A compressed code for memory discrimination

Dale Zhou[1,2,3,*], Sharon M. Noh[3], Nora C. Harhen[4], Nidhi V. Banavar[5], C. Brock Kirwan[6], Michael A. Yassa[1,2], and Aaron M. Bornstein[2,3]

[1]University of California, Irvine, Neurobiology and Behavior, 519 Biological Sciences Quad, Irvine, 92697, United States

[2]University of California, Irvine, Center for the Neurobiology of Learning and Memory, Qureshey Research Laboratory, Irvine, 92697, United States

[3]University of California, Irvine, Department of Cognitive Sciences, Social Science Lab 334, Irvine, 92697, United States

[4]New York University, Department of Psychology, 6 Washington Place, New York, NY 10003, United States

[5]University of California, Berkeley, Department of Political Science, 210 Social Sciences Building, Berkeley, CA 94720, United States

[6]University of Pennsylvania, MindCORE Neuroimaging Facility, 3401 Grays Ferry Ave, Philadelphia, 19146, United States

[*]Corresponding author: dale.zhou@uci.edu

October 12, 2025

Abstract

The ability to discriminate similar visual stimuli has been used as an important index of memory function. This ability is widely thought to be supported by expanding the dimensionality of relevant neural codes, such that neural representations for the similar stimuli are maximally distinct, or "separated." An alternative hypothesis is that discrimination is supported by lossy compression of visual inputs, efficiently coding sensory information by discarding seemingly irrelevant details. A benefit of compression, relative to expansion, is that it allows the individual to efficiently retain fewer essential dimensions underlying stimulus variation—a process linked to higher-order visual processing—without hindering discrimination. Under the compression hypothesis, pattern separation is facilitated when more information from similar stimuli can be discarded, rather than preserving more information about distinct stimulus dimensions. We test the compression versus expansion hypotheses by predicting performance on the canonical mnemonic similarity task. First, we train neural networks to compress perceptual and semantic factors of stimuli, and measure lossiness of those representations using the mathematical framework underlying compression. Consistent with the compression hypothesis, and not the expansion hypothesis, we find that greater lossiness predicts the ease and performance of lure discrimination, particularly in later layers of convolutional neural networks shown to predict brain activity in the higher-order visual stream. We then empirically confirm these predictions across two sets of images, four behavioral datasets, and alternative metrics of lossiness. Finally, using task fMRI data, we identify signatures of lossy compression—neural dimensionality reduction and information loss—in the higher-order visual stream regions V4 and IT as well as hippocampal subregions dentate gyrus/CA3 and CA1 associated with lure discrimination performance. These results suggest lossy compression may support mnemonic discrimination behavior by discarding redundant and overlapping information.

Keywords: memory reconstruction; efficient coding; false memory; rate-distortion theory; novelty detection

1 Introduction

Many behaviors, from value-based decisions [1, 2, 3, 4] and associative learning [5] to perceptual inference and memory [6, 7, 8, 9, 10], require recognizing whether the perception of a current situation is familiar or novel. This process is challenging because memory is constructive. From a percept with partial information, memory integrates prior experiences to fill in gaps but, in doing so, introduces distortions that can create incorrect impressions of prior experience [11, 12]. Discriminating whether the current situation is novel compared to remembered experiences depends on pattern separation, the ability to distinguish between

highly similar inputs with distinct responses [13, 14]. Pattern separation occurs when brain regions, such as the hippocampus, transform similar inputs that would produce aligned activity patterns into output patterns that are more distinct **(Figure 1A-C)** [15, 16, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Distinctness is often measured by the degree of linear independence where maximally distinct patterns are orthogonal. Although successful behavioral pattern separation is thought to be supported by reducing redundant overlap and keeping distinct details between inputs [26, 27, 28], it is not fully understood how computations support orthogonalization nor what input properties are orthogonalized [29].

The longstanding Marr-Albus hypothesis suggests two strategies to orthogonalize representations: expand the encoding ensemble and sparsen encoding activity [30, 31]. Both expanding the number of encoding units (neurons) using divergent projections from a small to a large population (a ratio of about 1:5 in the hippocampus) and inhibiting the population activity to have few active neurons within a relevant timespan (around 5% of neurons active) can decorrelate the statistical structure of inputs to separate across distinct ensembles of neurons [30, 31, 18, 16, 32, 27, 33, 34, 35, 5, 36]. However, expansion, sparsity, and decorrelation can have varying effects on pattern separation depending on the task and type of input, and it is difficult to disentangle their varying effects due to shared biological bases [37, 5].

A less explored hypothesis for pattern separation is lossy compression, a computation that discards redundant information to produce efficient representations with a tolerable level of error [38, 39, 40]. In contrast to the expansion strategy, compression suggests roles for reducing an encoding ensemble to create a physical or information bottleneck that encourages orthogonal representations of distinct features that dominate variation across inputs [41, 42]. In contrast to the sparsity strategy, compression benefits from different (sometimes lower) levels of sparsity due to better flexibility and expressivity with denser, mixed codes [43]. Excessive sparsity can encumber code diversity while reducing sparsity can avoid oversensitive responses to inconsequential variations in the input [44, 45, 46, 47, 48, 49].

The computational kinship between pattern separation and lossy compression can be made more apparent mathematically, distinguishing this account from alternative hypotheses such as sparsity **(Figure 1D-E)**. Pattern separation has been proposed to be the computation that decreases an arbitrary similarity metric, $S(X_1, X_2)$, of the degree of overlap between given inputs X_1 and X_2 [29]. Pattern separation occurs when $S(X_1, X_2)$ decreases across neural regions or neuronal populations A, B, and C, such that X_1 and X_2 are progressively decorrelated:

$$S_A(X_1, X_2) > S_B(X_1, X_2) > S_C(X_1, X_2). \tag{1}$$

Although S is commonly conceived as a linear correlation, S can also be defined as the mutual information I to capture both linear and nonlinear dependencies between inputs. Under this definition of S, Equation 1 directly parallels the data processing inequality [50]:

$$I_A(X_1; X_2) \geq I_B(X_1; X_2) \geq I_C(X_1; X_2), \tag{2}$$

stating that physical processing from $A \rightarrow B \rightarrow C$ cannot create new information about the original source. The inequality points to a trade-off where information is either retained at some cost or lost for more lightweight but error-prone transmission, a core computational problem of memory. Memory needs to reconstruct arbitrary traces yet cannot preserve all of the information and structure of inputs [26].

How many bits of information should be allocated to more precise high-fidelity memory versus saved for more approximate gist memory [51, 52]? Lossy compression provides a framework to determine the optimal solution to this trade-off. Optimally, lossy compression is the joint minimization of (1) the *rate* of information R needed to encode an input X_1 as a compressed representation X_2 and (2) the amount of *distortion* D caused by information lost from compressing X_1 into X_2 [53]:

$$R(D) = \min I(X_1; X_2) \text{ subject to } d(X_1, X_2) \leq D. \tag{3}$$

Rate-distortion theory shows how lossy compression forces approximations such that $D > 0$. We propose that reducing $S(X_1; X_2)$ for pattern separation involves minimizing existing redundancy, $I(X_1; X_2)$, which is marked by detectable increases in distortion $d_A(X_1, X_2) < d_B(X_1, X_2) < d_C(X_1, X_2)$.

Here, we test if lossy compression can explain performance in a behavioral task designed to tax pattern separation, the Mnemonic Similarity Task (MST) [25]. Participants incidentally encode information from a single exposure to a "target" image of an everyday object, then discriminate that memory from a similar yet distinct "lure" image and a novel "foil" image in a surprise discrimination test **(Figure 1)**. Discrimination performance, measured by the proportion of correctly identified lures relative to foils, is thought to assess detail knowledge or specific recollection [27, 54, 55]. We analyzed performance on 1,152 target-lure pairs across five previously published datasets [56, 57, 9, 58, 59]: a cross-sectional university sample ($n = 208$), a longitudinal university sample that performed lure discrimination immediately and after 1 week ($n = 78$), a cross-sectional sample of youths ($n = 92$, ages 8 to 25, average of 15.80 ± 5.12 years), a cross-sectional lifespan aging sample ($n = 297$, ages 18 to 86, average of 47.41 ± 19.61 years), and a cross-sectional sample who underwent fMRI scanning while performing the task ($n = 48$, 22.9 ± 3.6 years old). We focus on two pairs

of trials—corresponding target and lure trials, as well as first presentation and repeat trials. Using these data, we investigated how lossy compression contributes to the orthogonalization of target and lure images in support of pattern separation. Because visual information reaches the hippocampus after processing by visual and semantic cortical pathways [14], we extracted various perceptual and/or semantic features by processing images through neural networks trained for pixel reconstruction (perceptual), image classification (perceptual and semantic), or image-to-text conversion (semantic) [41, 60, 61, 62, 63]. These models have been predictive of the neural activity of temporal and visual cortices [64, 65, 66], which may help separate inputs by transforming the dimensionality of representations [67, 48, 68]. The lossiness of compression is operationalized by several convergent approaches. In behavioral data, we use an information-theoretic algorithm that we modify to estimate lossiness from a cosine similarity metric of orthogonalization [38] and auto-associative networks that measure lossiness as item reconstruction errors [41, 69, 51, 70]. In neural data, we use dimensionality reduction and the information rate (mutual information) of the evoked neural representations for targets and lures as operationalizations of lossy compression [39, 71].

To accomplish pattern separation, the expansion hypothesis proposes that a sharpened representation of the total information reduces overlap, whereas the compression hypothesis posits that a blurred representation discards overlapping information. Across images, we test whether lossiness explains why pattern separation is sometimes more difficult, as previously defined by binning performance in an independent sample [72] **(Figure 1B)**. Across individuals, we test if lossiness explains pattern separation performance, measured as the lure discrimination index: the proportion of correct rejection minus a response bias for "similar." We hypothesize that pattern separation is more difficult and poorer when more bits of information are needed to preserve high-fidelity information about subtle differences in detail **(Fig 1C)**. Conversely, if the targets and lures are more dissimilar, then pattern separation is easier and better because lossy compression can aggressively discard more bits for greater efficiency. However, aggressive compression that discards more bits per unit of lossiness increases false alarms by blurring together gist-like memories which are more susceptible to noise (**Figure 1D)**. Finally, we examine whether stimulus-evoked responses in a putative hierarchy of neural regions reflects continually increasing or decreasing dimensionality, in line with the expansion or compression hypotheses. We find evidence in support of the lossy compression hypothesis, replicated across image datasets, image features, compression methods, and participant datasets. Lossier gist-like features capturing higher-level perceptual features were more strongly related to pattern separation than low-level perceptual features, linking lossy compression with theories of object detection in the ventral visual stream by learning the most essential and invariant features in a lower-dimensional space [73, 74, 75, 48, 76]. Together,

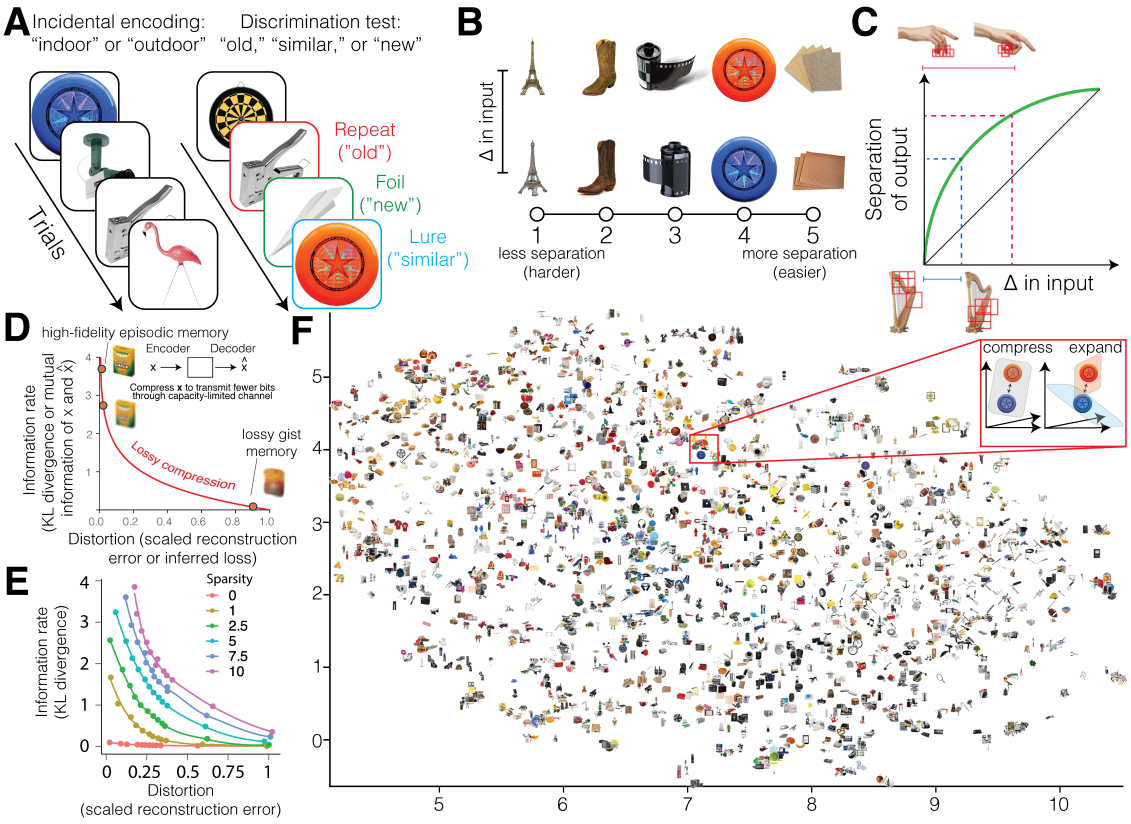

Figure 1: **Task schematic and lossy compression function.** **(A)** Task stimuli during the encoding and test phases with corresponding trial type and correct response. **(B)** Ease of pattern separation between target and lure images. **(C)** Pattern separation function where differences in the input are encoded as distinct outputs. Distinguishing between extremely similar harps may require encoding of more information (blue-dashed lines) than distinguishing between more dissimilar hand gestures (red-dashed lines). Red bounding boxes indicate patches of maximal importance for discrimination by AlexNet. **(D)** Putative lossy compression computation supporting pattern separation by discarding similarities in inputs according to a rate-distortion function, the mathematical basis of compression. A high-fidelity memory might be associated with the semantic representation like: "a new yellow and green box of crayon chalk in white". A moderately compressed memory loses some detail: "a new yellow and green box of crayons." Finally, a very aggressively compressed memory loses essential details, resulting in false memory: "a used, crumpled yellow and red bag." **(E)** Key differences between the sparsity and compression hypotheses can be seen by increasing sparsity constraints on rate-distortion functions generated by β-variational autoencoders that are tasked with reconstructing each image. Data points are averaged across images per β. Sparsity makes lossy compression worse because it increases the information cost as well as the distortion by enforcing usage of a restricted set of encoding units. **(F)** Different neural networks lossily compress images into different feature representations. Here a UMAP representation space is visualized for images embedded by their semantic representation from a vision-text transformer. *Inset*: How does lossily compressed dimensionality reduction (similar images represented in one plane) versus dimensionality expansion (similar images represented in more orthogonal planes) affect their memory discriminability?

these findings support the idea that lossy compression, rather than expansion, supports pattern separation.

