under a cloudy sky.
Gentle waves roll onto a soft, sandy beach under a lightly clouded blue sky.	A simple white gazebo with a gray shingled roof stands in a garden surrounded by plants and trees.
Clear turquoise waves gently meet a bright white sandy beach under a vivid blue sky with light clouds.	A charming white gazebo with a pointed shingled roof stands surrounded by lush greenery and colorful flowers in a garden.
Calm turquoise water meets a soft, white sandy beach under a lightly clouded sky.	A white wooden gazebo with a shingled roof and seating inside stands surrounded by lush greenery on a sunny day.
A wide, smooth beach reflects the blue sky as gentle waves roll in from the deep ocean.	A small white gazebo with a gray tiled roof stands on a wooden platform surrounded by grass and trees.
Gentle waves lap against a sandy beach under a bright, lightly clouded blue sky.	An elegant white gazebo with lattice columns and a tall dark roof stands surrounded by lush green bushes and trees.
Clear, rippling water gently washes onto a pebbly, shell-strewn beach under a bright sky with a few small clouds.	A tall white gazebo with a pointed green roof stands elevated on gray steps, surrounded by neatly trimmed bushes and trees.
Bright turquoise waves crash onto a golden sandy beach under a vivid blue sky with a few scattered clouds.	A large white gazebo with a dark curved roof and dome top stands elevated among autumn trees and colorful shrubs.
Soft pink sand meets clear turquoise water under a lightly clouded sky, with gentle waves lapping the shore.	An ornate white gazebo with a dark shingled roof stands elevated, surrounded by colorful bushes and tall green trees.
Crystal-clear water gently laps onto smooth, light-colored sand under a cloudless blue sky.	A white gazebo with a gray shingled roof stands surrounded by green grass, bushes, and blooming pink flowers.
Foamy waves roll onto a smooth, golden beach under a bright blue sky with a distant boat on the horizon.	A grand white gazebo with thick columns and a green-tiled domed roof stands in a lush, tree-filled park.
Soft waves meet a pale, sandy beach under a bright blue sky, with turquoise water stretching to the horizon.	A simple white gazebo with lattice sides and a metal roof stands on a concrete pad in a grassy park.
Dark turquoise waves crash onto a sandy beach under a cloudy, dramatic sky stretching across the horizon.	An elevated, ornate white gazebo with a steep wooden roof stands in an open field, with houses visible in the background.
Calm waves gently meet a dark sandy beach under a sky filled with soft, scattered clouds and patches of blue.	A white wooden gazebo with a dark shingled roof stands on a brick patio, surrounded by trees with autumn foliage.
A soft wave rolls onto a smooth golden beach under a hazy, blue-green sky.	A light gray wooden gazebo with a shingled roof stands in a grassy park, surrounded by colorful autumn trees.
Bright turquoise waves gently wash onto a pristine white sandy beach under a vivid blue sky dotted with fluffy clouds.	A white wooden gazebo with a shingled roof stands on a grassy hill, surrounded by vibrant green trees.
Soft waves meet a smooth beach under a pastel sky, blending shades of pink, purple, and blue at sunset.	A small white gazebo with a weathered metal roof stands among tall trees and lush greenery in a peaceful park.
A calm beach with gentle waves stretches under a clear, bright blue sky on a sunny day.	A large white gazebo with a brown shingled roof sits surrounded by bushes and trees near a paved pathway.

