PNAS
www.pnas.org

Supplementary Information for:

Pinpointing the neural signatures of single-exposure visual recognition memory

Vahid Mehrpour, Travis Meyer, Eero P. Simoncelli and Nicole C. Rust
Correspondence: Nicole C. Rust, nrust@psych.upenn.edu

This PDF file includes:

Supplementary Methods

Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta)*, with weights of 11 kg and 12 kg and estimated ages of 10 years and 7 years, respectively. Both animals were implanted with head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

The single-exposure, contrast-invariant visual memory task:

All behavioral training and testing were performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli were presented on an LCD monitor with an 85 Hz refresh rate using customized software (http://mworks-project.org).

As an overview of the monkeys' task, each trial involved viewing one image for 500 ms, after which the monkeys indicated whether it was novel (never seen before) or repeated (seen exactly once prior) with an eye movement to one of two response targets. Images were never presented more than twice during the entire training and testing period of the experiment. Trials were initiated when the monkey fixated on a small red square (0.25°) on the center of a gray screen followed by a 200 ms delay before a 4° image appeared within a circular aperture, positioned at the center of gaze. The monkeys needed to maintain fixation of the stimulus for 500 ms, at which time the red square turned green (the go cue) and the targets appeared. The monkeys then made a saccade to a target indicating whether the stimulus was novel or repeated, and correct responses were rewarded with juice. Targets were positioned 8° above or below the stimulus. The association between the target (up vs. down) and the report (novel vs. repeated) was swapped between the two animals.

The images used in these experiments collected via an automated procedure that gathered images from the internet. Images smaller than 96*96 pixels were not considered. Eligible images were cropped to be square and resized to 256*256 pixels. Duplicate images were removed. Colored images were converted to grayscale and were presented at two contrasts ("low (L)" and "high (H)") in all possible combinations as novel and repeated (novel-repeated = low-low (LL); high-high (HH); low-high (LH); high-low (HL)). Contrast modifications were applied in a manner that did not adjust image luminance (L_v), the mean pixel intensity. Images with L_v outside the range 0.25 – 0.75 were excluded. The computation of contrast began by first computing the median of the pixel intensities that fell above and below L_v, $L_{v\text{-hi}}$ and $L_{v\text{-lo}}$. The native contrast for each image C_{native} was computed as:

$$C_{\text{native}} = (L_{v\text{-hi}} - L_v) + (L_v - L_{v\text{-lo}})$$

Each image was manipulated to produce a high contrast version (C_{hi} = 0.4) and low contrast version (C_{lo} = 0.2) via a procedure that maintained L_v for each image. Adjustments to contrast involved: 1) subtracting the mean pixel value, 2) rescaling the residual pixel values all by the same amount, and 3) adding back the mean. When the procedure resulted in the saturation of more than 10% of pixels beyond their maximal value (black and white), that image was excluded.

Trial locations for novel images and their repeats were presented with a uniform distribution of the subset of n-back used in the experiment. The n-back distribution was adjusted for each monkey based on training history to approximately equate overall performance between the two animals: n-back = 1, 4, 16, and 32 for monkey 1, and n-back = 1, 2, 4, and 8 for monkey 2. The specific random sequence of images presented during each session was generated offline before the start of the session. Uniform n-back distributions were achieved by constructing a sequence slightly longer than what was anticipated to be needed for the session, and by iteratively populating the sequence with novel images and their repeats at positions selected randomly from all the possibilities that remained unfilled. Because the longest n-back values (8 or 32) were the most difficult to fill, a fixed number of those were inserted first. In the relatively rare cases that the algorithm did not converge, it was restarted. The result was a partially populated sequence in which 83% of the trials were occupied. Next, the remaining 17% of trials were examined to determine whether they could be filled with novel/repeated pairs from the list of possible n-back options. The very small number of trials that remained after all possibilities had been extinguished (e.g. a 3-back scenario) were filled with ‘off n-back’ novel/repeated image pairs and these trials were disregarded in later analyses.

The monkeys’ behavioral patterns were computed for each condition after collapsing across n-back. The degree of contrast invariance reflected in each monkey’s session-averaged behavioral pattern was computed as the mean of contrast invariance computed for the novel and repeated memory conditions separately. Within each memory condition M, contrast invariance (I) of the behavioral pattern X in either memory condition was defined by:

$$\mathrm{I} = 1 - \frac{\mathrm{Var(X)}}{\mathrm{Var}(\mathrm{X}_{max})} \tag{1}$$

Where, $\mathrm{Var(X)}$ is the variance of pattern X, and $\mathrm{Var}(\mathrm{X}_{max})$ is the maximum possible variance associated with contrast in memory condition M given monkeys’ overall performance in the same memory condition M. For example, the X_{max} for an overall performance across the repeated conditions of 85% would correspond to 70%, 100%, 100% and 70% for HH, LL, HL and LH, respectively.

Neural recording:

The activity of neurons in IT was recorded via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images in both monkeys. The region of IT recorded was located on the ventral surface of the brain, over an area that spanned 5 mm lateral to the anterior middle temporal sulcus and 14-17 mm anterior to the ear canals. Recording sessions began after the monkeys were fully trained on the task and behavioral performance had plateaued. The depth and extent of IT was mapped within the recording chamber in a previous experiment [1]. Combined recording and behavioral training sessions happened 2-5 times per week across a span of 4 weeks (monkey 1) and 6 weeks (monkey 2). Neural activity was recorded with 24-channel U-probes (Plexon, Inc.) with linearly arranged recording sites spaced with 100 μm intervals. Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored using the Grapevine Data Acquisition System (Ripple, Inc.). Spike sorting was done manually offline (Plexon Offline Sorter). At least one candidate unit was identified on each recording channel, and 2-3 units were occasionally identified on the same channel. Spike sorting was performed blind to any experimental conditions to avoid bias. A multi-channel recording session was included in the analysis if: (1) the recording session was stable, quantified as the grand mean firing rate across channels changing less than 3-fold across the session; (2) over 50% of neurons were visually responsive (a loose criterion based on our previous experience in IT), assessed by a visual inspection of the rasters; and (3) the number of successfully completed novel/repeated pairs of trials exceeded 100. In monkey 1, 19 sessions were recorded and five were removed (one based on criterion 1 and four based on criterion 3). In monkey 2, 15 sessions were recorded and one was removed (based on criterion 1). The resulting data set included 14 sessions for monkey 1 (n = 427 candidate units), and 14 sessions for monkey 2 (n = 429 candidate units). The sample size (number of successful sessions recorded) was chosen based on our previous work [1].