2 Results

Lossy compression relates to easier and better mnemonic pattern separation

Can the lossiness of compressing inputs explain mnemonic pattern separation performance? We calculate the lossiness from the perceptual and semantic feature representations of the stimuli **(Figure 2A)**. Of particular interest are the pairs of stimuli evoking (1) a single-exposure memory of a target image during the study phase and (2) the subsequent exposure to a similar lure image during the test phase. Each neural network learns representations as a point in an internal model. Traversing the space of this internal model is a process of memory retrieval and reconstruction, where similar images are encoded more closely together according to the features learned by each neural network. A simple implementation of this traversal is a linear interpolation between points in the internal model. Using prior information theoretic methods on the generalization and discriminability of perceptual stimuli, we infer lossiness from a confusion matrix constructed using the cosine distances between the points (targets, lures, and intermediate representations) of the retrieval and reconstruction process [38]. Lossiness values are similar for similar neural networks (e.g. AlexNet and VGG-16, Spearman's $\rho = 0.73, p < 0.001$) and uncorrelated across perceptual models and other generative models such as the OpenCLIP image-to-text transformer model ($\rho < 0.07, p > 0.09$), suggesting good coverage of convergent and divergent representations of differing sensory and semantic features **(Figure 2B)**.

Consistent with the compression hypotheses, we find that target and lure images that are easier to pattern separate tend to be those where compressing targets into lures has greater lossiness (all p-values Bonferroni corrected) **Figure 2C**). This effect reproduces across two image sets and all perceptual ($0.25 < \rho < 0.40$, $p < 0.001$) and semantic models ($0.09 < \rho < 0.19$, $p < 0.04$) that compress targets into lures, but not with the β-VAE models trained to reconstruct individual items ($\rho < 0.07, p = 1$). This suggests pattern separation is more related to the retrieval of compressed information about target and lure pairs, rather than efficient reconstruction of items in the pair themselves. We find further evidence consistent with the compression hypothesis, such that individuals tended to perform better when stimuli were more lossily compressed (linear regression $0.02 < \beta < 0.38$ Bonferroni-corrected $p < 0.006$, **Figure 2D**). This effect largely replicates in 21 out of 24 tests across four datasets and the perceptual and semantic models. Next, we assess if age moderates the relationship between lossiness and lure discrimination performance. To this end,

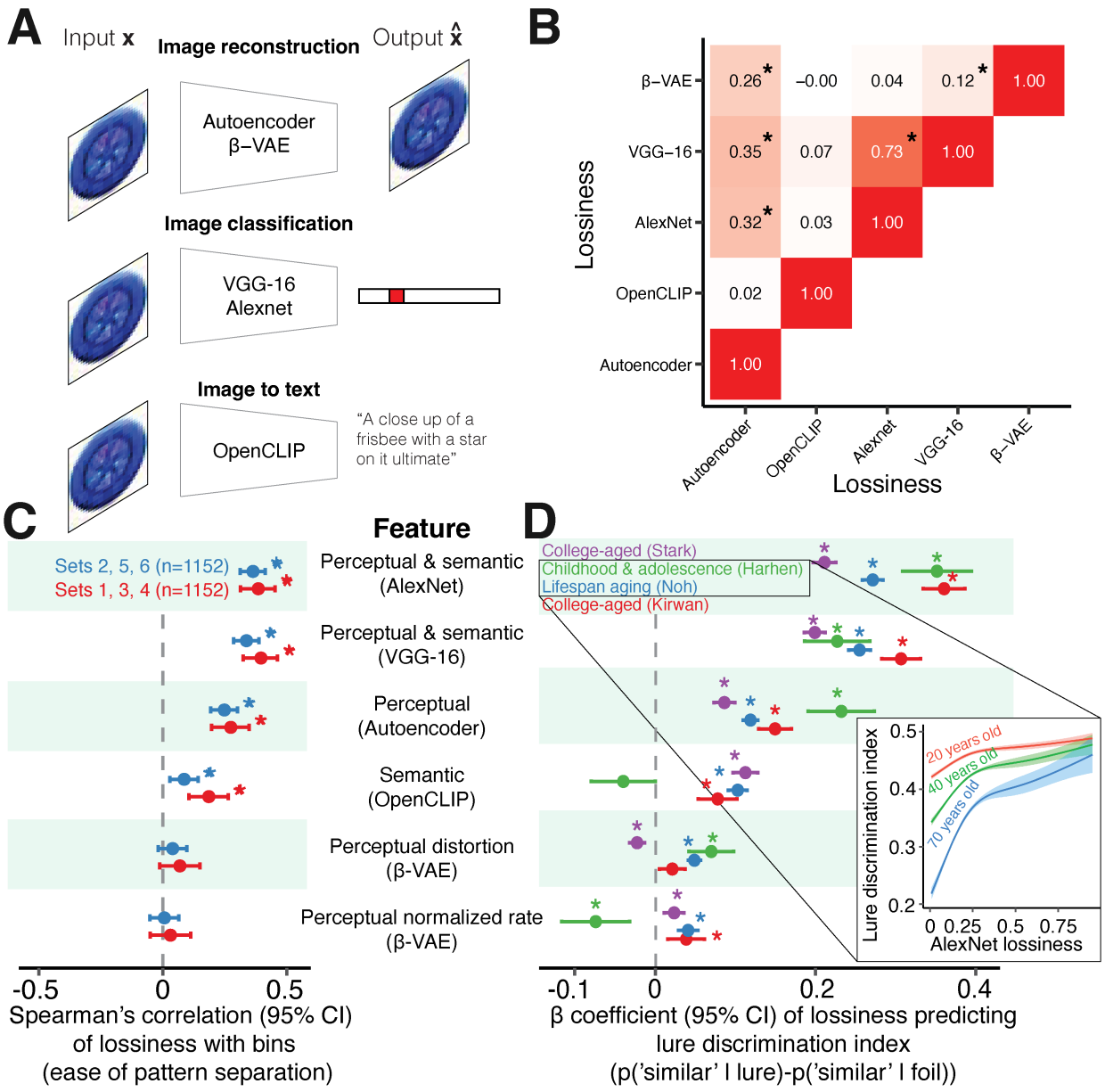

Figure 2: **Lossiness of perceptual and semantic compression predicts pattern separation ease and performance. (A)** Neural networks labeled by their respective tasks. We applied each network to all images x in order to obtain the sensory or semantic feature vectors $\hat{x}$ used to perform those tasks. **(B)** We measure the lossiness of compressing the $\hat{x}$ of a lure to the $\hat{x}$ of a target image by applying an information theoretic algorithm to simulated confusion matrices based on the cosine distance between $\hat{x}_{lure}$ and $\hat{x}_{target}$. The correlation table shows similarities and differences between the lossiness across models. **(C)** Consistent with our hypothesis, greater lossiness is correlated with easier pattern separation across two separate image sets. This effect was strongest for perceptual and sensory features. We did not observe an effect of lossiness using the β-VAE model. **(D)** Consistent with our hypotheses, greater lossiness is correlated with better pattern separation performance across datasets. The effect was greatest for perceptual and semantic features. The effect was less consistent for only semantic features and for the β-VAE model. Inset: nonparametric generalized additive modeling was used to flexibly model how lossiness and pattern separation interact with age. Individuals better pattern separate images with greater lossiness across the lifespan. The steeper logarithmic form for older adults ($F = 732.1, p < 0.001$, lossiness-by-age interactions $p < 0.001$) suggests that lossiness may especially help reduce performance gaps between older compared to younger adults.

we used generalized additive models with penalized splines, a method which allows for statistically rigorous modeling of linear and nonlinear effects while minimizing over-fitting [77]. Lure discrimination performance by older adults was more strongly related to the lossiness of compression than younger participants ($F = 732.1, p < 0.001$, lossiness-by-age interactions $p < 0.001$), consistent with the notion that older adults can use semantic memory to compensate for degradations in episodic memory as their semantic knowledge increases [78, 79]. A small amount of lossiness can contribute significantly to improvement, but its benefits exhibit diminishing returns. In support of the compression hypothesis, lossiness predicts pattern separation ease and performance. Perceptual and semantic features were most related to the ease and performance of pattern separation.

Aggressive compression creates gist-like memories related to increased false alarms

Discarding information can enhance pattern separation by strategically targeting particular features of the image. Optimizing a neural network to retain essential features to remain distinguishable while compressing away non-critical correlations has been called a kind of "optimal forgetting" and adaptive distortion [51, 80]. While we already found some evidence for this hypothesis in pattern separation ease and performance, here we investigate whether the hypothesis also explains errors in performance. Errors in pattern separation are calculated by the lure false alarm rate, wherein individuals mistakenly confuse a lure for a previously studied image. We were also interested in how the different compressibility of unique images could explain how errors occur when there is memory interference introduced by a delayed test. Increasing lossiness corresponds to differing levels of information discarded according to rate-distortion curves unique to each image. We quantify this difference by calculating the normalized rate per image, or the slope of the rate-distortion function defined as the amount of information discarded per unit of distortion **(Figure 3)A**. More negative, steeper slopes characterize forgetting that discards more information and resembles a more aggressive compression. More positive, shallower slopes characterize forgetting that retains more information and resembles a more conservative compression. While we test all datasets, we were particularly interested in the case of the longitudinal dataset where a substantial amount of interference and forgetting occurs after 1-week delayed test **(Figure 3)B**. While the trial difficulty no longer predicts the ease of pattern separation after 1 week, the normalized rate was associated with performance in both the immediate and delayed tests. Individuals tended to have worse pattern separation performance on images with more aggressive compression. Consistent with the compression hypothesis, greater lossiness correlated with lower lure false alarms **(Figure 3)C**. This relationship largely replicated in 20 out of 24 tests across 4 datasets and the

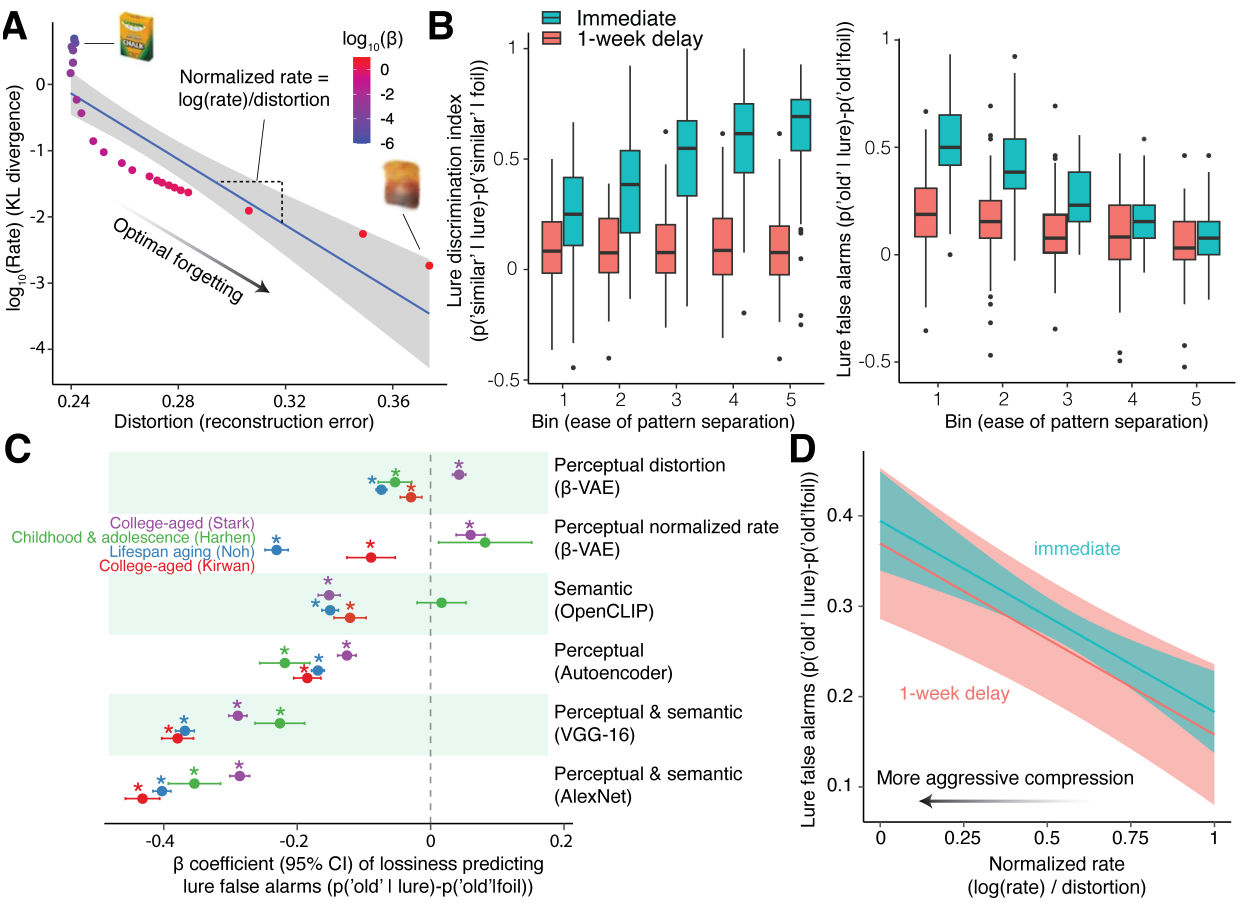

Figure 3: **Gist-like memories due to aggressive compression explains increased false alarms. (A)** Each image has lossy compression applied such that the information rate and the reconstruction error between an input and output are optimized according to different rate-distortion trade-offs, β, using β-VAEs. Discarding information in a principled manner according to the information theory underlying lossy compression may characterizes a process of optimal forgetting by introducing adaptive distortions, but overly aggressive compression can lead to false memory due to gist-like representations (e.g., "a new yellow and green box of crayon chalk in white" versus "a used, crumpled yellow and red bag"). We test the benefits and drawbacks of distortion and the loss of information on lure false alarms. The optimal forgetting function can be quantified by calculating the slope of the rate-distortion function, which differs across images. We refer to this slope as the normalized rate because it is the amount of information discarded normalized by distortion for lossy compression. Steeper slopes indicate more aggressive compression that discards more information per unit of distortion, while shallower slopes indicate more conservative compression that preserves more information per unit of distortion. **(B)** *Left:* Pattern separation performance degrades after 1 week. Pattern separation performance is better on easier stimuli in the immediate test but ease is not predictive of performance after 1 week. *Right:* Lure false alarms occur more on harder stimuli in the immediate test but difficulty is not predictive of performance after 1 week. False alarms decreased after 1 week for harder stimuli. **(C)** Consistent with our hypothesis, greater lossiness across people is related to reduced lure false alarms, consistent with the idea that redundancy reduction decreases overlap across an individual's memories. **(D)** Consistent with our hypothesis, people viewing images with a lower normalized rate tended to have more lure false alarms in both the immediate and delayed tests. More aggressive compression discards more information for compact gist-like memory representations at the cost of a greater risk of false memory.

perceptual and semantic features tested. However, the effect of the normalized rate was less consistent, suggesting that lossiness is a more robust metric.

