

1 **Title:** The integration of visual and target signals in V4 and IT during visual object search

2

3

4

5 **Authors:** Noam Roth and Nicole C. Rust

6 **Affiliation:** Department of Psychology, University of Pennsylvania, Philadelphia, PA USA

7

8

9

10 **Running title:** V4 and IT during visual object search

11

12

13

14

15

16 **Corresponding author:**

17 Nicole Rust

18 Department of Psychology

19 University of Pennsylvania

20 Goddard Labs, 428

21 Philadelphia, PA 19104

22 Phone: (215) 898-4587

23 Email: nrust@psych.upenn.edu

24

25

26

27

28

29

30

31

32 **Acknowledgements:** We thank Margot P. Wohl and Krystal Henderson for their technical
33 contributions. This work was supported by the National Eye Institute of the US National
34 Institutes of Health (award number R01EY020851), the Simons Foundation (through an award
35 from the Simons Collaboration on the Global Brain), and the McKnight Endowment for
36 Neuroscience.

37

38

48 **ABSTRACT**

49

50 Many everyday tasks require us to flexibly map incoming sensory information onto behavioral
51 responses based on context. One example is the act of searching for a specific object, which
52 requires our brain to compare the items in view with a remembered representation of a sought
53 target to determine whether a target match is present. During object search, this comparison is
54 thought to be implemented, in part, via the combination of top-down modulations reflecting
55 target identity with feed-forward visual representations. However, it remains unclear whether
56 these top-down signals are integrated at a single locus within the ventral visual pathway (e.g.
57 V4) or at multiple stages (e.g. both V4 and inferotemporal cortex, IT). To investigate, we
58 compared neural responses in V4 and IT recorded as rhesus monkeys performed a task that
59 required them to identify when a target object appeared across variation in position, size and
60 background context. We found non-visual, task-specific signals in both V4 and IT. To evaluate
61 the plausibility that V4 was the only locus for the integration of top-down signals, we evaluated a
62 number of feed-forward accounts of processing from V4 to IT, including a model in which IT
63 preferentially sampled from the best V4 units, as well as a model that allowed for nonlinear IT
64 computation. IT task-specific modulation could not be accounted for by any of these feed-
65 forward descriptions, suggesting that during object search, top-down signals are integrated
66 directly within IT itself.

67

68

69

70 **SIGNIFICANCE**

71

72 To find specific visual objects, the brain must combine top-down information reflecting the
73 identity of a sought target with visual information about objects in view. While top-down signals
74 are known to exist at multiple stages along the ventral visual pathway, the route with which they
75 arrive in each brain area is unclear. Here we present evidence that task-relevant signals in one
76 high-level visual brain area, IT, cannot be described as simply being inherited from an earlier
77 stage of processing, V4, and thus must be integrated directly within IT itself. This study is the
78 first to systematically compare the responses of V4 and IT during an object search task in which
79 objects can appear in different real-world configurations, and it provides important constraints on
80 the neural computations responsible for finding visual targets.

81

82

83 INTRODUCTION

84

85 Finding a sought object, such as our car keys, requires our brains to perform at least two non-
86 trivial computations. First, we must determine the identities of the objects in view, across
87 variation in details such as their position, size, and background context. Second, we must
88 compare this visual representation (of what we are looking at) with a remembered
89 representation (of what we are looking for) to determine whether our target is in view.
90 Considerable evidence suggests that computations in the primate ventral visual pathway,
91 including brain areas V1, V2, V4 and IT, support the process of invariant object recognition
92 (reviewed by DiCarlo et al. 2012). Within V4 and IT, many neurons are also modulated by
93 information about target identity as well as whether an image is a target match (Haenny et al.
94 1988; Maunsell et al. 1991; Eskandar et al. 1992; Leuschow et al. 1994; Gibson and Maunsell
95 1997; Chelazzi et al. 1998; Chelazzi et al. 2001; Bichot et al. 2005; Pagan et al. 2013; Kosai et
96 al. 2014; Roth and Rust 2018). However, the route by which these signals arrive in V4 and IT
97 remains unclear.