The data reported here correspond to the subset of images for which the monkeys' behavioral reports were recorded for both novel and repeated presentations (e.g. trials in which the monkeys did not prematurely break fixation during either the novel or the repeated presentation of an image). Accurate estimate of population response magnitude requires many hundreds of units, and when too few units are included, magnitude estimates are dominated by the stimulus selectivity of the sampled units. To perform our analyses, we thus concatenated units across sessions to create larger pseudopopulations. When creating these pseudopopulations, we aligned data across sessions in a manner that preserved whether the trials were presented as novel or repeated and their experimental contrast condition. To prevent artificial correlations from influencing our results, analyses were performed after re-randomizing the responses within each condition for each unit to create many pseudopopulations. To deal with varying data sizes across sessions, the number of images included in the analysis was selected to balance incorporating data of equal sizes across sessions with not needlessly discarding data. NaNs were used as place holders for the more limited sessions in which data did not exist. The resulting pseudopopulations consisted of the responses to 180 images presented as both novel and repeated (i.e. 45 images per condition: HH, LL, HL, and LH). Spikes were counted in a temporal window over the range 100-500 ms following stimulus onset.

Contrast and memory modulations:

Contrast (c) and memory (m) modulations were computed from the grand mean firing rate (GMFR) across all units and images as:

$$c = 100 \times \frac{\text{GMFR}_H - \text{GMFR}_L}{\text{GMFR}_H} \tag{2}$$

$$m = 100 \times \frac{\text{GMFR}_H - \text{GMFR}_{HH}}{\text{GMFR}_H} \tag{3}$$

In the Results, we present both raw and baseline-corrected contrast and memory modulations. To determine the modulations after correcting for pre-stimulus baseline activity, the pre-stimulus GMFR in a 200-ms pre-stimulus time window was subtracted from the GMFR for each condition.

Linear population decoders:

For all decoders, the population response was quantified as the vector $\mathbf{x}$ containing spike counts on a given trial. To ensure that the decoder did not erroneously rely on visual selectivity, the decoder was trained on balanced pairs of novel/repeated trials in which monkeys viewed the same image (regardless of behavioral outcome or experimental contrast condition).

Cross-validated training and testing:

We applied the same, iterative cross-validated procedure for each linear decoder. On each iteration of the resampling procedure, the responses for each unit were randomly shuffled within each experimental condition to ensure that artificial correlations (e.g. between the neurons recorded in different sessions) were removed. Each iteration also involved setting aside the responses to one randomly selected image within each contrast condition (presented as both novel and repeated, for 8 trials in total) for testing classifier performance. The remaining trials were used to train one of the linear decoders to distinguish novel versus repeated images invariant to contrast, where the novel and repeated classes included the data corresponding to all n-backs and all trial outcomes. A neural prediction of the proportion of trials on which "repeated" would be reported was computed as the proportion of each distribution that took on a value less than the criterion. Finally, the predicted response pattern was rescaled by a rescaling parameter (see below) as a proxy for adjusting the population size to consider.

All decoders in this study took the general form of linear discriminators. The class (novel/repeated) of a population response vector, $\mathbf{x}$ was determined by the sign of:

$$f(\mathbf{x}) = \mathbf{w}.\mathbf{x} - b \tag{4}$$

where $\mathbf{w}$ is an N-dimensional weight vector in the N-dimensional IT neural space (N is the number of units), and *b* is decision criterion, given by:

$$b = \frac{1}{2}\mathbf{w}.(\boldsymbol{\mu}_N + \boldsymbol{\mu}_R) \tag{5}$$

Here $\boldsymbol{\mu}_N$ and $\boldsymbol{\mu}_R$ are the mean population response vectors across novel and repeated images in the training set, respectively. A population response vector $\mathbf{x}$ was classified as "novel" if $f(\mathbf{x}) > 0$, and "repeated" if $f(\mathbf{x}) < 0$.

Spike count classifier (associated with repetition suppression, RS):

Arguably the simplest classifier, the total spike decoder uses a homogeneous weight vector:

$$\mathbf{w}_{RS} = \mathbf{1} = (1, 1, \ldots, 1) \tag{6}$$

Fisher Linear Discriminant (iFLD):

The iFLD used in this study follows our previous implementation[1]. The Fisher Linear Discriminant (FLD) is defined as:

$$\mathbf{w}_{FLD} = \Sigma^{-\mathbf{1}}(\boldsymbol{\mu}_N - \boldsymbol{\mu}_R) \tag{7}$$

where Σ^{-1} is the inverse of the average covariance matrix across novel and repeated conditions:

$$\Sigma = \frac{1}{2}(\Sigma_N + \Sigma_R) \tag{8}$$

The dimensionality of our neural populations is high enough that we do not have enough data to obtain reliable covariance estimates (the amount of data needed for acceptable estimates of the off-diagonal entries is >10x what we collected in a single session). As such, we assume independence of the stimulus responses within conditions (i.e., we set the off-diagonal entries to zero). The resulting iFLD uses a weight for each unit that is proportional to its visual memory discriminability (d'):

$$\mathbf{w}_{iFLD} = \sum_{i=1}^{N} \mathbf{e}_{\mathbf{i}} \left(\frac{\mu_N^{(i)} - \mu_R^{(i)}}{\sigma_i{}^2} \right) \tag{9}$$

where $\mathbf{e}_{\mathbf{i}}$ is the unit vector along *i*-th dimension (*i*-th unit's response); N is the number of units; $\mu_N^{(i)}$ and $\mu_R^{(i)}$ are *i*-th unit's mean responses to novel and repeated images, respectively; and $\sigma_i{}^2$ is the *i*-th unit's average response variance across novel and repeated conditions:

$$\sigma_i{}^2 = \frac{1}{2}\left(\sigma_N^{(i)^2} + \sigma_R^{(i)^2} \right) \tag{10}$$

Family of contrast-corrected linear decoders:

The family of contrast-corrected linear decoders are based on weight vectors that are rotated within the plane containing the RS decoder ($\mathbf{1}$) and a contrast decoder, $\mathbf{w}_{\boldsymbol{c}}$:

$$\hat{\mathbf{w}}(\theta) = (\cos\theta - \cot\gamma \sin\theta)\hat{\mathbf{1}} + (\csc\gamma\ \sin\theta)\hat{\mathbf{w}}_{\boldsymbol{c}}\,; \quad \theta \in [\gamma - \pi, \gamma] \tag{11}$$

where, $\hat{\mathbf{w}}(\theta)$, $\hat{\mathbf{1}}$ and $\hat{\mathbf{w}}_{\boldsymbol{c}}$ are the unit vectors representing the decoding axis, RS decoder, and contrast decoder, respectively. θ is the angle between the decoder axis ($\hat{\mathbf{w}}(\theta)$) and the RS axis ($\mathbf{1}$), and γ is the angle between the RS and contrast axes. The contrast weight vector $\mathbf{w}_{\boldsymbol{c}}$ was defined as:

$$\mathbf{w}_{\boldsymbol{c}} = (\boldsymbol{\mu}_H - \boldsymbol{\mu}_L) \tag{12}$$

where $\boldsymbol{\mu}_H$ and $\boldsymbol{\mu}_L$ are the mean population response vectors across high and low contrast images in train set, respectively. This is a simple form of FLD that arises when the average covariance is a multiple of the identity, and is sometimes called a "prototype classifier". We define the SRS decoder as the axis that is orthogonal to contrast, i.e.

$$\theta_{SRS} = \gamma - \frac{\pi}{2} \tag{13}$$

Variant decoding scheme that incorporates a contrast correction:

We also evaluated a variant decoding scheme that corrected for contrast modulation (*SI Appendix*, Fig. S6). This decoder operated by correcting modulations caused by contrast along the RS axis by estimating and then subtracting the mean of population response across novel images for each contrast condition. The estimate of the mean population response at each contrast was computed after classifying novel images by contrast based on the training data, using the same prototype linear decoder used for SRS (Eq. 12).

Family of linear decoders in a 3D subspace spanned by SRS, RS, and iFLD:

To compute the 3D plots presented in *SI Appendix*, Fig. S8, we considered a 3D subspace of our high dimensional neural space spanned by SRS, RS, and iFLD where plane $\boldsymbol{\pi}_1\colon \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}$ intersects with plane $\boldsymbol{\pi}_2\colon \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{iFLD}}$ along SRS ($\wedge$ denotes exterior product; see *SI Appendix*, Fig. S8a). In this 3D subspace a decoding axis is given by:

$$\mathbf{w}(\theta, \phi) \ = \cos\theta\, \mathbf{w}_{\boldsymbol{SRS}} + \sin\theta (\cos\phi\, \mathbf{w}_{\boldsymbol{c}} + \sin\phi\, \mathbf{w}_{\mathbf{0}}) \tag{14}$$

where θ and ϕ are polar and azimuthal angles in a spherical coordinate system measured relative to $\mathbf{w}_{\boldsymbol{SRS}}$ (zenith direction) and $\mathbf{w}_{\boldsymbol{c}}$ in the plane $\boldsymbol{\pi}_1\colon \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}$ (azimuth reference), respectively. Furthermore:

$$\mathbf{w}_{\boldsymbol{c}} = \csc\gamma_1\, \mathbf{w}_{\boldsymbol{RS}} - \cot\gamma_1\, \mathbf{w}_{\boldsymbol{SRS}} \tag{15}$$

$$\mathbf{w}_{\mathbf{0}} = \csc\gamma_2 \csc\gamma_3\, \mathbf{w}_{\boldsymbol{iFLD}} - (\cot\gamma_2 \csc\gamma_3 + \cot\gamma_1 \cot\gamma_3)\mathbf{w}_{\boldsymbol{SRS}} - \csc\gamma_1 \cot\gamma_3\, \mathbf{w}_{\boldsymbol{RS}} \tag{16}$$

and

$$\gamma_1 = \operatorname{acos}(\mathbf{w}_{\boldsymbol{SRS}} \cdot \mathbf{w}_{\boldsymbol{RS}}) \tag{17}$$

$$\gamma_2 = \operatorname{acos}(\mathbf{w}_{\boldsymbol{SRS}} \cdot \mathbf{w}_{\boldsymbol{iFLD}}) \tag{18}$$

γ_3 is the angle between planes $\boldsymbol{\pi}_1\colon \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}$ and $\boldsymbol{\pi}_2\colon \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{iFLD}}$. This angle can be determined in exterior algebra as:

$$\gamma_3 = \operatorname{acos}\left(\frac{\langle \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}\, , \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{iFLD}} \rangle}{(\langle \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}\, , \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{RS}}\, \rangle\, \langle \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{iFLD}}, \mathbf{w}_{\boldsymbol{SRS}} \wedge \mathbf{w}_{\boldsymbol{iFLD}} \rangle)^{1/2}}\right) \tag{19}$$

where

$$\langle \mathbf{a} \wedge \mathbf{b}, \mathbf{a} \wedge \mathbf{c} \rangle \equiv \det\left(\begin{bmatrix} \langle \mathbf{a}, \mathbf{a} \rangle & \langle \mathbf{a}, \mathbf{c} \rangle \\ \langle \mathbf{b}, \mathbf{a} \rangle & \langle \mathbf{b}, \mathbf{c} \rangle \end{bmatrix}\right) \tag{20}$$

$\langle .\,, . \rangle$ and det(.) are the operators of inner product and determinant, respectively.

Rescaling parameter and prediction quality (PQ):

Comparing IT population decoding performance with behavior depends on the neural population size under consideration, and there is no good way to choose this *a priori.* We thus applied a fitting approach for each decoder. After confirming that performance using all recorded units in our dataset fell below saturation, we simulated increases in population size by fitting a single rescaling parameter (α) to minimize the MSE between the neural predictions and actual behavioral patterns. We emphasize that while this adjustment changed the overall performance, it did not impact the shape of the predicted behavioral patterns. The minimization of MSE yields the standard analytical solution for α:

$$\alpha = \frac{\sum_{i=1}^{6} \hat{y}_i y_i}{\sum_{i=1}^{6} {y_i}^2} \tag{21}$$

where $\hat{y}_i$ and y_i are the actual and neutrally predicted performance for condition i, respectively, and i indexes each of six conditions {HH, LL, HL, LH, H, L}.