In the longitudinal dataset, the ease of separating stimuli no longer predicts performance nor false alarms after a 1-week delayed test. Yet, a lower normalized rate was associated with more lure false alarms in both the immediate and delayed session ($\beta = -0.21, p < 0.001$), indicating that aggressive compression can give rise to efficient but false gist-like memories. "Optimally forgetting" information according to lossy compression appears to improve the overall discriminability of memories by reducing redundancies across representations; at the same time, aggressive compression can also generate false memories.

Lossily compressed high-level perceptual and semantic representations associated with easier and better mnemonic pattern separation

We next investigated which kinds of image features were most related to pattern separation ease and performance, comparing low-level features about spatial detail to high-level features about semantic abstraction. More compression of all perceptual features related to easier pattern separation ($0.26 < \rho < 0.40, p < 0.001$). Higher-order semantic features (representations in deeper layers) had greater effects on the ease of pattern separation performance across image sets than lower-level features about spatial details (representations in shallower layers; $r = 0.97, p < 0.001$; **Figure 4A**). Moreover, we find a similar relationship when analyzing pattern separation performance ($0.08 < \beta < 0.83, p < 0.001$ **Figure 4B**). Individuals tended to have better performance when either the deepest or most shallow features were more lossily compressed.

Whether the deepest or most shallow features are more important may be driven by age. In the dataset of our youngest participants including children and adolescents, performance was more strongly related to low-level sensory details than higher-level features, as they may have not yet formed the richer semantic memories of older adults. To further test whether the order of importance of feature layers is explained by age, we conducted an analysis controlling for age as a covariate. After adjusting for age, the predictive strength of high-level features again surpassed that of low-level features **(4B inset)**, indicating that age may moderate the observed shift from reliance on perceptual to semantic processing ($0.21 < \beta < 0.32, p < 0.001$). This pattern of results suggests that lossily compressing high-level semantic features supports pattern separation, perhaps characterizing how the ventral visual stream discards inessential information to detect objects [73, 74, 75, 48, 76].

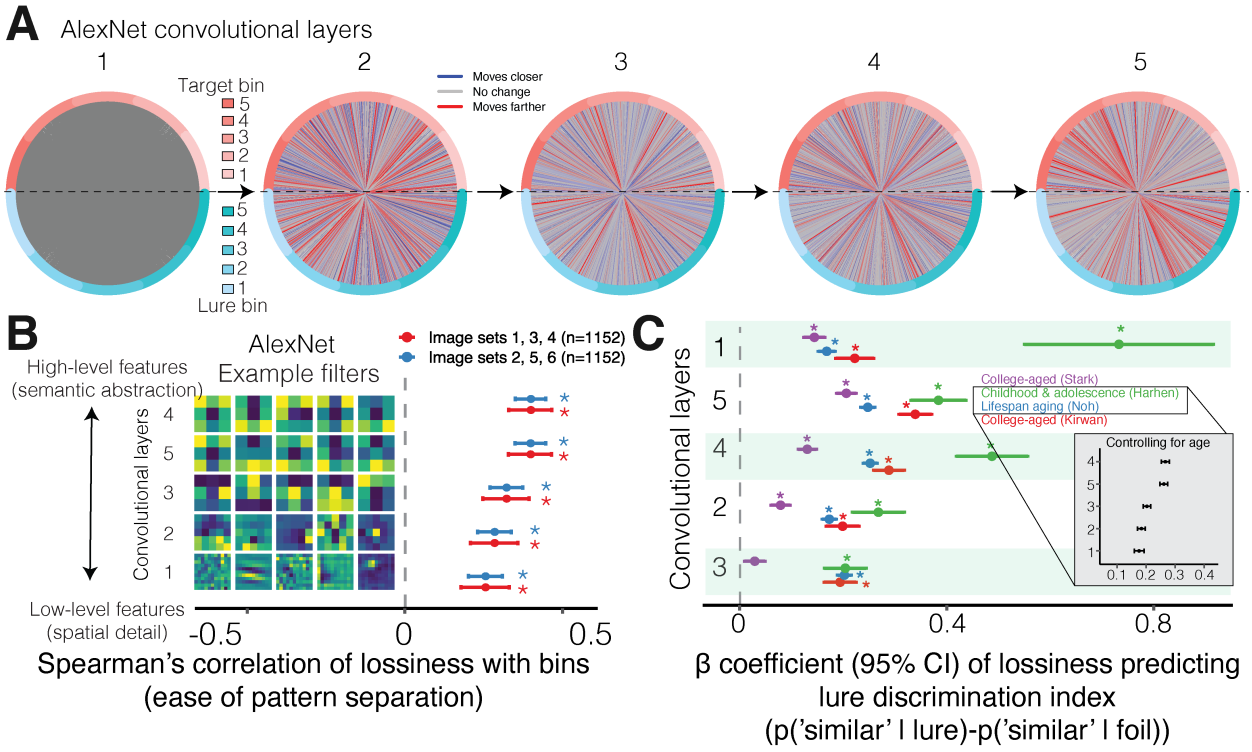

Figure 4: **Lossiness of high-level semantic abstractions are more important than low-level spatial details for pattern separation ease and performance. (A)** What happens to the distance between target-lure pairs as they are processed through progressively deeper layers of AlexNet? Targets (pink) and lures (cyan) are shown as nodes, shaded by their ease of separation (bin). Edges indicate the relative change in Euclidean distance between target–lure pairs, based on dimensionality-reduced representations at each layer compared to the preceding one. Visually, earlier layers, known to process spatial details and orientation, appear to separate more difficult stimuli; later layers, known to process more semantically abstract information, appear to separate easier stimuli. **(B)** Left inset: AlexNet contains a hierarchy of convolutional layers that process different features of an image, including low-level perceptual features that represent edges, shapes, and textures (fine-grained filter); high-level perceptual features that represent conceptual and semantic abstractions for object categorization; and semantic features that represent verbal descriptions of images (coarse-grained filters). Right: Compressions of targets into lures features that incur greater lossiness tended to be better pattern separated. The lossiness of higher-level features characterizing semantic abstraction have a stronger effect on the ease of pattern separation than low-level features characterizing spatial detail. **(C)** Individual differences in pattern separation performance is most strongly predicted by the lowest-level layers for children and adolescents and highest-level layers of processing for older participants. Layers are visualized in descending order by average effect size. Inset: controlling for age results in the higher-level layers being more predictive of performance than lower-level layers.

Neural signatures of lossy compression in higher-order visual stream and hippocampus associated with better pattern separation

Lastly, in light of the computational and behavioral evidence associating lossy compression and pattern separation, we investigate neural signatures of lossy compression for separating evoked representations of target and lure stimuli **(Figure 5)**. Two signatures were of interest: the dimensionality of the representations for target and lure stimuli pairs and the mutual information between the representations. While dimensionality expansion has been proposed to support pattern separation and object discrimination by increasing the neural separability of distinct clusters of information [5, 76], the lossy compression framework predicts the opposite whereby dimensionality reduction helps to efficiently retain only a few of the most separable dimensions. Evidence supporting the compression hypothesis involves reduced dimensionality for correct trials (lure correct rejections) compared to incorrect trials (lure false alarms) that correlates with behavioral pattern separation performance. However, reduced dimensionality does not necessarily mean that there is less total information but rather that information is clustered around fewer dimensions. The compression hypothesis further predicts that the lower-dimensional representations exhibit a loss of information in correct trials compared to incorrect trials, a lower mutual information characteristic of lossiness. We tested 9 regions of interest, including V1, V2, V3, V4, IT, anterolateral entorhinal cortex, DG/CA3, CA1, and subiculum, as well as 2 reference regions that were not hypothesized to be directly involved in perceptual compression (primary somatosensory cortex and primary motor cortex). All comparisons are reported using FDR correction.

A circuit that performs pattern separation is proposed to exhibit a progressively stronger dissimilarity signature, such that input stimuli diverge more distinctly as population activity is measured across regions [29]. Indeed, the dimensionality reduction for correct versus incorrect trials progressively strengthens from the ventral visual stream to the hippocampal circuit ($\rho = -0.31, p < 0.001$; **Figure 5B**). Consistent with a compression hypothesis, we find that better pattern separation performance was associated with dimensionality reduction in higher-order visual stream regions V4 ($\rho = -0.37, p = 0.043$; **Figure 5C-D**) and IT ($\rho = -0.36, p = 0.043$; **Figure 5E-F**) as well as hippocampal DG/CA3 ($\rho = -0.39, p = 0.043$) and CA1 ($\rho = -0.44, p = 0.031$). We did not find any correlations between dimensionality expansion and pattern separation performance, nor did we find any associations involving the reference regions ($\rho > -0.29, p > 0.09$). Bootstrap resampling with 10,000 iterations confirmed post-hoc that both M1 and S1 had significantly weaker correlations than the mean of other regions related to pattern separation: M1 vs. others = -0.10 [95% CI: 0.08, 0.13], $p < 0.001$; S1 vs. others = -0.17 [95% CI: 0.15, 0.21], $p < 0.001$). Together, the lossy compression

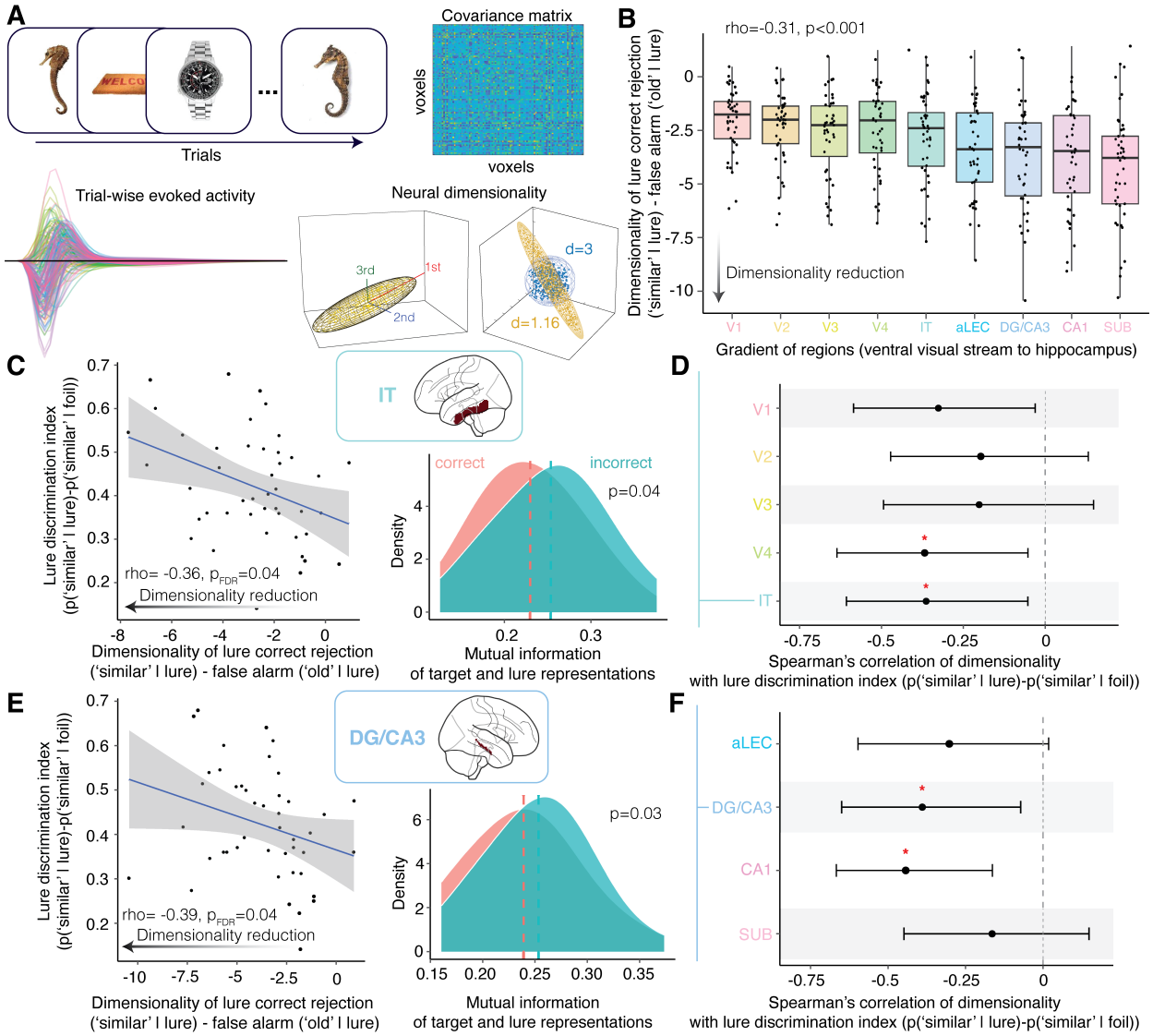

Figure 5: **Signatures of lossy compression across the higher-order visual stream and hippocampus associated with pattern separation performance. (A)** *Top-left*: Participants completed a continuous recognition paradigm, with no separate study and test phase for incidental encoding. *Bottom-left*: For each trial, GLMsingle was used to estimate voxel-wise evoked activity magnitudes (βs) and an optimized hemodynamic response function from a library of canonical forms to repoduce trial-wise evoked activity time series [81]. Data visualized from one participant. *Top-right*: The time series was used to calculate the voxel-wise covariance matrix. *Bottom-right*: The magnitudes of principal components are the eigenvectors of the covariance matrix. The neural dimensionality (participation ratio) quantifies how many principal components effectively contribute to the variance captured along each direction (eigenvalues). When some voxels dominate the variance, there is low dimensionality. When voxels act more independently, there is higher dimensionality. **(B)** There is a gradient of progressively stronger dimensionality reduction from the ventral visual stream to the hippocampus. **(C-D)** Dimensionality reduction in higher-order visual regions in the ventral visual stream processing objects predicts pattern separation performance. In IT, correct trials had lower mutual information between target and lure reprsentations than incorrect trials. **(E-F)** Dimensionality reduction in hippocampal DG/CA3 and CA1 predict pattern separation performance. In DG/CA3, correct trials had lower mutual information between target and lure reprsentations than incorrect trials.

of semantic abstractions and combinations of a few discriminable perceptual features in the ventral visual stream may support pattern separation in the hippocampus.