98

99 Here we present two classes of proposals describing how top-down signals reflecting the
100 identity of a sought target and/or whether the object in view is a target match might arrive within
101 V4 and IT during target search. In the first (Fig 1a), V4 serves as the sole locus of the
102 combination of visual and top-down information, and IT receives this information via feed-
103 forward propagation from V4. In the second (Fig 1b), top-down information is integrated directly
104 in IT. This class includes proposals in which top-down information is integrated in both V4 and
105 IT (Fig 1b, left) as well as proposals in which IT serves as the sole locus for the integration of
106 top-down information, and V4 receives this information from IT through feedback (Fig 1b, right).

107

108

109

110 **Figure 1.** *Proposals for how top-down target and/or target match information might arrive within*
111 *V4 and IT during object search. a)* The class of “IT: inherited” proposals predict that top-down
112 information is integrated only in V4, and this information is then inherited by IT via feed-forward
113 propagation. *b)* The class of “IT: integrated” proposals predict that top-down information is
114 integrated directly in IT. This class includes proposals in which top-down information is
115 integrated in both V4 and IT (*left*) as well as proposals in which top-down information is
116 integrated exclusively in IT but is then fed-back to V4 (*right*).

117

118

119

120 At least some evidence exists to support all of the proposals presented in Figure 1, albeit
121 sometimes indirect. Support for multi-locus descriptions (Fig 1b, left) comes from studies
122 reporting that non-visual, task-relevant signals increase in a gradient-like fashion across the
123 early visual hierarchy (i.e. V1, V2 and V4) during covert spatial attention and feature-based
124 attention tasks (reviewed by Noudoost et al. 2010), consistent with the integration of top-down
125 signals at multiple stages. By extension, top-down signals could be integrated in both V4 and IT
126 during visual target search. Importantly, if a gradient of top-down modulation were to exist
127 between V4 and IT, this would not necessarily imply multiple stages of top-down integration, as
128 a gradient is also consistent with top-down integration in IT followed by feedback to V4 (Fig 1b,
129 right). Evidence supports this scheme in V1, V2 and V4 (Buffalo et al. 2010).

130

131 In contrast to the proposals presented in Fig 1b, proposals in which V4 serves as the sole locus
132 of top-down integration (Fig 1a) predict matched amounts of non-visual, task-relevant
133 modulation between V4 and IT, and the few studies that have measured it are most consistent
134 with this prediction, both during visual target search (Chelazzi et al. 1998; Chelazzi et al. 2001)
135 as well as one covert spatial attention task (Moran and Desimone 1985). Additional support for
136 a single locus of top-down integration within the ventral visual pathway comes from
137 comparisons of target match signals in IT and a stage of processing just beyond it, perirhinal
138 cortex, where perirhinal target match information is reported to be well-accounted for via purely
139 feed-forward input from IT (Pagan et al. 2013; Pagan and Rust 2014; Pagan et al. 2016).

140

141 Here we focus on differentiating between the proposals presented in Figure 1 by probing the
142 responses of V4 and IT neurons during a visual object search task that capitalizes on
143 differences in how V4 and IT represent object identity across identity-preserving
144 transformations.

145 **METHODS**

146

147 **Experimental Design**

148

149 Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta*)
150 with implanted head posts and recording chambers. All procedures were performed in
151 accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and
152 Use Committee. A portion of the data recorded in one brain area (IT) was presented in an earlier
153 report (Roth and Rust 2018).

154

155 **The invariant delayed-match-to-sample (IDMS) task**

156

157 Monkey behavioral training and testing utilized standard operant conditioning, head stabilization
158 and infrared video eye tracking. Custom software (<http://mworks-project.org>) was used to
159 present stimuli on an LCD monitor with an 85 Hz refresh rate.

160

161 The monkeys performed an invariant delayed-match-to-sample task (Fig 2). As an overview, the
162 task required the monkeys to make a saccade when a target object appeared within a sequence
163 of distractor images (Fig 2a). Objects were presented at differing positions, sizes and
164 background contexts (Fig 2b). Stimuli consisted of a fixed set of 20 images that included 4 target
165 objects, each presented at 5 different identity-preserving transformations (Fig 2c). Each short
166 block (~3 min) was run with a fixed target object before another target was pseudorandomly
167 selected. Our design included two types of trials: cue trials and test trials (Fig 2a). Only test
168 trials were analyzed for this report.