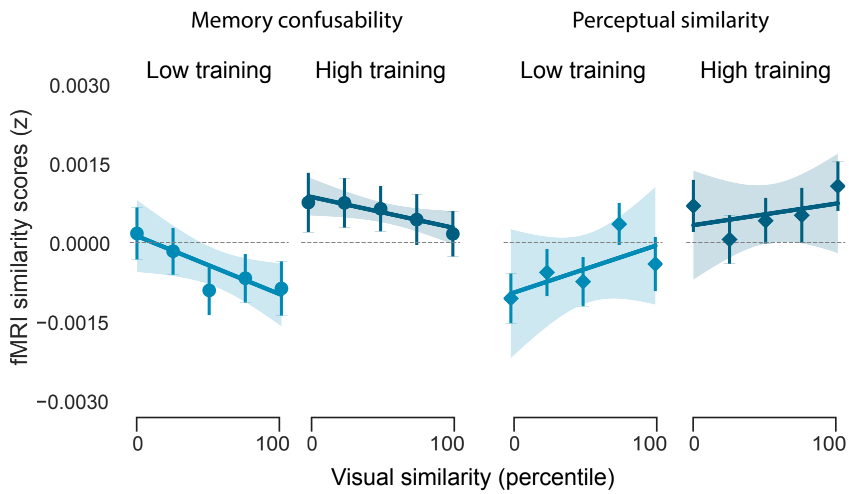
Extended Data Fig. 1 | Text descriptions, generated by OpenAI's ChatGPT (GPT-4-turbo; gpt-4o variant), for each scene image. Text descriptions were used for the Natural Language Processing (NLP) similarity metric.

	Memory	Percept.	CLIP	NLP	VGG MP1	VGG MP3	VGG MP5	Alex.	GIST	SSIM
PC1	0.240	0.191	0.236	-0.081	0.427	0.442	0.332	0.426	-0.127	0.399
PC2	0.438	0.515	0.272	0.482	-0.106	-0.146	0.113	0.005	0.325	-0.293
PC3	-0.236	-0.124	-0.252	-0.154	-0.064	0.166	0.534	0.104	0.705	-0.128
PC4	-0.410	-0.443	0.570	0.537	0.038	0.082	0.118	-0.015	0.007	0.020
PC5	-0.017	0.044	0.689	-0.662	-0.157	-0.119	-0.064	-0.005	0.139	-0.150
PC6	0.615	-0.511	0.044	-0.023	0.146	0.176	-0.154	-0.408	0.304	0.154
PC7	-0.369	0.365	0.028	0.004	0.568	0.131	-0.412	-0.315	0.343	0.068
PC8	0.093	-0.278	-0.062	0.022	0.277	-0.049	-0.425	0.665	0.148	-0.432
PC9	0.057	-0.087	-0.014	-0.101	0.545	-0.192	0.432	-0.284	-0.310	-0.531
PC10	-0.019	0.081	-0.005	-0.007	-0.245	0.805	-0.116	-0.140	-0.183	-0.467

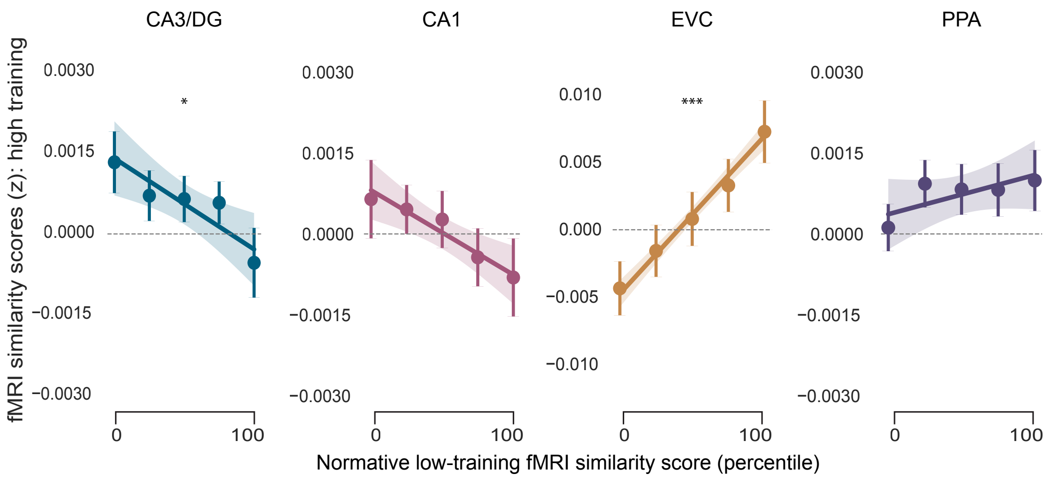
Extended Data Table 1 | Loadings for all 10 Principal Components (PCs) from the Principal Component Analysis described in Fig. 1. Rows correspond to the PC number and columns correspond to each of the 10 similarity metrics.