Discussion

In summary, we found behavioral, computational, and neural evidence supporting a lossy compression account of memory discrimination. The lossiness of compressing the semantic and perceptual input features explained the ease, performance, and errors in pattern separation. The relationship between lossiness and pattern separation was strongest when semantic abstractions were distilled from perceptual details in networks trained to perform image classification. This result is consistent with how the ventral visual stream first processes low-level spatial detail then high-level semantic abstractions to perform object detection [74, 75, 64]. Moreover, layer-wise dimensionality reduction—a form of lossy compression—improves the separability of inputs in intermediate layers of neural network models associated with the later regions in the ventral visual stream [48, 64, 65, 66]. Consistent with this computational result, we found that the dimensionality reduction of neural activity was associated with better pattern separation in high-level V4 and IT, but not low-level, visual stream regions for intermediate processing of abstract, semantic qualities of objects that are then later input into and further reduced within hippocampal DG/CA3 and CA1 to support pattern separation. The importance of high-level versus low-level information may be related to age. Young participants' performance was uniquely related to the lossiness of the lowest-level features, a relationship that was moderated with age. Pattern separation in children may be supported by the neurodevelopment of high-resolution memories for perceptual details [82, 83, 84], while loss of precision in mnemonic representations is related to episodic memory decline during aging [85]. Indeed, performance in our youngest sample including children most related to the compression of perceptual details. Our results suggest that perceptual inputs which are more amenable to lossy compression can produce representations with strategically placed imprecision that still retain some essential dimensions that support memory. Consistent with this idea, older adults' performance was closer to that of younger adults when the stimuli had lossier compression of perceptual and semantic features. These results are consistent with the notion that the development of semantic knowledge may help organize stimuli in memory, providing a scaffold for perceptual memory encoding [86, 87, 88, 89, 90, 91]. The lossy compression framework explains why perceptual details should be lost faster than higher-level semantic constructs for memory discrimination.

Due to optimizing how overlapping information is discarded by separating input features [92, 41, 42],

lossy compression can be viewed as a process of "optimal forgetting" for different computational goals [69, 93, 51]. The distortion caused by discarding information is not simply noise but is an adaptive distortion or error signal that can serve useful memory computations [94, 95, 38, 96, 97, 70, 51, 80, 98, 99, 52]. A variety of memory and learning phenomena can be understood under a shared framework by viewing memories as lossy compressions by the hippocampus [26, 37, 100, 65, 101], prefrontal, and perceptual processing streams which disentangle or repulse inputs to create the distances that support pattern separation [73, 37, 80, 39, 29, 14, 102, 103, 104, 105, 106, 107, 76, 57, 108, 109]. Indeed, the lossiness of compressing perceptually similar stimuli explains when those stimuli are more perceptually discriminable [38], as the sensory system can afford more aggressive compression without needing to invest additional resources to retain finer details with high precision.

These results can be understood more broadly in the context of the efficient coding principle. Efficient coding is when the brain maximizes the amount of information transmitted in an economical form by reducing redundancy [110, 111]. This coding strategy maximizes efficiency by allocating limited resources to where they are most needed for the task at hand [110]. For example, information is believed to be compressed even before the primary visual cortex, where 10^9 bits/second in the photoreceptors are compressed to 10^7 bits/second across retinal ganglion cells subject to the constraints of the capacity-limited optic nerve [112], and in a manner that is metabolically efficient [113, 114, 115, 40]. Arriving into the visual cortex, signals processed by neural network models of ventral visual stream computation suggest progressive dimensionality reduction of inputs [48]. However the efficient coding principle should not be read as always compressing or reducing, but adaptively switching between compression and expansion modes depends on the complexity and dimensionality of the task [46, 39, 116]. For instance, in our task we tested rapid, single-shot encoding from only one exposure to pairs of stimuli, whereas other tasks tax multiple repetitions and sessions across multi-dimensional sets of stimuli [117]. Indeed, compression and expansion computations both support learning because both help effectively orthogonalize representations dependent on the intrinsic dimensionality of the task and its inputs [69, 93, 118, 67, 68, 36, 71, 40, 119, 120, 121]. Hence, switching between differing levels of expansion and reduction may complement switching between different levels of inhibition and sparsity in support of adaptive memory function [44, 122].

In the context of memory and learning, compression and expansion processes can transform the dimensionality of the memory representations to support their separability or generalization [123]. Such neural processes in the IT and perirhinal cortex may separate representations by category, novelty, and familiarity [124, 125, 121]. This separability hinges on the transformation of representational space creating separa-

tion due to the change in dimensionality defining the location of representational points in the space [76, 5]. Novelty detection is thought to involve ventral tegmental and hippocampal loops that compare predictions generated by the memory traces in hippocampal regions with the new inputs from the cortex [126, 127]. A future experiment could build upon our methods by using supervised or self-supervised redundancy reduction methods based on an information bottleneck approach to compression [118, 128]. Selective attention to task-relevant factors shape perceptual feature representations in a similar manner as top-down modulation by hippocampal representations for memory and learning [129, 130, 39, 91, 131, 132]. In addition to reconstructed retrieval and top-down signals generated by the hippocampus and medial temporal region, regions earlier in the processing hierarchy in the visual and orbitofrontal cortex can support single-exposure perceptual memories that may be lossily compressed to support pattern separation [133, 134, 135]. Eye-tracking, information bottleneck, and neural recordings could help elucidate how early compressive computations can support the pathway of pattern separation processes. The efficient coding hypothesis predicts sampling and storage of highly relevant features of information. For example, future work could investigate relationships between optimal features to remember and the features contributing to the memorability of objects [136, 137, 138, 139, 89, 140]. Tantalizingly these approaches can provide a precise and general framework that explains how such salient features are extracted and later used if they are *relevant*, where relevance is defined in the information theoretic terms of how much information can efficiently retained and discarded according to task demands [128, 141, 142]. This relevance computation can be used to unify computational and neural processes linking memory, control [143, 1, 144, 145, 146, 57, 147], and category learning [148, 149, 150, 151].

While compression and expansion have complementary functions, compression may prove more useful when stimuli contain redundancy (already have intermediate or heavy overlap) and are intrinsically low-dimensional [37, 48, 76]. In our behavioral pattern separation task, key stimulus dimensions include orientation, color, brightness, detail, number, scale, and/or shape factors [91]. Compressed representations may support pattern separation in the dentate gyrus of the hippocampus for rapid, one-shot learning [65, 20, 152], whereas more prolonged processing is supported by activity in the ventral visual stream, determining the parts of the input that are invariant to changes in these dimensions, and other regions requiring larger changes in inputs across longer timespans [19, 153, 154, 73, 74, 14, 75, 103, 117, 57]. Prior work has already explored how regions contribute to pattern separation beyond the hippocampus according to complementary learning systems theory [14, 155, 22, 103, 152, 57, 156, 147]. What the compression framework provides is a cohesive mathematical account for when differing levels of expansion/reduction and sparsity can directly

and indirectly support pattern separation and related processes of general recognition, detailed recollection, and perceptual discrimination across the brain [29, 102, 105, 157, 108, 107, 57].

Our results can also be interpreted from a retrieval and pattern completion account, as memory discrimination tasks are not process pure [13]. Pattern completion is the reactivation of a stored memory trace in response to a partial cue [15]. Closely related to efficient coding and our compression methods is the analysis-by-synthesis theory, which posits that perception involves inferring a compact set of latent causes that can reconstruct the input [158, 159, 160]. Such inference parallels memory reconstruction, where the learned latent structure enables efficient recall. The reconstruction error or distortion from this process characterize the (im)precision of the memory representation which can serve as an uncertainty or confidence signal for learning and recollection [161], further shaping both perception and memory [98, 162].

Computationally, retrieval and completion perspectives offer a complementary view to compressed encoding by accounting for how recollection processes supports memory discrimination and novelty detection via a recall-to-reject process [163, 164, 27, 54, 165, 166, 167]. Neurally, later regions in the ventral visual and occipital temporal stream as well as hippocampal CA3 and CA1 are regions involved in pattern completion and retrieval processes [168, 169], overlapping with the regions we identified. Our framework, like complementary systems models [17, 170], involve both encoding and retrieval processes: inputs are compressed at encoding into a latent space, where latent variables are the "causes" of pixels in the stimuli. Traversing a distance in the latent space is akin to using the latent variables as a partial cue to reconstruct and recollect an associated stimulus (target-lure pair). What our framework provides is a normative model to determine the nature of what semantic or perceptual details should be encoded, retrieved, or discarded and at what level of precision or distortion given particular goals of a task. Future work could investigate these processes by quantifying the information cost of reconstruction given the partially distorted compression as a cue compared to attentional templates or memory schemas that represent goals, interacting with the memory and control processes of frontoparietal and association regions [143, 144, 145, 146, 57, 147, 171].

This work has several limitations, some of which motivate future research. First, lossiness was inferred from an information theoretic method developed to fit empirical data from perceptual identification experiments, such as on a range of tones, colors, line lengths, and shapes [38]. To adapt this method to our pattern separation task, we simulated confusion matrices on a range of memory representations according to a putative retrieval process that linearly interpolates over cosine distances between internal memory representations. A consequence is that, in contrast to prior work showing that lossy compression provides an alternative to the theories dependent on psychological distance representations such as multidimensional

scaling, our inferences of lossiness explicitly use a psychological space defined using cosine distance. This concern is partly addressed by using alternative variational methods for estimating lossiness. Nevertheless, the question remains: what does a compression framework add beyond the simpler cosine dissimilarity metric of the separability of representations? Compression provides a process model where the lossiness of compression is intrinsically linked to the orthogonalization (cosine dissimilarity) of representations, while also yielding novel testable predictions about compression that extend beyond simpler dissimilarity metrics of overlap. In future work, a diffusion process that stochastically samples from latent manifolds could ground this process in retrieval and memory reconstruction processes [172, 173, 59]. Moreover, approaches adapted from compression models of visual working memory could be explored [174, 175]. Second, it is unclear how general this framework is because our experiment focused on single-shot encoding of concrete, everyday objects. With a basis in efficient coding, we expect similar results with naturalistic images as they contain statistics that the brain is attuned to efficiently process via redundancy reduction [176, 177]. It would also be interesting to test pattern separation performance on sketches, drawings that focus on the gists of objects retrieved from memory. Noted as early as the origins of efficient coding theory [111], sketches drawn with a limited number of strokes have little information to compress, yet contain strokes that are essential for object recognition and should be retained while non-essential strokes should be discarded [178, 179]. Third, our analysis is limited by the temporal and spatial scale of fMRI data. For example, dentate gyrus and CA3 are not separable at this resolution. Furthermore, while we took statistical measures to limit the confounding effect of differing sizes of brain regions of interest on our metrics, measures of dimensionality remain affected by spatiotemporal scale, motivating the usage of scale-dependent measures [180]. Fourth and finally, all of our models encoded feature representations in fewer dimensions than the raw number of dimensions of the inputs (number and color of pixels) but the intrinsic dimensionality of the dataset is far lower due to redundancies among images and the predefined multi-dimensional variation between targets and lures. A stronger test of the compression and expansion hypotheses could involve determining the intrinsic dimensionality of our task's image dataset using similarity judgment tasks and computational approaches [48, 181, 68, 182]. The usage of information bottleneck methods can elucidate how learning and forgetting different parts of inputs can support pattern separation by compressing key factors of variation across a greater number of image repetitions or experimental sessions [183, 100, 184, 185].

3 Method

Experimental task

Mnemonic similarity task

Participants were instructed to judge object images as indoor or outdoor during 128 study trials (**Figure 1A**). This study phase is intended to have participants incidentally encode the stimuli in memory while performing the indoor and outdoor cover task. Next, during a test phase, participants judge stimuli as "new," "similar," or "old," when the presented stimulus was already seen, never seen, or slightly altered (repeat, foil, and lure trials, respectively) over 192 test trials. Several performance metrics can be calculated that index different memory processes. Mnemonic discrimination performance is measured as the lure discrimination index, or the proportion of correct rejections adjusted by a response bias: $p(\text{"similar"}|\text{lure}) - p(\text{"similar"}|\text{foil})$. The lure discrimination index in an independent sample was used to discretize the "mnemonic dissimilarity" of original and lure images into 5 bins of ease, where 5 is the easiest and most dissimilar and 1 is the hardest and most similar (**Figure 1B**). A recognition score is defined as correct detections corrected by the false alarm rate: $p(\text{"old"}|\text{repeat}) - p(\text{"old"}|\text{foil})$. We primarily use the recognition score for quality control purposes, as we focus here on pattern separation whereas this score centers pattern completion processes [186]. Finally, lure false alarms are defined as as the proportion of trials where lures were mistaken as a repeat corrected for a response bias: $p(\text{"old"}|\text{lure}) - p(\text{"old"}|\text{foil})$. This task allows us to probe pattern separation ease, performance, and errors.

Data and pre-processing

We analyzed four behavioral datasets. Two datasets covered a wide range of ages: a sample of participants across the lifespan (n =366; 46 ± 19 years old; 218 women, 144 men, 4 other) [56] and a childhood and adolescent sample (n =92; 16 ± 5 years old; 54 women, 37 men). The lifespan datasest was recruited from Amazon Mechanical Turk. Child data was recruited from Hartley Lab Participant database. Participants were recruited from the Hartley lab database which includes individuals recruited through ads on social media (e.g., Facebook and Instagram), word of mouth, local science fairs, and flyers on New York University's campus. Participants who had not previously completed an in-person study with the lab completed a brief Zoom call with a researcher. During this call, participants (and their parent or guardian, if the potential participant is under 18 years) were required to be on camera and confirm their full name and date of birth. Adult participants and parents of child and adolescent participants were additionally required to show photo

identification. We also assessed two datasets of undergraduates recruited in a university setting: a sample of undergraduate students who participated for course credit (n =208) [58] and a longitudinal sample (n =78) collected immediately and after 1 week. Participants in the longitudinal sample completed two study and two test phases, counterbalanced across two sessions. Half of the participants completed the test immediately after study, followed by the second study phase in the same session. They then took their second test one week later. The other half of the participants completed only the study in the first session and then completed the test one week after. They then completed the second study and second test during the second session.

We performed several quality control steps to arrive at our final sample. First, we performed trial-level quality control. If a response was faster than 300 milliseconds or slower than 3 seconds, then the trial was excluded. Next, we performed participant-level quality control. If the participant missed more than 20% of trials, the participant was excluded. Participants also were required to meet minimum scores on the lure discrimination index and recognition metrics indicating minimal engagement with the task. A minimum lure discrimination index of 0 was required; this minimum occurs when response bias was equal to correct rejections indicating chance-level performance. A minimum recognition score of 0.5 was required; this calculated as the correct detection minus a response bias, where chance performance is again 0. For the longitudinal dataset, we only applied the performance thresholds to exclude participants based on their immediate test.

fMRI mnemonic similarity task

Participants (n=48; 22.9 ± 3.6 years old, 27 female, 21 male) completed a continuous recognition version of the mnemonic discrimination task [187, 25]; see previous publications using this dataset for more details [57]. A series of images of everyday objects were shown in sequence. The image set differs and predates those from those published in the prior online implementations of the task, which were used in our behavioral data above [188]. Participants judged whether each image was new (foil trials), similar (lure trials), or old (repeat trials). The delay between the first and repeated presentation of lure or repeated objects varied with a mean lag of 19 trials. Six blocks of 107 stimuli (total of 642 images), were presented. Each block contained 32 first presentations, 16 targets, 16 similar lures, and 43 unrelated foils. Participants were required to respond within 3.0 s, after which time the stimulus was replaced by a blank screen with a fixation cross for .5 s followed by the next stimulus.