Next, to quantify the quality of the fit after rescaling the predicted pattern, we computed a measure of prediction quality (PQ):

$$\mathrm{PQ} = 1 - \frac{MSE}{MSE_{max}} \tag{22}$$

where *MSE* and *MSE_{max}* denote the mean squared error of the rescaled predicted pattern and the pattern with maximum MSE that was matched in overall performance, respectively, i.e.

$$MSE = \frac{1}{6}\sum_{i=1}^{6}(\hat{y}_i - y'_i)^2 \tag{23}$$

and

$$MSE_{max} = \max_{\delta_i}\left(\frac{1}{6}\sum_{i=1}^{6}(\hat{y}_i - \delta_i)^2\right) \tag{24}$$

$\hat{y}_i$ and y'_i are the actual and rescaled predicted performance for condition i, respectively, and i corresponds to each of six conditions including HH, LL, HL, LH, H, and L. Each δ_i was chosen to be either 1 or $2\bar{y} - 1$ (with $\bar{y}$ the mean performance across all six conditions), in order to maximize MSE. The upper bound of PQ = 1 reflects a neural prediction that perfectly replicates the actual behavioral pattern. A value of PQ = 0 reflects the worst possible predicted behavioral pattern that is matched in overall performance. Negative PQ values reflect predicted behavioral patterns that could not be rescaled with α to match overall performance because one or more entries were pinned at saturation (e.g., as a consequence of extreme contrast modulation).

Covariance error ellipse:

Error ellipses (shown in Fig. 3a, 5a, and *SI Appendix*, Fig. S7a) were computed by first projecting the neural response vectors onto the non-orthogonal discriminant axes $\hat{\mathbf{1}}$ and $\hat{\mathbf{w}}_{\boldsymbol{c}}$, producing coordinates (u, v). These were transformed to orthogonal coordinates using a transformation matrix (derived from Eq. 11):

$$\mathrm{R} = \begin{bmatrix} 1 & 0 \\ -\cot\gamma & \csc\gamma \end{bmatrix} \tag{25}$$

where γ is the angle between the two discriminant axes:

$$\gamma = \angle(\mathbf{w_1}, \mathbf{w_2}) = \mathrm{acos}(\mathbf{w_1} \cdot \mathbf{w_2}) \tag{26}$$

We then rotate this coordinate system in the plane by angle φ, using transformation matrix:

$$\mathrm{R}_\varphi = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \tag{27}$$

Combining Eq. 25 and 27 gives an expression for the (x, y) coordinates of the projected neural responses:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ -\cot\gamma & \csc\gamma \end{bmatrix} \times \begin{bmatrix} u \\ v \end{bmatrix} \tag{28}$$

For each condition, the covariance matrix of the transformed data was computed, and the eigenvectors of this matrix provide the major and minor axes of the associated ellipse. To determine the dimensions of the ellipse, we multiplied the square root of the eigenvalues by a scale factor equal to the square root value of the cumulative chi-square distribution function (CDF) for 2-degrees of freedom evaluated at 95%.

Decomposition of total variance into variance due to identity/trial variability and contrast:

For Figs. 4, 5, and *SI Appendix*, Fig. S7, we used the following equations to decompose the average variance across novel (N) and repeated (R) conditions, $\sigma_{avg}{}^2 = \frac{1}{2}(\sigma_N{}^2 + \sigma_R{}^2)$, into the variance due to image identity and trial variability (ID) and contrast (C):

$$\sigma_{avg}{}^2 = \frac{1}{2}(\sigma_{ID}{}^2 + \sigma_C{}^2)$$

Where:

$$\sigma_{\mathrm{ID}}{}^2 = \frac{1}{2}(\sigma_H{}^2 + \sigma_L{}^2) + \frac{1}{4}(\sigma_{HH}{}^2 + \sigma_{LL}{}^2 + \sigma_{HL}{}^2 + \sigma_{LH}{}^2)$$

$$\sigma_{\mathrm{C}}{}^2 = \frac{1}{2}[(\mu_H{}^2 - \mu_N{}^2) + (\mu_L{}^2 - \mu_N{}^2)] + \cdots$$
$$\frac{1}{4}[(\mu_{HH}{}^2 - \mu_R{}^2) + (\mu_{LL}{}^2 - \mu_R{}^2) + (\mu_{HL}{}^2 - \mu_R{}^2) + (\mu_{LH}{}^2 - \mu_R{}^2)] \tag{29}$$

In each condition, σ and μ denote the standard deviation and mean of the corresponding distribution, respectively.

Fitting the four-parameter tuning model to each unit and synthesizing data:

In Fig. 5, and *SI Appendix*, Fig. S8 we assessed the population geometry in the limit of infinite samples by fitting a model to each unit that we recorded, and then using these models to synthesize population data. A 4-parameter model was used to describe the mean spike count response of each unit:

$$y(x; M, C) = A.m.c.\exp(-ax) \tag{30}$$

where x is stimulus rank, M is image memory condition (novel or repeated), C is image contrast (high or low), A is amplitude, m is memory modulation (set to 1 for novel images, and a fitted value for repeated images), c is contrast modulation (set to 1 for high contrast images, and fitted value for low contrast), and a controls stimulus selectivity.

We estimated each unit's tuning curve parameters by maximizing the likelihood (MLE) of observing the spike count data from 100 ms to 500 ms relative to stimulus onset using the techniques introduced in ref. [2]. If $\{\nu_1, \nu_2, \ldots, \nu_n\}$ is the spike count data for a unit in all six conditions (n trials in total), the log-likelihood of observing the data is given by:

$$\log \mathcal{L}(\nu|\, A, \alpha, m, c) = \sum_{i=1}^{n} \log\big(P(\nu_i|A, \alpha, m, c)\big) \tag{31}$$

where

$$P(\nu|A, \alpha, m, c) = \begin{cases} 1 + \dfrac{1}{a} \displaystyle\sum_{k=1}^{\infty} \dfrac{(-1)^k {A_X}^k}{k.k!} \left(1 - e^{-ka} \right) & ;\nu = 0 \\ \dfrac{1}{a.\nu} \displaystyle\sum_{k=0}^{\nu-1} \dfrac{{A_X}^k}{k!} \left(e^{-(ka + A_X e^{-a})} - e^{-A_X} \right) & ;\nu \neq 0 \end{cases} \tag{32}$$

and

$$A_X = \begin{cases} A & ;X\text{: H} \\ c.A & ;\ X: L \\ m.A & ;X: HH, \text{or } LH \\ m.c.A & ;X: LL, \text{or } HL \end{cases} \tag{33}$$

We estimated four tuning parameters of the unit by maximizing Eq. 31 with respect to the parameters A, a, m, and c.