Extended Data Fig. 2 | Relationships between visual similarity, training level, and fMRI similarity scores in visual cortical regions. We tested whether fMRI similarity scores in early visual cortex (EVC, **a**) and parahippocampal place area (PPA, **b**) were related to visual similarity on each dimension (PC1 and PC2 scores) and/or training level (low vs. high training). This was tested via linear mixed-effects models that were identical, in structure, to the models used for CA3/DG and CA1 (**Fig. 4c-d**). **a,** For EVC, there were significant main effects of PC1 ($t_{(28692)}$ = -4.812, $P < 0.001$) and PC2 ($t_{(28692)}$ = 6.461, $P < 0.001$), however these effects were in opposite directions (negative effect for PC1, positive effect for PC2). While the effect for PC1 was qualitatively similar across training levels (negative in both cases: $t_{(14348)}$'s < -3.619, P's < 0.001), there was a significant interaction between training level and PC1 ($t_{(28692)}$ = 3.437, P = 0.005). The effect for PC2 was positive across training levels ($t_{(14348)}$'s > 5.777, P's < 0.001) and there was no interaction between training level and PC2 ($t_{(28692)}$ = -1.914, P = 0.056). **b**, For PPA, there was a significant main effect for PC2 ($t_{(28692)}$ = 2.589, P = 0.010), but not PC1 ($t_{(28692)}$ = -1.158, P = 0.247). There was a significant interaction between training and PC1 ($t_{(28692)}$ = 2.636, P = 0.008) but no significant interaction between training and PC2 ($t_{(28692)}$ = 1.666, P = 0.096). For PC1, there was a negative effect of PC score for the low training condition ($t_{(14348)}$ = -2.491, P = 0.013) and no effect in the high training condition ($t_{(14348)}$ = -0.106, P = 0.915). For PC2, there was no effect of PC score in the low training condition ($t_{(14348)}$ = 1.677, P = 0.094), but there was a positive effect for the high training condition ($t_{(14348)}$ = 2.971, P = 0.003). In **a** and **b**, circles represent effects of PC1 and diamonds represent effects of PC2. For visualization purposes, PC scores are sorted and divided into five bins, as in **Fig. 2c**. Each circle or diamond marker represents the mean similarity score within a given similarity bin, and the accompanying error bar indicates S.E.M. Lines of best fit—established from linear mixed-effects models—are shown for reference, with shaded areas representing S.E.M. Asterisks above plots reflect the main effect of PC scores obtained from linear mixed-effects models (*P < 0.05, ***P < 0.001).