Neural network feature representations

All models were implemented using PyTorch. Autoencoder models were trained on 2,252 images across 6 image sets used in the mnemonic discrimination task. Images were originally 400x400 (x3 RGB color channels) and were downsampled to rescale according to smaller input dimensions required by the first layer of each neural network and to reduce computational costs of training. Images were normalized by the mean RGB value across all images to stabilize optimization and improves generalization. Image classification and image to text models were pre-trained on large datasets, described below. Given a stimulus from the mnemonic similarity task, the activations of the hidden layer in the autoencoders were used as the learned feature representation per image, while the activation of the convolutional or fully connected layers were used as feature representations or low-level to high-level perceptual and semantic features in the pre-trained models. The dimensionality of the layers are described in more detail below.

Image reconstruction

We use a convolutional autoencoder to learn compressed memory representations of color images in an under-complete latent space. Stimuli were rescaled to 32x32 (x3 RGB color channels). The encoder progressively reduces the input image to a compact latent representation using three convolutional layers with stride-based downsampling and GELU activations. The tensor produced by the last convolutional layer is inputted to a fully connected bottleneck layer flattening the tensor to a 256-dimensional latent space. The decoder uses transposed convolutions to upsample the latent representation back to the original spatial dimensions, with ReLU activations and a final Tanh activation to normalize the output. We used a reconstruction loss defined as the mean squared error and trained for 250 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 0.001. We refer to the 256-dimensional latent representation of each image as perceptual features.

We also use a β-VAE to learn a compressed latent representation z of the input images x [189, 41]. These neural networks learn a probabilistic generative model showing how the input data x depends on unobserved latent variables z and approximate the optimal posterior distribution $q_\phi(z|x)$ over the latent factors given the observations. The objective of a VAE is to minimize the distance (KL-divergence) between the approximate $q_\phi(z|x)$ and true posterior distribution $p(z|x)$ by maximizing the evidence lower bound, described in the next section. The latent variables z can also be used to reconstruct inputs x with a measurable error. The KL-divergence and reconstruction error term in combination forms the training objective of a variant of the VAE called the β-VAE: minimizing reconstruction error (binary cross-entropy)

and minimizing the KL-divergence term to varying degrees according to the scalar parameter β. The objective function is:

$$L_{\beta\text{-VAE}} = \text{reconstruction error} + \beta\,\mathrm{KL}(q_\phi(z|x) \parallel p(z)) \tag{4}$$

Increasing β enforces stronger constraints on the KL-divergence; in practice, this results in greater regularization of the latent space. In sparse β-VAEs, we further constrain the loss function to penalize a L1 regularization term on the magnitude of latent representations z.

The encoder consisted of four convolutional layers with ReLU activations, progressively reducing spatial dimensions while increasing feature depth. A fully connected bottleneck layer z with 512 dimensions parameterized the mean and covariance of the latent distribution, and a latent vector was sampled using the reparameterization trick using a Gaussian distribution parameterized by ϕ. The bottleneck layer factorizes the distribution over inputs x, forming a probabilistic representation that can disentangle the primary factors of variation across observations. The decoder, composed of transposed convolutional layers with ReLU activations, reconstructed images from the latent space. The final layer of the decoder applied a sigmoid activation to ensure output pixel values remained between 0 and 1. We trained the models over 5 random seeds using a batch size of 128 over 200 epochs with the Adam optimizer with a learning rate of 0.001. Each of the 5 seeds performed a sweep across 22 β values ranging from 10^{-6} to 10 to estimate the rate-distortion function of each image.

Image classification

We used two pre-trained convolutional neural networks called AlexNet and VGG-16 to encode different feature representations of the task stimuli. We refer to these representations as perceptual and semantic features. Stimuli were rescaled to 214x214 (x3 RGB color channels). Both networks were trained on a large number of images across a diverse set of categories in the ImageNet dataset, consisting of 1,281,167 training, 50,000 validation, and 100,000 test images across 1,000 semantic categories. The networks were trained to categorize images into the 1000 semantic categories.

AlexNet consists of nine layers. Briefly, the input layer maps unit activation to colors. Next, the output is processed sequentially by five convolutional layers, containing spatial filters that convolve with the input from the previous layer to produce the 3D tensor representing spatial activations of a particular visual pattern (see **Figure 4A**). Similar to the brain's ventral visual pathway, the hierarchical structure of convolutional layers are thought to extract low-level spatial details first then more and more abstract

higher-level semantic representations in each subsequent layer to support image categorization. Next, there are three fully connected layers which flatten the 3D tensor into a vector.

For each image in the mnemonic similarity task, we obtained six vectors which we refer to as perceptual and semantic feature representations. We obtained a 4096-dimensional vector from the penultimate fully connected layer that the following layer uses for semantic categorization. The penultimate layer is used instead of the last fully connected layer in order to obtain more generalizable, less task-specific, and richer features that are useful for transfer learning, clustering, and similarity tasks. In contrast, the last layer's features are tailored for categorizing images into 1,000 categories. We also obtained five feature representations with differing dimensionality from the activation of each convolutional layer. The feature representations were 96-dimensional in the first layer, 256-dimensional in the second layer, 384-dimensional in the third layer, 384-dimensional in the fourth layer, and 256-dimensional in the fifth layer.

VGG-16 is another convolutional neural network trained to categorize images on the ImageNet dataset but has a different architecture. The input dimensionality for images is again 224x224x3 and the network is again trained to classify images into 1000 categories. The input image is processed by 13 convolutional layers and 3 fully connected layers. For each image in the mnemonic similarity task, we obtained a 4096-dimensional feature representation from the penultimate fully connected layer.

Image to text

To obtain semantic feature representations, we used the contrastive language-image pre-training (CLIP) neural network framework with the ViT-H/14 image embedding variant, a hierarchical vision transformer. This model was pre-trained with the LAION-2B English dataset, a dataset of 2 billion English image-text pairs. Briefly, a vision encoder transforms images into vector representations, using pixel patches. The text encoder processes text descriptions through a transformer architecture, creating word-vector embeddings in the same semantic space as the visual features. The model is trained to optimize a contrastive objective, maximizing similarity between paired image-text embeddings while minimizing similarity with unpaired samples. This objective constrains the vision and text encoders to learn representations in a joint embedding space where semantically similar concepts are more similarly represented regardless of whether their source was an image or text. The resulting image embeddings from this model can be used for zero-shot image classification and other tasks without requiring task-specific fine-tuning. For example, the model performs zero-shot categorization of the Imagenet dataset with a 78.0% accuracy despite not being explicitly trained to do so. For each image stimulus in the mnemonic similarity task, we use this pre-trained model to obtain

a text description. This description is tokenized and inputted to the pre-trained text encoder to extract a 768-dimensional word-vector embedding. We refer to the 768-dimensional vectors as semantic features.

Calculating lossiness

Lossiness from a simulation of memory retrieval and identification

To compute lossiness for all models, except for the β-VAE, we adapted a previously published information theoretic algorithm [38]. The algorithm infers a loss function based on efficiently coding (compressing) an input X_1 into an output X_2. In that work, the lossiness of compression explained when two perceptually similar items are confused (generalized) or discriminated across many sensory modalities, such as a range of tones, colors, line lengths, and shapes.

In the context of the mnemonic similarity task, we determine the perceptual loss when compressing $X_{targets}$ and X_{lures} in latent space. To simulate the cost of remembering $X_{targets}$ given X_{lures}, we linearly interpolate between x_{target} and x_{lure} in latent space. Similar images are thought to exist on a low-dimensional manifold in this latent space [74, 76], and linearly interpolating between points on this manifold can be thought of as a trajectory along the manifold's surface from one image to another. This trajectory represents samples that transform according to a primary factor of variation, such as the rotation leftward versus rightward a face, the shape of a square to circle, and the color of a blue to orange object. After generating intermediate representations along this manifold, we calculate the cosine distance between all pairs of latent representations. Then, we convert this distance matrix to a similarity matrix and normalize values by the the maximum similarity across latent representations. Finally we simulate a confusion matrix that represents the perceptual information channel, where rows are input representations and columns are output representations, based by converting all values of the similarity matrix to probablities marginalizing over the output columns. Drawing 250 samples from this probability matrix simulates how often an input representation along the target-lure image manifold is mistaken for each output representation. Previously published R code was used to fit a model based on rate-distortion theory to the confusion matrix. Specifically, the lossiness was inferred using the Blahut algorithm for computing an optimal, capacity-limited information channel. This optimization procedure is grounded in rate-distortion theory and iteratively adjusts conditional probabilities to minimize the lossiness at a given information rate. Model fitting was performed over 100 iterations, ensuring convergence to a stable solution, by maximizing the log-likelihood of the observed confusion matrix under the inferred probability distributions, incorporating a prior over the lossiness that penalized deviations from symmetry and constrained diagonal elements. To improve robustness, multiple fitting attempts were

conducted, selecting the best fit based on the highest log-likelihood value. This procedure returns a cost matrix where each element is the lossiness or cost of error for confusing an input (row) with each output (column element). The lossiness used in the current paper is the average lossiness across the cost matrix per pair of target and lure images.

Optimal lossiness from variational methods

In a β-VAE, the loss function is defined with a trade-off that parallels rate-distortion trade-offs. Recall that the objective function in Equation 4 is $L_{\beta\text{-VAE}} = \text{reconstruction error} + \beta\,\text{KL}(q_\phi(z|x) \parallel p(z))$. Rate-distortion functions can be approximated by interpreting the reconstruction error (here defined as the binary cross entropy loss) as the distortion D, the information-limiting KL-divergence term as the rate R, and β as a Lagrange multiplier that scales the regularization. Larger β enforces more constraints on the information channel.

This objective is a special unsupervised case of the supervised information bottleneck method which separates task-relevant and irrelevant information [69]. Formulating the objective function in this way enables a controllable rate-distortion trade-off:

$$L_{\text{rate-distortion}} = D + \beta\, R. \tag{5}$$

This rate-distortion loss function characterizes how a better lossy compression minimizes both the distortion and the information rate needed to transmit the source. Here, increasing β enforces more conservative information rates, or more aggressive compression characterizing a more capacity-limited information channel. Using this method, the lossiness of each image was measured as the average reconstruction error of the target and lure images across β for the seed with the lowest train/test loss. The normalized rate was measured by calculating the slope from a linear function that best fit the rate-distortion function plotted in a semi-log plot: $\frac{log_{10}(R)}{D}$.

MRI data acquisition

Neuroimaging was performed on a Siemens 3 Tesla TIM Trio scanner using a 32-channel receive-only head coil. Structural images were acquired using a T1-weighted MP-RAGE sequence with the following parameters: 192 interleaved slices; total acquisition = 8:55 min; TR = 20 ms; TE = 4.92 ms; flip angle = 25°; field of view = 256 mm; slice thickness = 1 mm; voxel resolution = 1.0 × 1.0 × 1.0 mm; 1 average. A T2-weighted

pulse sequence was acquired with the following parameters: 35 interleaved slices; total acquisition = 9:44 min; TR = 6,000 ms; TE = 64 ms; flip angle = 129°; field of view = 200 mm; slice thickness = 2 mm; voxel resolution = 0.4 × 0.4 × 2.0 mm; averages = 2. This scan was positioned perpendicular to the longitudinal axis of the hippocampus for each participant prior to acquisition.

High-resolution functional images were acquired using gradient-echo echoplanar, T2*-weighted pulse sequence utilizing a multiband (MB) technique [190]; six functional scans were conducted that coincided with six blocks of the task. These scans had the following parameters: 72 interleaved slices; TA = 6:25 min; TR = 875 ms; TE = 43.6 ms; flip angle 55°; field of view = 180 mm; slice thickness = 1.8 mm; voxel resolution = 1.8 × 1.8 × 1.8 mm; MB factor = 8; measurements = 428. These scans were aligned with the longitudinal axis of the hippocampus for each participant. The first four TRs of each run were discarded to allow for T1 equilibration.

fMRI preprocessing

MRI data were analyzed using the Analysis of Functional NeuroImages (AFNI) software (version 22.1.09; [191]) following the first steps of a prior publication using the same data [57]. Given the short acquisition time, functional scans were not slice-time corrected. Motion correction for functional scans was calculated to align each volume to the single volume of the experiment with the smallest number of outlier voxel values. Structural scans were also aligned with the minimum-outlier functional volume. Rotated structural scans were then skull-stripped and warped into MNI space using a nonlinear diffeomorphic transformation. The motion correction and MNI normalization spatial transformations were concatenated and applied to the functional scans in a single step, thus resulting in a single spatial transformation for functional data. Functional data were scaled by the mean of the overall signal for each run. No blurring was done to functional data as part of preprocessing and spatial resolution of functional scans was maintained at 1.8 mm^3.

Using this minimally pre-processed data, we performed trial-wise GLM estimation using the GLMsingle toolbox implemented in Python [81]. This toolbox identifies an optimal HRF from a library of 20 canonical functional forms, performs automated denoising using data-driven nuisance regressors from selected repeat trials, and implements fractional-ridge regularization to regularize estimates based on voxel-wise reliabilities [192]. This pipeline does not benefit from regressing out putative nuisance components of fMRI data, such as motion, white matter, or cerebrospinal fluid, prior to running GLMsingle because those steps can bias the data-driven learning of nuisance components which potentially overlap with signal components of interest. For these reasons, our pre-processing departed from the steps in the prior work using this dataset [57],

i.e., we also did not use the motion scrubbing (framewise censoring of TRs with greater than 0.3° of rotation or 0.6 mm of translation in any direction, as well as the immediately preceding TR) or coverage masks removing very low EPI signal.

Region of interest maps were created for the ventral visual stream (V1, V2, V3, V4, and IT) that progressively extract low-level spatial details and high-level semantic abstractions, as well as the hippocampal subregions (anterolateral entorhinal cortex, DG/CA3, CA1, and subiculum) that classically support pattern separation for inputs from the object processing stream.

For each subject, we masked all ROIs. Then, for 1 of the 6 runs, we split the 2nd run for each participant into two half runs because GLMsingle requires repeat trials across runs for cross-validation and the original task only includes repeat trials within a run. With stimulus duration = 2.5 s and repetition time = 0.875 s, we ran GLMsingle with default options. Outputs used for further analysis included the per-trial evoked activity β maps and map of HRF indices. Split-half reliability on βs were computed voxelwise, though we did not exclude any low-reliability voxels to avoid potential double dipping, biasing selection for voxels showing stronger memory reactivation for repeats.

fMRI analysis

To analyze representations of distinct stimuli, we used the trial-wise evoked activity magnitudes β as well as a data-driven selection of a hemodynamic response function (HRF) from a library of canonical functions. We use the magnitude and HRF to construct the activity time series per voxel evoked by each stimulus. The metrics calculated using the neural data scale strongly with the size of the input. Therefore, to obtain regional metrics which are more fairly comparable across ROIs of substantially different size (e.g. larger primary visual cortex versus smaller dentate gyrus/CA3), we sampled 100 voxel-wise time series over 100 iterations to compute averaged regional signatures of lossy compression.