Goodness of fit was assessed by comparing the actual and predicted grand mean spike counts, and only accepting units whose predicted grand mean spike counts fell in the range 0.83-1.2x of the actual values. Of 856 units, 661 units fulfilled this criterion.

Finally, we used the tuning parameters for each unit to synthesize the responses to 1000 images per condition. For each unit, we sampled x in Eq. 30 as 1000 draws from a uniform distribution between 0 and 1 and used those values to compute spike count rates, which were converted to spike counts by drawing from a Poisson process.

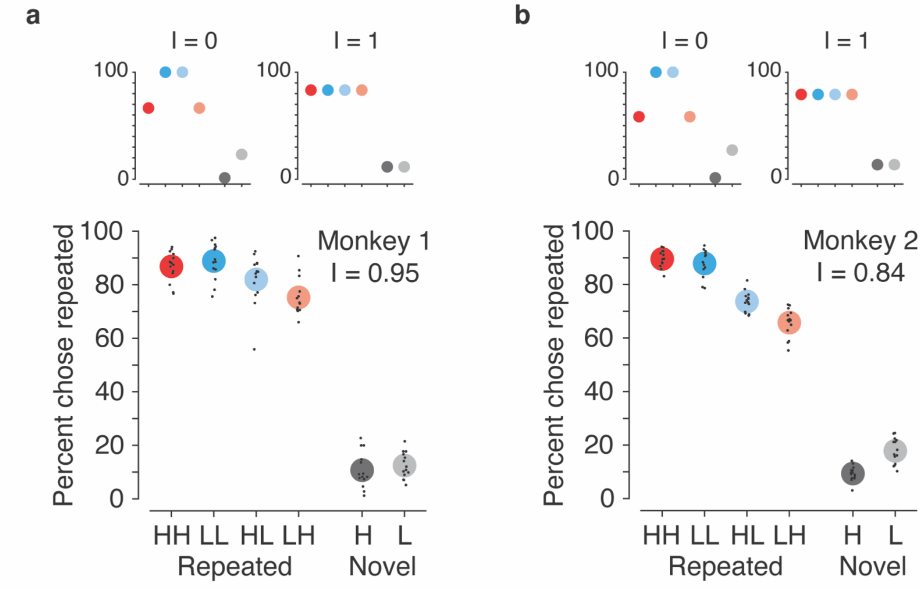

Fig. S1. *Behavioral performance patterns for individual monkeys.* **(a-b)** Fig. 1c replotted for two animals. Small black dots indicate average performance for an individual session and large colored dots indicate the average performance across sessions (14 sessions per animal). The contrast invariance reflected in each behavioral pattern (**I**) is labeled in each plot. Insets correspond to behavioral patterns with maximal (**I** = 0) and minimal (**I** = 1) contrast confusion, matched for overall performance.

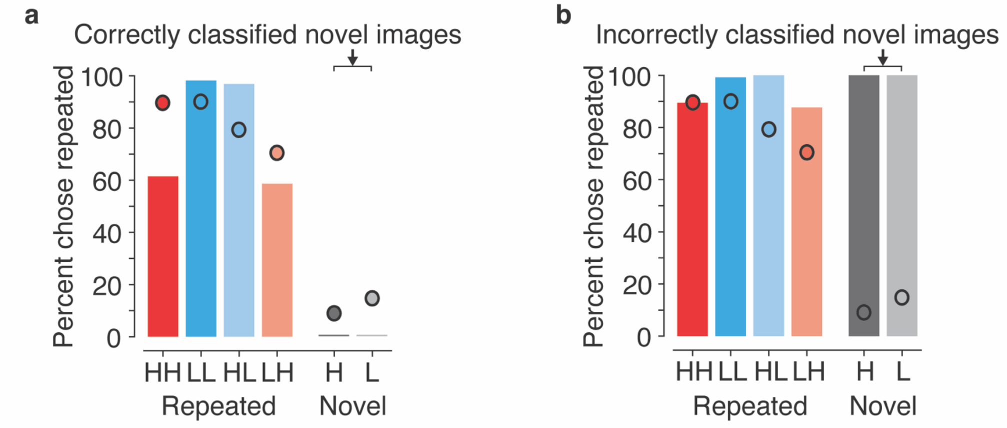

Fig. S2. *RS decoding performance, sorted by decoding performance on novel trials.* Shown is a breakdown of the behavioral pattern predicted by the RS decoder depicted in Fig. 2a, sorted by images that were **(a)** correctly, and **(b)** incorrectly classified when presented as novel. These results indicate that even when novel images are correctly classified (panel a), the predicted behavioral pattern for repeated images reflects contrast confusions.

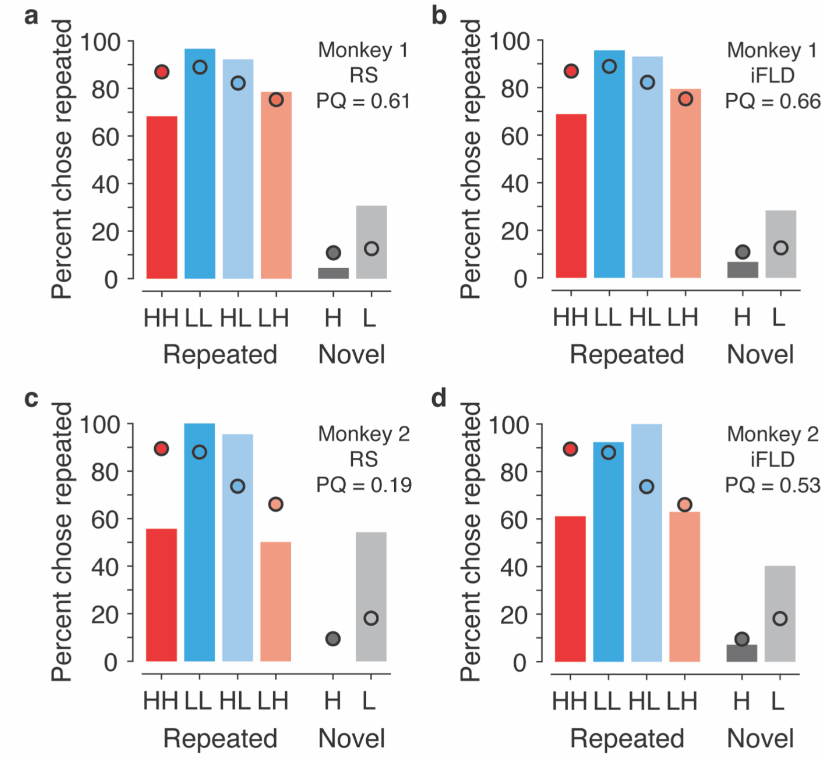

Fig. S3. *Classic linear decoders fail to map IT neural responses to behavior for each monkey.* **(a, c)** Fig. 2a replotted for each animal. **(b, d)** Fig. 2b replotted for each animal. In all panels, dots indicate the actual behavioral patterns and bars indicate the neural predictions of behavior for each type of linear decoder. Prediction quality (PQ) is indicated for each case.