Extended Data Fig. 3 | Effects of memory confusability and perceptual similarity on CA3/DG similarity scores. For the two human-based measures of similarity (memory confusability and perceptual similarity judgments), we tested for relationships with similarity scores in CA3/DG. Given the relevance of CA3/DG to memory, we hypothesized that CA3/DG similarity scores would specifically be sensitive to memory confusability. We ran a linear mixed-effects model with pairwise CA3/DG similarity scores (552 pairs per participant) as the dependent variable. The predictors (fixed effects) included memory confusability, perceptual similarity, their interaction with training (low vs. high), and the identity of the high-training category (beaches vs. gazebos). There was a significant negative main effect of memory confusability ($t_{(28692)}$= -2.157, P = 0.031), with a qualitatively positive main effect of perceptual similarity ($t_{(28692)}$ = 1.390, P = 0.165). There was no interaction with training with either measure of similarity ($t_{(28692)}$'s < 0.427, Ps > 0.669). Thus, CA3/DG inverted similarity as defined by memory confusability, but not as defined by perceptual judgments. The fact that memory confusability did not interact with training level (which contrasts with effects for PC1 and PC2 shown in **Fig. 4c**) is presumably because the memory confusability metric captured dimensions of similarity that were relevant across training levels. Indeed, memory confusability positively loaded on both PC1 and PC2 (see **Fig. 1i** and **Extended Data Table 1**). In the figure, circles represent memory confusability effects and diamonds represent perceptual similarity effects. For both metrics, similarity was divided into five bins, solely for visualization purposes. Each circle or diamond marker represents the mean similarity score within a given similarity bin, and the accompanying error bar indicates S.E.M. Lines of best fit—established from linear mixed-effects models—are shown for reference, with shaded areas representing the S.E.M.



Extended Data Fig. 4 | Transformation of representational structure across training levels. Here, we tested the extent to which the representational structure associated with high training preserved vs. transformed the representational structure associated with low training, independent of any similarity metric or dimension (see Methods – 'Relationships between experience, similarity dimensions, and the hippocampus'). For each ROI, we ran linear mixed-effects models for which the dependent measure was each participant's fMRI similarity scores for the high-training category (276 pairwise values). The key predictor (fixed effect) was normative low-training fMRI similarity scores. The normative low-training fMRI similarity scores reflected the mean similarity score for each scene pair across all *other* participants for which that scene category was associated with low training. The identity of the high-training category (beaches vs gazebos) was included as a fixed effect. The relationship between normative low-training similarity scores and (participant specific) high-training similarity scores was negative for CA3/DG ($t_{(14348)}$= -1.970, P = 0.049), but strongly positive for EVC ($t_{(14348)}$= 10.408, P < 0.001). Effects were not significant for CA1 ($t_{(14348)}$= -1.468, P = 0.142) or PPA ($t_{(14348)}$= 0.899, P = 0.369). Thus, whereas EVC robustly *preserved* within-category representational structure across stages of learning, CA3/DG *transformed* representational structure across stages of learning. In the figure, normative low-training fMRI similarity scores are divided into five bins, solely for visualization purposes. Each circle marker represents the mean similarity score within a given similarity bin, and the accompanying error bar indicates S.E.M.. Lines of best fit—established from linear mixed-effects models—are shown for reference, with shaded areas representing the S.E.M.

	Main effect of PC score		PC score x training interaction	
	$t_{(28681)}$	P	$t_{(28681)}$	P
PC1	0.001	0.999	5.761	< 0.001
PC2	-1.025	0.305	-4.696	< 0.001
PC3	0.950	0.342	1.663	0.096
PC4	-1.692	0.091	-0.174	0.862
PC5	0.217	0.828	1.640	0.101
PC6	-2.099	0.036	0.052	0.958
PC7	-0.699	0.484	-1.753	0.080
PC8	0.241	0.809	0.289	0.773
PC9	-1.012	0.312	0.644	0.520
PC10	0.833	0.405	-0.609	0.543

Extended Data Table 2 | Relationships between visual similarity dimensions, training level, and similarity scores in CA3/DG. Whereas the linear mixed effect model related to **Fig. 4c** included two dimensions of visual similarity (PC1, PC2) and training level (low vs. high), here we ran an expanded model that included all 10 dimensions of similarity (PC1 – PC10 scores). Each row in the table shows the main effect of PC score (for each similarity dimension) as well as the interaction between PC scores (for each similarity dimension) and training level. As in **Fig. 4c**, the model included a fixed effect representing the identity of the high-training category (beaches vs. gazebos). To correct for multiple comparisons (separately for main effects and interactions), we used a Bonferroni-adjusted significance threshold of $\alpha = 0.005$. At this threshold, the only significant effects were interactions between PC score and training level for PC1 and PC2.