Neural dimensionality of target and lure representations

The first signature of lossy compression we investigated treats dimensionality reduction as a form of lossy compression, because reconstructing an input using a reduced number of dimensions introduces distortions. The key quantity is the representational dimensionality of target-lure stimuli, calculated as the participation ratio of the covariance matrix of the target and lure time series.

$$\text{Dimensionality} = \frac{\left(\sum_{i=1}^{N} \lambda_i\right)^2}{\sum_{i=1}^{N} \lambda_i^2} \tag{6}$$

where λ_i are the eigenvalues of the covariance matrix of the target and lure time series, and N is the number of dimensions. Participation ratios are like a continuous and normalized version of the discrete principal components that explain all variance in the data. Higher participation ratios indicate neural representations with greater dimensionality and a larger capacity to represent more features or stimuli [43, 46, 71] which may support the separability of distinct inputs or input categories [5, 76]. In contrast, a lossy compression account of pattern separation predicts that separability is supposed by lower dimensionality, selectively discarding less relevant features to only retain essential information.

A signature of lossy compression is therefore dimensionality reduction for correct versus incorrect trials: the difference between the dimensionality of trials with correct lure responses ('similar') minus the dimensionality of trials with incorrect lure false alarm responses ('old').

$$\Delta\text{Dimensionality} = \text{Dimensionality}_{\text{correct}} - \text{Dimensionality}_{\text{incorrect}} \tag{7}$$

where $\text{Dimensionality}_{\text{correct}}$ is the participation ratio of lure trials with correct "similar" responses, and $\text{Dimensionality}_{\text{incorrect}}$ is the participation ratio of lure trials with incorrect "old" responses.

Information rate between target and lure representations

The second signature of lossy compression we investigated is the mutual information between the neural representations of targets and lures. The mutual information is the amount of information retained between representations and should be lower for correct trials versus incorrect trials according to a lossy compression account. Mutual information was calculated by discretizing values into fixed bins and using the mutual information score between the resulting discretized distributions [193]:

$$I(X;Y) = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \frac{n_{ij}}{n} \log\left(\frac{n_{ij} \cdot n}{n_{i\cdot}\, n_{\cdot j}}\right) \tag{8}$$

where n_{ij} is the number of co-occurrences of bin i in X and bin j in Y, $n_{i\cdot}$ and $n_{\cdot j}$ are the corresponding marginal counts, and n is the total number of samples. To calculate within-region information rates for incorrect trials and correct trials, we concatenated and flattened all voxel-wise time courses for targets and separately for lures. Then, we separated these lure and target time courses by whether the trial was evaluated

as a correct lure response ('similar') and incorrect lure false alarm response ('old'). Finally, for correct trial mutual information, we defined X as the target time courses for the correct trials and Y as the lure time courses for the correct trials. For incorrect trial mutual information, we did the same but with only the incorrect trials. The direction of results did not differ when using more advanced adaptive discretization algorithms which are better able to capture non-linear relationships [194, 195].

Statistical analysis

To test the effect of lossiness on the ease of pattern separation performance, we used a Spearman correlation between lure bin and lossiness. To test the effect of lossiness on pattern separation performance, we tested a linear model with mixed effects at the trial-level nested by participant:

$$\text{Lure discrimination index} = \beta_0 + \beta_1 \cdot \text{lossiness} + \beta_2 \cdot \text{Age} + (1|\text{participant}) + \epsilon \tag{9}$$

The lure discrimination index was calculated across trials within each lure bin. The lossiness was the mean across the images presented in each lure bin. We only included age as a covariate for the dataset that spanned the lifespan and the dataset containing children and adolescents. Finally, the effects of age are often non-linear. To test both linear and non-linear effects of age on the relationship between lossiness and pattern separation performance, we used generalized additive models with penalized splines, a method which allows for statistically rigorous modeling of linear and nonlinear effects while minimizing over-fitting [77]. We tested the model:

$$\text{Lure discrimination index} = \beta_0 + s(\text{Lossiness, by age}, k = 4) + \epsilon \tag{10}$$

The model included k as a smooth term for lossiness, capturing non-linear effects of lossiness across different age groups, using 4 basis functions and fit using restricted maximum likelihood estimation and fixed effects.

Citation diversity statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field [196, 197, 198, 199, 200, 201, 202, 203, 204]. Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we

obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman [200, 205]. By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 12.27% woman(first)/woman(last), 12.63% man/woman, 21.27% woman/man, and 53.84% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color [206, 207]. By this measure (and excluding self-citations), our references contain 6.18% author of color (first)/author of color(last), 12.90% white author/author of color, 22.74% author of color/white author, and 58.18% white author/white author. This method is limited in that a) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

4 Acknowledgments

D.Z. acknowledges funding from the George E. Hewitt Foundation for Medical Research. A.M.B. acknowledges funding from the National Institute on Aging (R21AG072673 and R01AG088306). M.Y. and A.M.B. acknowledge funding from the National Institute for Mental Health (R01MH128306).

5 Data availability

All fMRI data are available at OpenNeuro and pre-processing scripts used in data analyses are available at GitHub.

References

[1] S. M. Noh, V. X. Yan, M. S. Vendetti, A. D. Castel, and R. A. Bjork, "Multilevel induction of categories: Venomous snakes hijack the learning of lower category levels," *Psychological science*, vol. 25, no. 8, pp. 1592–1599, 2014.

[2] M. Botvinick, A. Weinstein, A. Solway, and A. Barto, "Reinforcement learning, efficient coding, and the statistics of natural tasks," *Current opinion in behavioral sciences*, vol. 5, pp. 71–77, 2015.

[3] A. M. Bornstein, M. Aly, S. F. Feng, N. B. Turk-Browne, K. A. Norman, and J. D. Cohen, "Associative memory retrieval modulates upcoming perceptual decisions," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 23, no. 3, pp. 645–665, 2023.

[4] W. W. Pettine, D. V. Raman, A. D. Redish, and J. D. Murray, "Human generalization of internal representations through prototype learning with goal-directed attention," *Nature human behaviour*, vol. 7, no. 3, pp. 442–463, 2023.

[5] N. A. Cayco-Gajic and R. A. Silver, "Re-evaluating Circuit Mechanisms Underlying Pattern Separation," *Neuron*, vol. 101, pp. 584–602, Feb. 2019. Publisher: Elsevier.

[6] J. C. Hulbert and K. Norman, "Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice," *Cerebral Cortex*, vol. 25, no. 10, pp. 3994–4008, 2015.

[7] G. Kim, K. A. Norman, and N. B. Turk-Browne, "Neural differentiation of incorrectly predicted memories," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2022–2031, 2017.

[8] A. Khoudary, M. A. Peters, and A. M. Bornstein, "Precision-weighted evidence integration predicts time-varying influence of memory on perceptual decisions," *Cognitive Computational Neuroscience*, 2022.

[9] S. M. Noh, U. K. Singla, I. J. Bennett, and A. M. Bornstein, "Memory precision and age differentially predict the use of decision-making strategies across the lifespan," *Scientific Reports*, vol. 13, no. 1, p. 17014, 2023.

[10] O. Bein, C. Gasser, T. Amer, A. Maril, and L. Davachi, "Predictions transform memories: How expected versus unexpected events are integrated or separated in memory," *Neuroscience & Biobehavioral Reviews*, vol. 153, p. 105368, 2023.

[11] F. C. Bartlett, *Remembering: A study in experimental and social psychology.* Cambridge university press, 1995.

[12] E. F. Loftus, "Memory distortion and false memory creation," *Journal of the American Academy of Psychiatry and the Law Online*, vol. 24, no. 3, pp. 281–295, 1996.

[13] M. A. Yassa, J. W. Lacy, S. M. Stark, M. S. Albert, M. Gallagher, and C. E. Stark, "Pattern separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in nondemented older adults," *Hippocampus*, vol. 21, no. 9, pp. 968–979, 2011.

[14] S. E. Motley and C. B. Kirwan, "A parametric investigation of pattern separation processes in the medial temporal lobe," *Journal of Neuroscience*, vol. 32, no. 38, pp. 13076–13084, 2012.

[15] A. Bakker, C. B. Kirwan, M. Miller, and C. E. Stark, "Pattern separation in the human hippocampal ca3 and dentate gyrus," *science*, vol. 319, no. 5870, pp. 1640–1642, 2008.

[16] J. K. Leutgeb, S. Leutgeb, M.-B. Moser, and E. I. Moser, "Pattern separation in the dentate gyrus and ca3 of the hippocampus," *science*, vol. 315, no. 5814, pp. 961–966, 2007.

[17] D. Marr, D. Willshaw, and B. McNaughton, *Simple memory: a theory for archicortex*. Springer, 1991.

[18] R. C. O'Reilly and J. L. McClelland, "Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off," *Hippocampus*, vol. 4, no. 6, pp. 661–682, 1994.

[19] J. L. McClelland and N. H. Goddard, "Considerations arising from a complementary learning systems perspective on hippocampus and neocortex," *Hippocampus*, vol. 6, no. 6, pp. 654–665, 1996.

[20] R. C. O'Reilly and J. W. Rudy, "Conjunctive representations in learning and memory: principles of cortical and hippocampal function.," *Psychological review*, vol. 108, no. 2, p. 311, 2001.

[21] R. C. O'Reilly and K. A. Norman, "Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework," *Trends in cognitive sciences*, vol. 6, no. 12, pp. 505–510, 2002.

[22] N. A. Suthana, N. N. Parikshak, A. D. Ekstrom, M. J. Ison, B. J. Knowlton, S. Y. Bookheimer, and I. Fried, "Specific responses of human hippocampal neurons are associated with better memory," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, pp. 10503–10508, 2015.

[23] H. R. Dimsdale-Zucker, M. Ritchey, A. D. Ekstrom, A. P. Yonelinas, and C. Ranganath, "Ca1 and ca3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields," *Nature communications*, vol. 9, no. 1, p. 294, 2018.

[24] J. J. Sakon and W. A. Suzuki, "A neural signature of pattern separation in the monkey hippocampus," *Proceedings of the national academy of sciences*, vol. 116, no. 19, pp. 9634–9643, 2019.

[25] S. M. Stark, C. B. Kirwan, and C. E. Stark, "Mnemonic similarity task: A tool for assessing hippocampal integrity," *Trends in cognitive sciences*, vol. 23, no. 11, pp. 938–951, 2019.

[26] A. Treves and E. T. Rolls, "Computational analysis of the role of the hippocampus in memory," *Hippocampus*, vol. 4, no. 3, pp. 374–391, 1994.

[27] K. A. Norman, "How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model," *Hippocampus*, vol. 20, no. 11, pp. 1217–1227, 2010.

[28] S. L. Leal and M. A. Yassa, "Integrating new findings and examining clinical applications of pattern separation," *Nature neuroscience*, vol. 21, no. 2, pp. 163–173, 2018.

[29] A. Santoro, "Reassessing pattern separation in the dentate gyrus," *Frontiers in Behavioral Neuroscience*, vol. 7, 2013.

[30] D. Marr, "A theory of cerebellar cortex. 202: 437–470," 1969.

[31] J. S. Albus, "A theory of cerebellar function," *Mathematical biosciences*, vol. 10, no. 1-2, pp. 25–61, 1971.

[32] C. E. Myers and H. E. Scharfman, "A role for hilar cells in pattern separation in the dentate gyrus: a computational approach," *Hippocampus*, vol. 19, no. 4, pp. 321–337, 2009.

[33] C. E. Myers and H. E. Scharfman, "Pattern separation in the dentate gyrus: a role for the ca3 backprojection," *Hippocampus*, vol. 21, no. 11, pp. 1190–1215, 2011.

[34] G. Billings, E. Piasini, A. Lőrincz, Z. Nusser, and R. Silver, "Network Structure within the Cerebellar Input Layer Enables Lossless Sparse Encoding," *Neuron*, vol. 83, pp. 960–974, Aug. 2014. Publisher: Elsevier.

[35] S. Chavlis, P. C. Petrantonakis, and P. Poirazi, "Dendrites of dentate gyrus granule cells contribute to pattern separation by controlling sparsity," *Hippocampus*, vol. 27, no. 1, pp. 89–110, 2017.

[36] S. J. Guzman, A. Schlögl, C. Espinoza, X. Zhang, B. A. Suter, and P. Jonas, "How connectivity rules and synaptic properties shape the efficacy of pattern separation in the entorhinal cortex–dentate gyrus–CA3 network," *Nature Computational Science*, vol. 1, pp. 830–842, Dec. 2021.

[37] A. J. Chanales, A. Oza, S. E. Favila, and B. A. Kuhl, "Overlap among spatial memories triggers repulsion of hippocampal representations," *Current Biology*, vol. 27, no. 15, pp. 2307–2317, 2017.

[38] C. R. Sims, "Efficient coding explains the universal law of generalization in human perception," *Science*, vol. 360, no. 6389, pp. 652–656, 2018.

[39] M. L. Mack, A. R. Preston, and B. C. Love, "Ventromedial prefrontal cortex compression during concept learning," *Nature communications*, vol. 11, no. 1, p. 46, 2020.

[40] D. Zhou, C. W. Lynn, Z. Cui, R. Ciric, G. L. Baum, T. M. Moore, D. R. Roalf, J. A. Detre, R. C. Gur, R. E. Gur, *et al.*, "Efficient coding in the economics of human brain connectomics," *Network Neuroscience*, vol. 6, no. 1, pp. 234–274, 2022.

[41] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[42] I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, and M. Botvinick, "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons," *Nature communications*, vol. 12, no. 1, p. 6456, 2021.

[43] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[44] M. E. Hasselmo, "The role of acetylcholine in learning and memory," *Current opinion in neurobiology*, vol. 16, no. 6, pp. 710–715, 2006.

[45] P. Gao and S. Ganguli, "On simplicity and complexity in the brave new world of large-scale neuroscience," *Current opinion in neurobiology*, vol. 32, pp. 148–155, 2015.

[46] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," *BioRxiv*, p. 214262, 2017.

[47] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *Nature*, vol. 571, no. 7765, pp. 361–365, 2019.

[48] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, "Separability and geometry of object manifolds in deep neural networks," *Nature communications*, vol. 11, no. 1, p. 746, 2020.

[49] R. S. Koolschijn, A. Shpektor, W. T. Clarke, I. B. Ip, D. Dupret, U. E. Emir, and H. C. Barron, "Memory recall involves a transient break in excitatory-inhibitory balance," *Elife*, vol. 10, p. e70071, 2021.