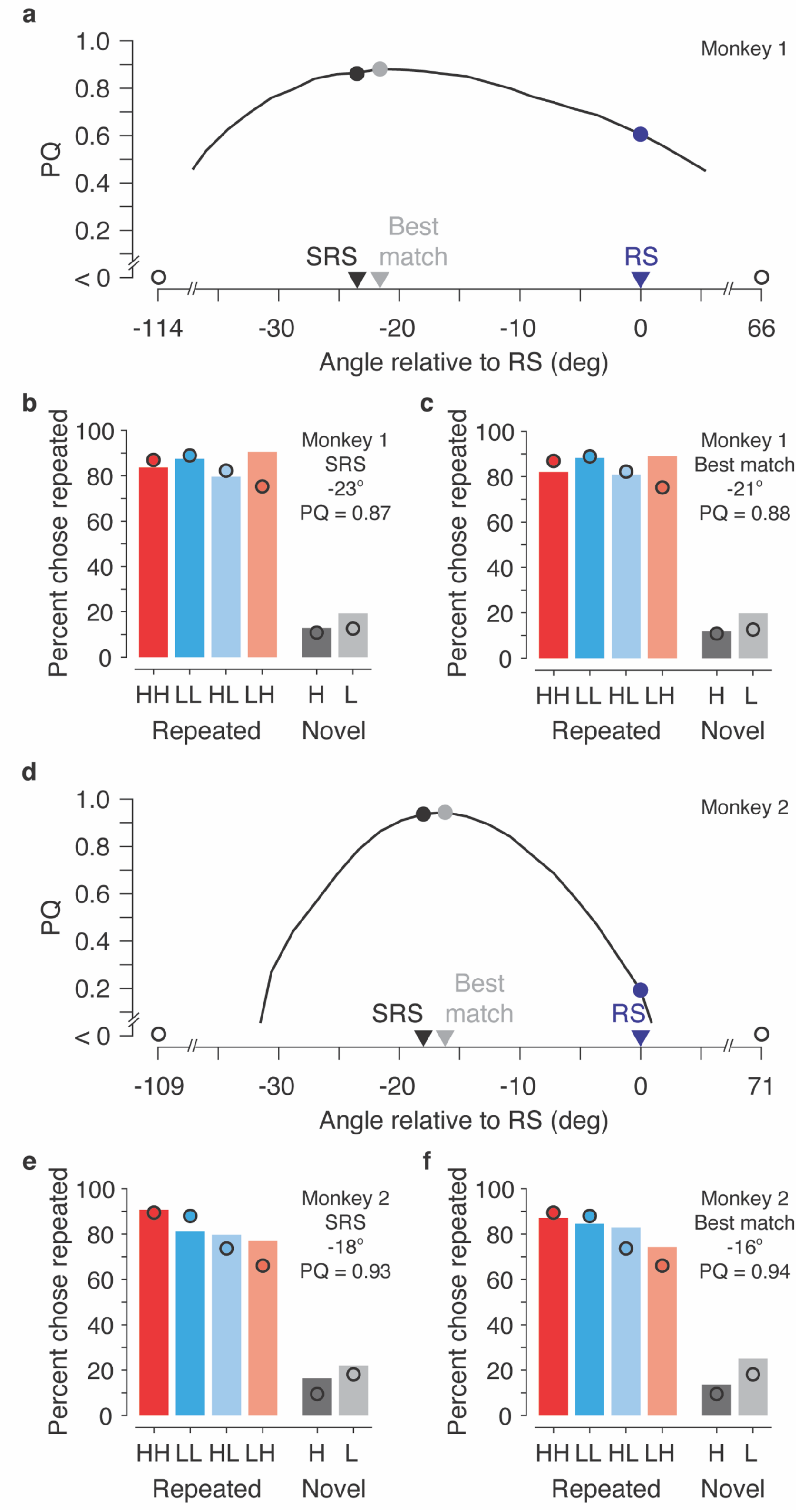

Fig. S4. *Neural predictions of behavior for a family of weighted linear decoders that include SRS, plotted for each monkey.* **(a, d)** Fig. 3b, plotted for each animal: prediction quality (PQ) for the family of linear decoders that lie on the plane spanned by RS and contrast axis (Fig. 3a). Markers correspond to SRS

(black), RS (blue), and the linear decoder with largest PQ (grey). **(b, e)** Fig. 3c, plotted for each animal: the alignment of the actual behavioral pattern and the SRS prediction **(c, f)** Fig. 3d, plotted for each animal: the alignment of the actual behavioral pattern and the decoder with the highest PQ on this plane. In b-f, dots indicate actual behavioral patterns and bars indicate the linearly decoded neural predictions of behavior. The decoder's direction relative to RS, and prediction quality (PQ) are labeled for each case.