169

170 A trial began when the monkey fixated on a red dot (0.15°) in the center of a gray screen, within
171 a square window of $\pm 1.5^\circ$. Fixation was followed by a 250 ms delay before a stimulus appeared.
172 Cue trials, which indicated the current target object, were presented at the beginning of each
173 short block or after three subsequent error trials. To minimize confusion, cue trials were
174 designed to be distinct from test trials and began with the presentation of an image of each
175 object that was distinct from the images used on test trials (a large version of the object
176 presented at the center of gaze on a gray background; Fig 2a). Test trials began with a
177 distractor image, and neural responses to the first distractor were discarded to minimize non-
178 stationarities such as stimulus onset effects. During the DMS task, all images were presented at
179 the center of gaze, in a circular aperture that blended into a gray background (Fig 2b).

180

181 In each block, 5 images were presented as target matches and the other 15 as distractors.
182 Distractor images were drawn randomly without replacement until each distractor was presented
183 once on a correct trial, and the images were then re-randomized. On most test trials, a target
184 match followed the presentation of a random number of 1-6 distractors (Fig 2a). On a small
185 fraction of trials, 7 distractors were shown, and the monkeys were rewarded for fixating through
186 all distractors. Each image was presented for 400 ms (or until the monkeys' eyes left the fixation
187 window) and was immediately followed by the presentation of the next stimulus. Monkeys were
188 rewarded for making a saccade to a response target within a window of 75 – 600 ms after the
189 target match onset. In monkey 1, the response target was positioned 10 degrees below fixation;
190 in monkey 2 it was 10 degrees above fixation. If the monkeys had not yet moved their eyes after
191 400 ms following target onset, a distractor stimulus was immediately presented. A trial was
192 classified as a 'false alarm' if the eyes left the fixation window via the top (monkey 1) or bottom

193 (monkey 2) outside the allowable correct response period and travelled more than 0.5 degrees.
194 In contrast, all other instances in which the eyes left the fixation window during the presentation
195 of distractors were characterized as fixation breaks. A trial was classified as a 'miss' when the
196 monkey continued fixating beyond 600 ms following the onset of the target match. Within each
197 block, 4 repeated presentations of each of the 20 images were collected, and a new target
198 object was then pseudorandomly selected. Following the presentation of all 4 objects as targets,
199 the targets were re-randomized. At least 10 repeats of each condition were collected on correct
200 trials. When more than 10 repeats were collected, the first 10 were used for analysis. Overall,
201 monkeys performed this task with high accuracy. Disregarding fixation breaks (monkey 1: 8% of
202 trials, monkey 2: 11% of trials), percent correct on the remaining trials was: monkey 1: 94%
203 correct, 2% false alarms, and 4% misses; monkey 2: 98% correct, ~1% false alarms, and ~1%
204 misses. Behavioral performance was comparable for the sessions corresponding to recordings
205 from the two areas (V4 percent correct overall = 96.5%; IT percent correct overall = 91.4%).
206

207 V4 receptive fields at and near the center of gaze are small: on average they have radii of 0.56
208 degrees at the fovea, extending to radii of 1.4 at an eccentricity of 2.5 degrees (Desimone and
209 Schein 1987; Gattass et al. 1988). We thus took considerable care to ensure that that the
210 images were approximately placed in the same region of these receptive fields across repeated
211 trials. In one monkey, fixational control was good after training (on average 85 and 97% of
212 presentations occurred within a radius of 0.56 and 1.4 degrees respectively). In a second
213 monkey, adequate fixational control could not be achieved through training. We thus applied a
214 procedure in which we shifted each image at stimulus onset 25% toward the center of gaze (e.g.
215 if the eyes were displaced 0.5 degrees to the left, the image was repositioned 0.125 degrees to
216 the left and thus 0.375 degrees from fixation). Image position then remained fixed until the onset
217 of the next stimulus. The resulting deviation across trials, measured relative to the mean
218 position across trials, was comparable to monkey 1: on average, 95, and 99% of presentations
219 occurred within windows with a radius of 0.56 and 1.4 degrees, respectively.