[50] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[51] D. G. Nagy, B. Török, and G. Orbán, "Optimal forgetting: Semantic compression of episodic memories," *PLoS Computational Biology*, vol. 16, no. 10, p. e1008367, 2020.

[52] D. G. Nagy, G. Orban, and C. M. Wu, "Interplay of episodic and semantic memory arises from adaptive compression," 2025.

[53] C. E. Shannon *et al.*, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1, 1959.

[54] A. P. Yonelinas, M. Aly, W.-C. Wang, and J. D. Koen, "Recollection and familiarity: Examining controversial assumptions and new directions," *Hippocampus*, vol. 20, no. 11, pp. 1178–1194, 2010.

[55] S. L. Leal, S. K. Tighe, and M. A. Yassa, "Asymmetric effects of emotion on mnemonic interference," *Neurobiology of learning and memory*, vol. 111, pp. 41–48, 2014.

[56] S. M. Noh, K. W. Cooper, C. E. Stark, and A. M. Bornstein, "Multi-step inference can be improved across the lifespan with individualized memory interventions," *PsyArXiv*, 2024.

[57] M. I. Nash, C. B. Hodges, N. M. Muncy, and C. B. Kirwan, "Pattern separation beyond the hippocampus: A high-resolution whole-brain investigation of mnemonic discrimination in healthy adults," *Hippocampus*, vol. 31, no. 4, pp. 408–421, 2021.

[58] C. E. Stark, J. A. Noche, J. R. Ebersberger, L. Mayer, and S. M. Stark, "Optimizing the mnemonic similarity task for efficient, widespread use," *Frontiers in behavioral neuroscience*, vol. 17, p. 1080366, 2023.

[59] N. V. Banavar, S. M. Noh, C. N. Wahlheim, B. S. Cassidy, C. B. Kirwan, C. E. Stark, and A. M. Bornstein, "A response time model of the three-choice mnemonic similarity task provides stable, mechanistically interpretable individual-difference measures," *Frontiers in Human Neuroscience*, vol. 18, p. 1379287, 2024.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[62] J. A. Lewis-Peacock, A. T. Drysdale, and B. R. Postle, "Neural evidence for the flexible control of mental representations," *Cerebral Cortex*, vol. 25, no. 10, pp. 3303–3313, 2015.

[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[64] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[65] M. K. Benna and S. Fusi, "Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence," *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2018422118, 2021.

[66] A. Y. Wang, K. Kay, T. Naselaris, M. J. Tarr, and L. Wehbe, "Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset," *Nature Machine Intelligence*, vol. 5, pp. 1415–1426, Dec. 2023.

[67] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown, "Dimensionality compression and expansion in deep neural networks," *arXiv preprint arXiv:1906.00443*, 2019.

[68] M. Farrell, S. Recanatesi, T. Moore, G. Lajoie, and E. Shea-Brown, "Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 564–573, 2022.

[69] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations*, 2017.

[70] C. J. Bates and R. A. Jacobs, "Efficient data compression in perception and perceptual memory.," *Psychological review*, vol. 127, no. 5, p. 891, 2020.

[71] D. Zhou, J. Z. Kim, A. R. Pines, V. J. Sydnor, D. R. Roalf, J. A. Detre, R. C. Gur, R. E. Gur, T. D. Satterthwaite, and D. S. Bassett, "Compression supports low-dimensional representations of behavior across neural circuits," *arXiv preprint arXiv:2211.16599*, 2022.

[72] J. W. Lacy, M. A. Yassa, S. M. Stark, L. T. Muftuler, and C. E. Stark, "Distinct pattern separation related transfer functions in human ca3/dentate and ca1 revealed using high-resolution fmri and variable mnemonic similarity," *Learning & memory*, vol. 18, no. 1, pp. 15–18, 2011.

[73] N. C. Rust and J. J. DiCarlo, "Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it," *Journal of Neuroscience*, vol. 30, no. 39, pp. 12978–12995, 2010.

[74] J. DiCarlo, D. Zoccolan, and N. Rust, "How Does the Brain Solve Visual Object Recognition?," *Neuron*, vol. 73, pp. 415–434, Feb. 2012. Publisher: Elsevier.

[75] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin, "The ventral visual pathway: an expanded neural framework for the processing of object quality," *Trends in cognitive sciences*, vol. 17, no. 1, pp. 26–49, 2013.

[76] S. Chung and L. Abbott, "Neural population geometry: An approach for understanding biological and artificial neural networks," *Current Opinion in Neurobiology*, vol. 70, pp. 137–144, 2021. Computational Neuroscience.

[77] S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 673–686, 2004.

[78] D. C. Park and P. Reuter-Lorenz, "The adaptive brain: aging and neurocognitive scaffolding," *Annual review of psychology*, vol. 60, pp. 173–196, 2009.

[79] L. Naspi, C. Stensholt, A. E. Karlsson, Z. A. Monge, and R. Cabeza, "Effects of aging on successful object encoding: Enhanced semantic representations compensate for impaired visual representations," *Journal of Neuroscience*, vol. 43, no. 44, pp. 7337–7350, 2023.

[80] A. J. Chanales, A. G. Tremblay-McGaw, M. L. Drascher, and B. A. Kuhl, "Adaptive repulsion of long-term memory representations is triggered by event similarity," *Psychological science*, vol. 32, no. 5, pp. 705–720, 2021.

[81] J. S. Prince, I. Charest, J. W. Kurzawski, J. A. Pyles, M. J. Tarr, and K. N. Kay, "Improving the accuracy of single-trial fmri response estimates using glmsingle," *Elife*, vol. 11, p. e77599, 2022.

[82] C. T. Ngo, N. S. Newcombe, and I. R. Olson, "The ontogeny of relational memory and pattern separation," *Developmental science*, vol. 21, no. 2, p. e12556, 2018.

[83] C. T. Ngo, Y. Lin, N. S. Newcombe, and I. R. Olson, "Building up and wearing down episodic memory: Mnemonic discrimination and relational binding.," *Journal of Experimental Psychology: General*, vol. 148, no. 9, p. 1463, 2019.

[84] K. L. Canada, C. T. Ngo, N. S. Newcombe, F. Geng, and T. Riggins, "It's all in the details: relations between young children's developing pattern separation abilities and hippocampal subfield volumes," *Cerebral Cortex*, vol. 29, no. 8, pp. 3427–3433, 2019.

[85] S. M. Korkki, F. R. Richter, P. Jeyarathnarajah, and J. S. Simons, "Healthy ageing reduces the precision of episodic memory retrieval.," *Psychology and Aging*, vol. 35, no. 1, p. 124, 2020.

[86] E. Rosch, "Cognitive representations of semantic categories.," *Journal of experimental psychology: General*, vol. 104, no. 3, p. 192, 1975.

[87] L. Luo, T. Hendriks, and F. I. Craik, "Age differences in recollection: three patterns of enhanced encoding.," *Psychology and aging*, vol. 22, no. 2, p. 269, 2007.

[88] M. Dubova and R. L. Goldstone, "The influences of category learning on perceptual reconstructions," *Cognitive Science*, vol. 45, no. 5, p. e12981, 2021.

[89] M. A. Kramer, M. N. Hebart, C. I. Baker, and W. A. Bainbridge, "The features underlying the memorability of objects," *Science advances*, vol. 9, no. 17, p. eadd2981, 2023.

[90] S. S. Cohen, C. T. Ngo, I. R. Olson, and N. S. Newcombe, "Pattern separation and pattern completion in early childhood," *Proceedings of the National Academy of Sciences*, vol. 122, no. 11, p. e2416985122, 2025.

[91] Z. Nemecz, A. Ilyés, L. Kerekes, H. Kis, M. Werkle-Bergner, and A. Keresztes, "Different object features shape mnemonic discrimination in younger and older adults,"

[92] S. Ganguli and H. Sompolinsky, "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis," *Annual review of neuroscience*, vol. 35, no. 1, pp. 485–508, 2012.

[93] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124020, 2019.

[94] D. L. Schacter, "The seven sins of memory: insights from psychology and cognitive neuroscience.," *American psychologist*, vol. 54, no. 3, p. 182, 1999.

[95] M. Wimber, A. Alink, I. Charest, N. Kriegeskorte, and M. C. Anderson, "Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression," *Nature neuroscience*, vol. 18, no. 4, pp. 582–589, 2015.

[96] T. H. Wang, K. Placek, and J. A. Lewis-Peacock, "More is less: increased processing of unwanted memories facilitates forgetting," *Journal of Neuroscience*, vol. 39, no. 18, pp. 3551–3560, 2019.

[97] C. W. Lynn, A. E. Kahn, N. Nyema, and D. S. Bassett, "Abstract representations of events arise from mental errors in learning and memory," *Nature communications*, vol. 11, no. 1, p. 2313, 2020.

[98] Q. Lin, Z. Li, J. Lafferty, and I. Yildirim, "Images with harder-to-reconstruct visual representations leave stronger memory traces," *Nature human behaviour*, vol. 8, no. 7, pp. 1309–1320, 2024.

[99] D. L. Schacter, A. C. Carpenter, A. L. Devitt, and P. P. Thakral, "1373 memory errors and distortion," *The Oxford Handbook of Human Memory, Two Volume Pack: Foundations and Applications*, pp. 1373–1399, 2024.

[100] G. Wanjia, S. E. Favila, G. Kim, R. J. Molitor, and B. A. Kuhl, "Abrupt hippocampal remapping signals resolution of memory interference," *Nature communications*, vol. 12, no. 1, p. 4816, 2021.

[101] E. Spens and N. Burgess, "A generative model of memory construction and consolidation," *Nature human behaviour*, vol. 8, no. 3, pp. 526–543, 2024.

[102] Z. M. Reagh and M. A. Yassa, "Object and spatial mnemonic interference differentially engage lateral and medial entorhinal cortex in humans," *Proceedings of the National Academy of Sciences*, vol. 111, no. 40, pp. E4264–E4273, 2014.

[103] L. M. Pidgeon and A. M. Morcom, "Cortical pattern separation and item-specific memory encoding," *Neuropsychologia*, vol. 85, pp. 256–271, 2016.

[104] P. E. Wais, S. Jahanikia, D. Steiner, C. E. Stark, and A. Gazzaley, "Retrieval of high-fidelity memory arises from distributed cortical networks," *NeuroImage*, vol. 149, pp. 178–189, 2017.

[105] Z. M. Reagh, J. A. Noche, N. J. Tustison, D. Delisle, E. A. Murray, and M. A. Yassa, "Functional imbalance of anterolateral entorhinal cortex and hippocampal dentate/ca3 underlies age-related object pattern separation deficits," *Neuron*, vol. 97, no. 5, pp. 1187–1198, 2018.

[106] Y. Zhao, A. J. Chanales, and B. A. Kuhl, "Adaptive memory distortions are predicted by feature representations in parietal cortex," *Journal of Neuroscience*, vol. 41, no. 13, pp. 3014–3024, 2021.

[107] S. A. Johnson, S. Zequeira, S. M. Turner, A. P. Maurer, J. L. Bizon, and S. N. Burke, "Rodent mnemonic similarity task performance requires the prefrontal cortex," *Hippocampus*, vol. 31, no. 7, pp. 701–716, 2021.

[108] T. Amer and L. Davachi, "Extra-hippocampal contributions to pattern separation," *elife*, vol. 12, p. e82250, 2023.

[109] G. Wanjia, S. Han, and B. A. Kuhl, "Repulsion of hippocampal representations driven by distinct internal beliefs," *Current Biology*, 2025.

[110] H. B. Barlow *et al.*, "Possible principles underlying the transformation of sensory messages," *Sensory communication*, vol. 1, no. 01, pp. 217–233, 1961.

[111] F. Attneave, "Some informational aspects of visual perception.," *Psychological review*, vol. 61, no. 3, p. 183, 1954.

[112] Z. Li, *Understanding vision: theory, models, and data*. Oxford University Press (UK), 2014.

[113] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, and P. Sterling, "How much the eye tells the brain," *Current biology*, vol. 16, no. 14, pp. 1428–1434, 2006.

[114] X. Pitkow and M. Meister, "Decorrelation and efficient coding by retinal ganglion cells," *Nature neuroscience*, vol. 15, no. 4, pp. 628–635, 2012.

[115] J. J. Harris, R. Jolivet, E. Engl, and D. Attwell, "Energy-efficient information transfer by visual pathway synapses," *Current Biology*, vol. 25, no. 24, pp. 3151–3160, 2015.

[116] L. L. Owen and J. R. Manning, "High-level cognition is supported by information-rich but compressible brain activity patterns," *Proceedings of the National Academy of Sciences*, vol. 121, no. 35, p. e2400082121, 2024.

[117] E. Tang, M. G. Mattar, C. Giusti, D. M. Lydon-Staley, S. L. Thompson-Schill, and D. S. Bassett, "Effective learning is accompanied by high-dimensional and efficient representations of neural activity," *Nature neuroscience*, vol. 22, no. 6, pp. 1000–1009, 2019.

[118] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019.

[119] M. Dubova and S. J. Sloman, "Excess capacity learning," in *Proceedings of the annual meeting of the cognitive science society*, vol. 45, 2023.

[120] T. Ito and J. D. Murray, "Multitask representations in the human cortex transform along a sensory-to-motor hierarchy," *Nature Neuroscience*, vol. 26, no. 2, pp. 306–315, 2023.

[121] T. Nigam and C. M. Schwiedrzik, "Predictions enable top-down pattern separation in the macaque face-processing hierarchy," *Nature Communications*, vol. 15, no. 1, p. 7196, 2024.

[122] J. C. Whittington, W. Dorrell, S. Ganguli, and T. E. Behrens, "Disentanglement with biological constraints: A theory of functional cell types," *arXiv preprint arXiv:2210.01768*, 2022.

[123] C. Kerrén, D. Reznik, C. F. Doeller, and B. J. Griffiths, "Exploring the role of dimensionality transformation in episodic memory," *Trends in Cognitive Sciences*, vol. 29, no. 7, pp. 614–626, 2025.

[124] C. Ranganath and G. Rainer, "Neural mechanisms for detecting and remembering novel events," *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 193–202, 2003.

[125] A. Jaegle, V. Mehrpour, and N. Rust, "Visual novelty, curiosity, and intrinsic reward in machine learning and the brain," *Current opinion in neurobiology*, vol. 58, pp. 167–174, 2019.

[126] J. E. Lisman and A. A. Grace, "The hippocampal-vta loop: controlling the entry of information into long-term memory," *Neuron*, vol. 46, no. 5, pp. 703–713, 2005.

[127] M. E. Hasselmo and B. P. Wyble, "Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function," *Behavioural brain research*, vol. 89, no. 1-2, pp. 1–34, 1997.

[128] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*, pp. 12310–12320, PMLR, 2021.

[129] V. A. Carr, S. A. Engel, and B. J. Knowlton, "Top-down modulation of hippocampal encoding activity as measured by high-resolution functional mri," *Neuropsychologia*, vol. 51, no. 10, pp. 1829–1837, 2013.