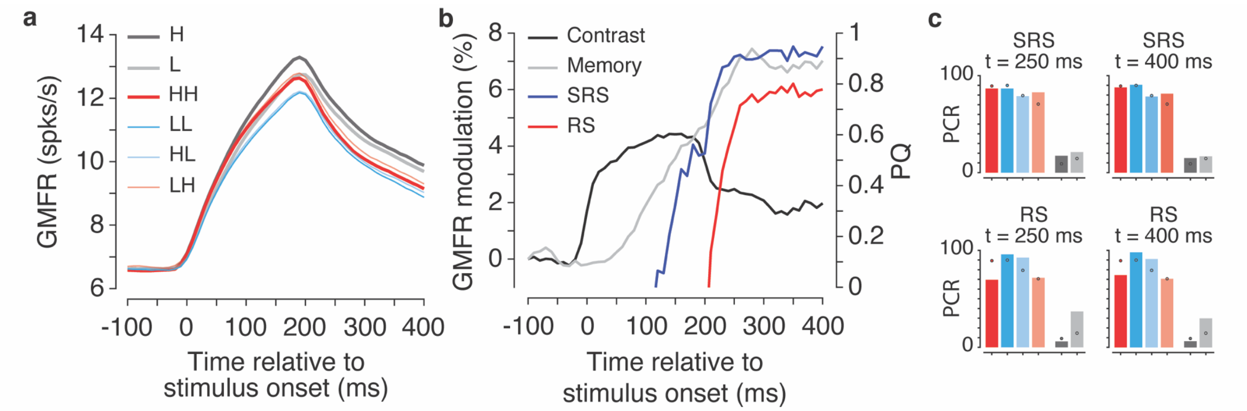

Fig. S5. *Temporal evolution of IT contrast and memory representations and their impact on decoding image memory.* **(a)** Grand mean firing rate (GMFR) of IT neurons as a function of time relative to stimulus onset. Traces in different colors correspond to different conditions (n = 856 units). **(b)** The evolution of contrast modulation (black) and memory modulation (grey), plotted along the left y-axis. We used traces depicted by thick lines in (a) to compute the modulations, such that: Contrast $= 100 \times \frac{\text{GMFR}_H - \text{GMFR}_L}{\text{GMFR}_H}$, and $\text{Memory} = 100 \times \frac{\text{GMFR}_H - \text{GMFR}_{HH}}{\text{GMFR}_H}$, where GMFR denotes grand mean firing rate in the subscripted condition. The time-course of prediction quality (PQ) for RS (red) and SRS (blue) decoders are also shown, plotted along the right y-axis. **(c)** The behavioral pattern predicted by RS (bottom row) and SRS (top row) decoders for two time points: t = 250 ms (first column), and t = 400 ms (second column). To perform these analyses, spikes were counted in 200 ms bins with bin centers that were shifted by 10ms.

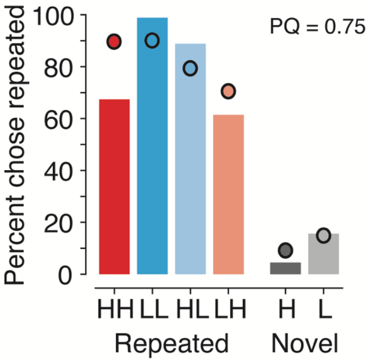

Fig. S6. *Variant decoder that incorporates a contrast correction*. Shown are the results for a decoder that operates by correcting the projection of IT responses along the RS axis by subtracting an estimate of the mean population response across novel images in each contrast condition (see Methods). The estimate of the mean population response at each contrast was computed after classifying novel images by contrast based on the training data, using the same prototype linear decoder used for SRS (see Methods). This scheme produced a lower prediction quality (PQ = 0.75) than SRS (PQ = 0.88; Fig. 3c).

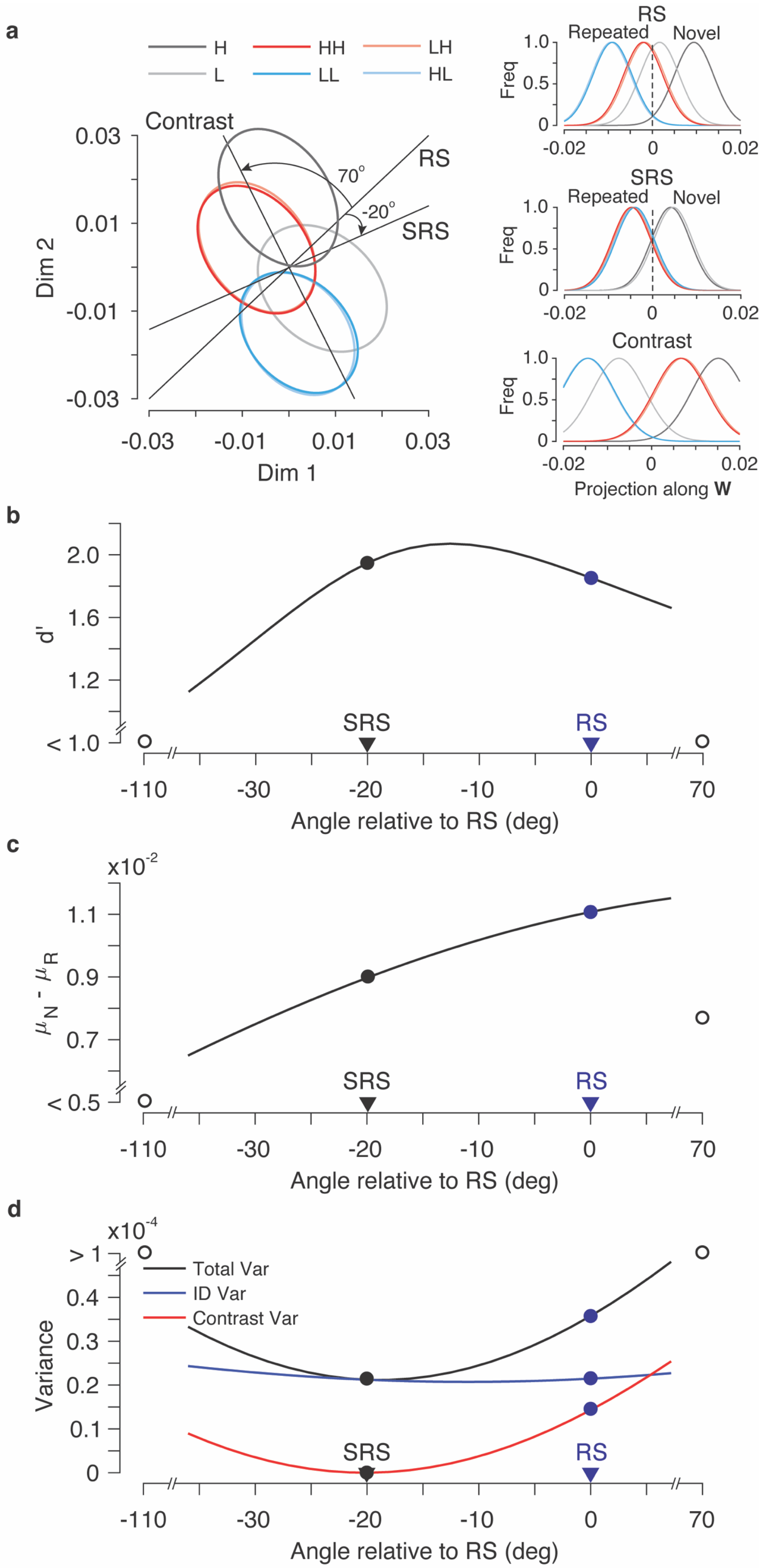
a
H
L
HH
LL
LH
HL
Contrast
RS
SRS
70°
-20°
0.03
0.01
-0.01
-0.03
Dim 2
-0.03
-0.01
0.01
0.03
Dim 1
RS
Repeated
Novel
Freq
1.0
0.5
0
-0.02
0
0.02
SRS
Repeated
Novel
Freq
1.0
0.5
0
-0.02
0
0.02
Contrast
Freq
1.0
0.5
0
-0.02
0
0.02
Projection along **W**
b
d'
2.0
1.6
1.2
< 1.0
SRS
RS
-110
-30
-20
-10
0
70
Angle relative to RS (deg)
c
x10^{-2}
$\mu_N - \mu_R$
1.1
0.9
0.7
< 0.5
SRS
RS
-110
-30
-20
-10
0
70
Angle relative to RS (deg)
d
x10^{-4}
> 1
0.4
0.3
0.2
0.1
0
Variance
Total Var
ID Var
Contrast Var
SRS
RS
-110
-30
-20
-10
0
70
Angle relative to RS (deg)

Fig. S7. *Synthetic data generated from the 4-parameter model recapitulates the actual data.* Simulations were performed for 650 units x 4K images (1K images/condition). All analyses were performed in the same manner as described for the physiological data. Plotted for the synthetic data **(a)** Fig. 3a **(b-d)** Fig. 4b-d.

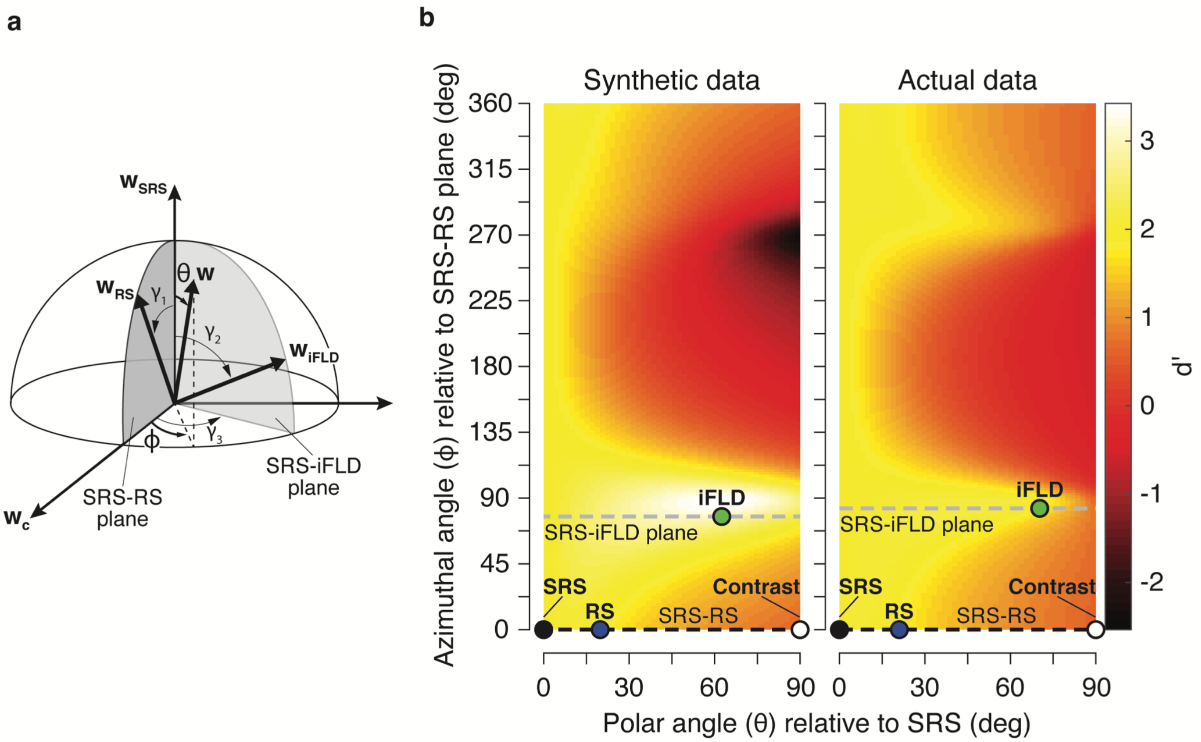

Fig. S8. *Decoding performance in the 3-D subspace spanned by the SRS, RS and iFLD linear decoders.* **(a)** Depiction of the 3-D subspace in a spherical coordinate system. A decoding axis (**w**) in this subspace is determined by the polar angle relative to SRS (θ) and the azimuthal angle (ϕ) relative to the contrast axis in the 2-D plane defined by SRS and RS. Because the 3-D subspace spanned by SRS, RS, and iFLD is a non-orthogonal coordinate system, we used angles $\boldsymbol{\gamma_1}$, $\boldsymbol{\gamma_2}$, and $\boldsymbol{\gamma_3}$ to transform the illustrated cartesian coordinate system to the coordinate system spanned by SRS, RS, and iFLD (see Methods). $\boldsymbol{\gamma_1}$ indicates the angle between SRS and RS, $\boldsymbol{\gamma_2}$ is the angle between SRS and iFLD, and $\boldsymbol{\gamma_3}$ represents the angle between SRS-RS and SRS-iFLD planes (see Methods). **(b)** Discriminability for image memory (d'), computed as described for Figs. 4b, 5b, and *SI Appendix,* Fig. S7b, plotted as a function of polar and azimuthal angles in the 3-D subspace (see Methods). Shown are the results applied to synthetic data taken from the model described for Fig. 5 and the actual data. Values corresponding to SRS, RS, iFLD, and contrast are labeled by black, blue, green, and open markers, respectively. Black and grey dashed lines mark the azimuthal angles associated with SRS-RS and SRS-iFLD planes, respectively.

SI References

1. Meyer, T. & Rust, N.C. Single-exposure visual memory judgments are reflected in inferotemporal cortex. *eLife* **7**, e32259 (2018).
2. Goris, R.L., Movshon, J.A. & Simoncelli, E.P. Partitioning neuronal variability. *Nat Neurosci* **17**, 858-865 (2014).