[130] M. Aly and N. B. Turk-Browne, "Attention promotes episodic encoding by stabilizing hippocampal representations," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. E420–E429, 2016.

[131] G. Son, D. B. Walther, and M. L. Mack, "Brief category learning distorts perceptual space for complex scenes," *Psychonomic Bulletin & Review*, vol. 31, no. 5, pp. 2234–2248, 2024.

[132] M. L. Mack and T. J. Palmeri, "Discrimination, recognition, and classification," 2024.

[133] M. Bar, K. S. Kassam, A. S. Ghuman, J. Boshyan, A. M. Schmid, A. M. Dale, M. S. Hämäläinen, K. Marinkovic, D. L. Schacter, B. R. Rosen, *et al.*, "Top-down facilitation of visual recognition," *Proceedings of the national academy of sciences*, vol. 103, no. 2, pp. 449–454, 2006.

[134] N. C. Hindy, F. Y. Ng, and N. B. Turk-Browne, "Linking pattern completion in the hippocampus to predictive coding in visual cortex," *Nature neuroscience*, vol. 19, no. 5, pp. 665–667, 2016.

[135] J. G. Kim, E. Gregory, B. Landau, M. McCloskey, N. B. Turk-Browne, and S. Kastner, "Functions of ventral visual cortex after bilateral medial temporal lobe damage," *Progress in neurobiology*, vol. 191, p. 101819, 2020.

[136] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," *Advances in neural information processing systems*, vol. 24, 2011.

[137] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," *Advances in neural information processing systems*, vol. 25, 2012.

[138] N. C. Rust and V. Mehrpour, "Understanding image memorability," *Trends in cognitive sciences*, vol. 24, no. 7, pp. 557–568, 2020.

[139] Z. Bylinskii, L. Goetschalckx, A. Newman, and A. Oliva, "Memorability: An image-computable measure of information utility," in *Human perception of visual information: Psychological and computational perspectives*, pp. 207–239, Springer, 2021.

[140] C. Revsine and W. A. Bainbridge, "Memorability reflects statistical regularities of the environment," *Current Opinion in Neurobiology*, vol. 94, p. 103095, 2025.

[141] K. A. Murphy and D. S. Bassett, "Interpretability with full complexity by constraining feature information," *arXiv preprint arXiv:2211.17264*, 2022.

[142] Z. Fang and C. R. Sims, "Humans learn generalizable representations through efficient coding," *Nature Communications*, vol. 16, no. 1, p. 3989, 2025.

[143] I. Kahn, L. Davachi, and A. D. Wagner, "Functional-neuroanatomic correlates of recollection: implications for models of recognition memory," *Journal of Neuroscience*, vol. 24, no. 17, pp. 4172–4180, 2004.

[144] B. A. Kuhl, M. K. Johnson, and M. M. Chun, "Dissociable neural mechanisms for goal-directed versus incidental memory reactivation," *Journal of Neuroscience*, vol. 33, no. 41, pp. 16099–16109, 2013.

[145] F. R. Richter, R. A. Cooper, P. M. Bays, and J. S. Simons, "Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory," *elife*, vol. 5, p. e18260, 2016.

[146] S. E. Favila, R. Samide, S. C. Sweigart, and B. A. Kuhl, "Parietal representations of stimulus features are amplified during memory retrieval and flexibly aligned with top-down goals," *Journal of Neuroscience*, vol. 38, no. 36, pp. 7809–7821, 2018.

[147] Z. Yang, X. Zhuang, K. A. Koenig, J. B. Leverenz, T. Curran, M. J. Lowe, and D. Cordes, "Pattern separation involves regions beyond the hippocampus in non-demented elderly individuals: A 7t object lure task fmri study," *Imaging Neuroscience*, vol. 2, pp. 1–15, 2024.

[148] R. M. Nosofsky and M. K. Johansen, "Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization," *Psychonomic Bulletin & Review*, vol. 7, no. 3, pp. 375–402, 2000.

[149] N. Sigala and N. K. Logothetis, "Visual categorization shapes feature selectivity in the primate temporal cortex," *Nature*, vol. 415, no. 6869, pp. 318–320, 2002.

[150] J. R. Folstein, T. J. Palmeri, and I. Gauthier, "Category learning increases discriminability of relevant object dimensions in visual cortex," *Cerebral Cortex*, vol. 23, no. 4, pp. 814–823, 2013.

[151] J. R. Folstein, T. J. Palmeri, A. E. Van Gulick, and I. Gauthier, "Category learning stretches neural representations in visual cortex," *Current directions in psychological science*, vol. 24, no. 1, pp. 17–23, 2015.

[152] J. L. Klippenstein, S. M. Stark, C. E. Stark, and I. J. Bennett, "Neural substrates of mnemonic discrimination: A whole-brain fmri investigation," *Brain and Behavior*, vol. 10, no. 3, p. e01560, 2020.

[153] K. A. Norman and R. C. O'Reilly, "Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach.," *Psychological review*, vol. 110, no. 4, p. 611, 2003.

[154] P. Rotshtein, R. N. Henson, A. Treves, J. Driver, and R. J. Dolan, "Morphing marilyn into maggie dissociates physical and identity face representations in the brain," *Nature neuroscience*, vol. 8, no. 1, pp. 107–113, 2005.

[155] M. Paleja, T. A. Girard, K. A. Herdman, and B. K. Christensen, "Two distinct neural networks functionally connected to the human hippocampus during pattern separation tasks," *Brain and Cognition*, vol. 92, pp. 101–111, 2014.

[156] S. Gattas, M. S. Larson, L. Mnatsakanyan, I. Sen-Gupta, S. Vadera, A. L. Swindlehurst, P. E. Rapp, J. J. Lin, and M. A. Yassa, "Theta mediated dynamics of human hippocampal-neocortical learning systems in memory formation and retrieval," *Nature communications*, vol. 14, no. 1, p. 8505, 2023.

[157] P. S. Davidson, P. Vidjen, S. Trincao-Batra, and C. A. Collin, "Older adults' lure discrimination difficulties on the mnemonic similarity task are significantly correlated with their visual perception," *The Journals of Gerontology: Series B*, vol. 74, no. 8, pp. 1298–1307, 2019.

[158] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, *Computer vision systems*. 1978.

[159] A. Yuille and D. Kersten, "Vision as bayesian inference: analysis by synthesis?," *Trends in cognitive sciences*, vol. 10, no. 7, pp. 301–308, 2006.

[160] B. A. Olshausen, G. Mangun, and M. Gazzaniga, "Perception as an inference problem," *The cognitive neurosciences*, pp. 295–304, 2014.

[161] I. M. Harlow and A. P. Yonelinas, "Distinguishing between the success and precision of recollection," *Memory*, vol. 24, no. 1, pp. 114–127, 2016.

[162] H. Jang and F. Tong, "Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks," *Nature Communications*, vol. 15, no. 1, p. 1989, 2024.

[163] C. M. Rotello and E. Heit, "Associative recognition: A case of recall-to-reject processing," *Memory & Cognition*, vol. 28, no. 6, pp. 907–922, 2000.

[164] D. A. Gallo, D. M. Bell, J. S. Beier, and D. L. Schacter, "Two types of recollection-based monitoring in younger and older adults: Recall-to-reject and the distinctiveness heuristic," *Memory*, vol. 14, no. 6, pp. 730–741, 2006.

[165] C. Wahlheim and L. Richmond, "A role for pattern completion in lure rejection evinced in subsequent order memory,"

[166] C. Wahlheim, I. Dobbins, and B. Wellons, "Mnemonic discrimination language evinces recollection rejection of similar lures,"

[167] G. F. DiRisio, C. Xue, and M. R. Cohen, "Neuronal signatures of successful one-shot memory in mid-level visual cortex," *bioRxiv*, 2025.

[168] M. E. Wheeler, S. E. Petersen, and R. L. Buckner, "Memory's echo: vivid remembering reactivates sensory-specific cortex," *Proceedings of the National Academy of Sciences*, vol. 97, no. 20, pp. 11125–11129, 2000.

[169] B. A. Kuhl, J. Rissman, M. M. Chun, and A. D. Wagner, "Fidelity of neural reactivation reveals competition between memories," *Proceedings of the National Academy of Sciences*, vol. 108, no. 14, pp. 5903–5908, 2011.

[170] B. McNaughton and R. O'reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory," *Psychol Rev*, 1995.

[171] O. Bein and Y. Niv, "Schemas, reinforcement learning and the medial prefrontal cortex," *Nature Reviews Neuroscience*, vol. 26, no. 3, pp. 141–157, 2025.

[172] G. Giguère and B. C. Love, "Limits in decision making arise from limits in memory retrieval," *Proceedings of the National Academy of Sciences*, vol. 110, no. 19, pp. 7613–7618, 2013.

[173] J. M. Wolfe, "Guided search 6.0: An updated model of visual search," *Psychonomic bulletin & review*, vol. 28, no. 4, pp. 1060–1092, 2021.

[174] T. F. Brady, T. Konkle, and G. A. Alvarez, "Compression in visual working memory: using statistical regularities to form more efficient memory representations.," *Journal of Experimental Psychology: General*, vol. 138, no. 4, p. 487, 2009.

[175] C. R. Sims, "The cost of misremembering: Inferring the loss function in visual working memory," *Journal of vision*, vol. 15, no. 3, pp. 2–2, 2015.

[176] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[177] Y. Karklin and E. Simoncelli, "Efficient coding of natural images with a population of noisy linear-nonlinear neurons," *Advances in neural information processing systems*, vol. 24, 2011.

[178] K. Mukherjee, X. Lu, H. Huey, Y. Vinker, R. Aguina-Kang, A. Shamir, and J. E. Fan, "Evaluating machine comprehension of sketch meaning at different levels of abstraction," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, 2023.

[179] B. Long, J. E. Fan, H. Huey, Z. Chai, and M. C. Frank, "Parallel developmental changes in children's production and recognition of line drawings of visual concepts," *Nature Communications*, vol. 15, no. 1, p. 1191, 2024.

[180] S. Recanatesi, S. Bradde, V. Balasubramanian, N. A. Steinmetz, and E. Shea-Brown, "A scale-dependent measure of system dimensionality," *Patterns*, vol. 3, no. 8, 2022.

[181] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker, "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements," *Nature human behaviour*, vol. 4, no. 11, pp. 1173–1185, 2020.

[182] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint arXiv:2306.09344*, 2023.

[183] G. Kim, J. A. Lewis-Peacock, K. A. Norman, and N. B. Turk-Browne, "Pruning of memories by context-based prediction error," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8997–9002, 2014.

[184] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, *et al.*, "A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence," *Nature neuroscience*, vol. 25, no. 1, pp. 116–126, 2022.

[185] K. A. Murphy and D. S. Bassett, "Information decomposition in complex systems via machine learning," *Proceedings of the National Academy of Sciences*, vol. 121, no. 13, p. e2312988121, 2024.

[186] J. Kim and M. A. Yassa, "Assessing recollection and familiarity of similar lures in a behavioral pattern separation task," *Hippocampus*, vol. 23, no. 4, pp. 287–294, 2013.

[187] C. B. Kirwan and C. E. Stark, "Overcoming interference: An fmri investigation of pattern separation in the medial temporal lobe," *Learning & Memory*, vol. 14, no. 9, pp. 625–633, 2007.

[188] C. E. Stark, G. D. Clemenson, U. Aluru, N. Hatamian, and S. M. Stark, "Playing minecraft improves hippocampal-associated memory for details in middle aged adults," *Frontiers in sports and active living*, vol. 3, p. 685286, 2021.

[189] D. P. Kingma, M. Welling, *et al.*, "Auto-encoding variational bayes," 2013.

[190] J. Xu, S. Moeller, E. J. Auerbach, J. Strupp, S. M. Smith, D. A. Feinberg, E. Yacoub, and K. Uğurbil, "Evaluation of slice accelerations using multiband echo planar imaging at 3 t," *Neuroimage*, vol. 83, pp. 991–1001, 2013.

[191] R. W. Cox, "Afni: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical research*, vol. 29, no. 3, pp. 162–173, 1996.

[192] A. Rokem and K. Kay, "Fractional ridge regression: a fast, interpretable reparameterization of ridge regression," *GigaScience*, vol. 9, no. 12, p. giaa133, 2020.

[193] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[194] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 6, p. 066138, 2004.

[195] J. T. Lizier, "Jidt: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI*, vol. 1, p. 11, 2014.

[196] S. M. Mitchell, S. Lange, and H. Brus, "Gendered citation patterns in international relations journals," *International Studies Perspectives*, vol. 14, no. 4, pp. 485–492, 2013.

[197] M. L. Dion, J. L. Sumner, and S. M. Mitchell, "Gendered citation patterns across political science and social science methodology fields," *Political Analysis*, vol. 26, no. 3, pp. 312–327, 2018.

[198] N. Caplar, S. Tacchella, and S. Birrer, "Quantitative evaluation of gender bias in astronomical publications from citation counts," *Nature Astronomy*, vol. 1, no. 6, p. 0141, 2017.

[199] D. Maliniak, R. Powers, and B. F. Walter, "The gender citation gap in international relations," *International Organization*, vol. 67, no. 4, pp. 889–922, 2013.

[200] J. D. Dworkin, K. A. Linn, E. G. Teich, P. Zurn, R. T. Shinohara, and D. S. Bassett, "The extent and drivers of gender imbalance in neuroscience reference lists," *bioRxiv*, 2020.

[201] M. A. Bertolero, J. D. Dworkin, S. U. David, C. L. Lloreda, P. Srivastava, J. Stiso, D. Zhou, K. Dzirasa, D. A. Fair, A. N. Kaczkurkin, B. J. Marlin, D. Shohamy, L. Q. Uddin, P. Zurn, and D. S. Bassett, "Racial and ethnic imbalance in neuroscience reference lists and intersections with gender," *bioRxiv*, 2020.

[202] X. Wang, J. D. Dworkin, D. Zhou, J. Stiso, E. B. Falk, D. S. Bassett, P. Zurn, and D. M. Lydon-Staley, "Gendered citation practices in the field of communication," *Annals of the International Communication Association*, 2021.

[203] P. Chatterjee and R. M. Werner, "Gender disparity in citations in high-impact journal articles," *JAMA Netw Open*, vol. 4, no. 7, p. e2114509, 2021.

[204] J. M. Fulvio, I. Akinnola, and B. R. Postle, "Gender (im)balance in citation practices in cognitive neuroscience," *J Cogn Neurosci*, vol. 33, no. 1, pp. 3–7, 2021.

[205] D. Zhou, E. J. Cornblath, J. Stiso, E. G. Teich, J. D. Dworkin, A. S. Blevins, and D. S. Bassett, "Gender diversity statement and code notebook v1.0," Feb. 2020.

[206] A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena, "Name-ethnicity classification from open sources," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 49–58, 2009.

[207] G. Sood and S. Laohaprapanon, "Predicting race and ethnicity from the sequence of characters in a name," *arXiv preprint arXiv:1805.02109*, 2018.