220

221 **Neural recording**

222

223 The activity of neurons in V4 and IT was recorded via a single recording chamber for each brain
224 area in each monkey. In both monkeys, chamber implantation and recording in IT preceded V4,
225 and the IT recording chamber was implanted on the right hemisphere whereas the V4 recording
226 chamber was implanted on the left hemisphere. While IT receptive fields span the vertical
227 meridian, thus allowing us to access the visual representation of both sides with a single
228 chamber, V4 receptive fields are confined to the contralateral hemifield. To simulate V4
229 coverage of the ipsilateral visual field, on roughly half of the V4 recording sessions, (n = 7/15
230 sessions in Monkey 1, n = 11/20 sessions in Monkey 2), we presented the images reflected
231 across the vertical axis. We then treated all V4 units recorded during these sessions as if they
232 were in the left hemisphere (and thus as receptive fields that were located in the right visual
233 field).
234

235 Chamber placement for the IT chambers was guided by anatomical magnetic resonance images
236 in both monkeys, and in one monkey, Brainsight neuronavigation ([https://www.rogue-](https://www.rogue-research.com/)
237 [research.com/](https://www.rogue-research.com/)). Both V4 chambers were guided by Brainsight neuronavigation. The region of IT
238 recorded was located on the ventral surface of the brain, over an area that spanned 4 mm
239 lateral to the anterior middle temporal sulcus and 15-19 mm anterior to the ear canals. Both V4
240 chambers were centered 1 mm posterior to the ear canals and 29 mm lateral to the midline,

241 positioned at a 30 degree angle. V4 recording sites were confirmed by a combination of
242 receptive field location and position in the chamber, corresponding to results reported previously
243 (Gattass et al. 1988). Specifically, we recorded from units within and around the inferior occipital
244 sulcus, between the lunate sulcus and superior temporal sulcus. V4 units in lower visual field
245 were confirmed as having receptive field centers that traversed from the vertical to horizontal
246 meridian as recordings shifted from posterior to anterior. As expected, V4 units in the fovea and
247 near the upper visual field were found lateral to those in the lower visual field, and had receptive
248 field centers that traversed from the horizontal meridian to the vertical meridian as recordings
249 traversed medial to lateral and increased in depth.

250
251 Neural activity was recorded with 24-channel U-probes and V-probes (Plexon, Inc) with linearly
252 arranged recording sites spaced with 100 mm intervals. Continuous, wideband neural signals
253 were amplified, digitized at 40 kHz and stored using the OmniPlex Data Acquisition System
254 (Plexon, Inc.). Spike sorting was done manually offline (Plexon Offline Sorter). At least one
255 candidate unit was identified on each recording channel, and 2-3 units were occasionally
256 identified on the same channel. Spike sorting was performed blind to any experimental
257 conditions to avoid bias. A multi-channel recording session was included in the analysis if the
258 animal performed the task until the completion of at least 10 correct trials per stimulus condition,
259 there was no external noise source confounding the detection of spike waveforms, and the
260 session included a threshold number of task-modulated units (>4 on 24 channels). The sample
261 size for IT (number of units recorded) was chosen to approximately match our previous work
262 (Pagan et al. 2013; Pagan and Rust 2014). The sample size for V4, was selected to be 3-fold
263 that number, to match the ratio between numbers of units estimated in V4 as compared to IT
264 (DiCarlo et al. 2012).

265
266 For many of the analyses presented in this paper, we measured neural responses by counting
267 spikes in a window that began 40 ms after stimulus onset in V4 and 80 ms after stimulus onset
268 in IT. We counted spikes in a 170 ms window in both areas, such that the spike counting
269 windows were of equal length. Counting windows always preceded the monkeys' reaction times.
270 On 7.7% of all correct target match presentations, the monkeys had reaction times faster than
271 250 ms, and those instances were excluded from analysis to ensure that spikes in both V4 and
272 IT were only counted during periods of fixation.

273
274 In IT, we recorded neural responses across 20 experimental sessions (Monkey 1: 10 sessions,
275 and Monkey 2: 10 sessions). In V4, we recorded neural responses across 35 experimental
276 sessions (Monkey 1: 15 sessions, and Monkey 2: 20 sessions). When combining the units
277 recorded across sessions into a larger pseudopopulation, we began by screening for units that
278 met three criteria. First, units needed to be modulated by our task, as quantified by a one-way
279 ANOVA applied to our neural responses (80 conditions * 10 repeats, $p < 0.01$). Second, units
280 needed to pass a loose criterion on recording stability, as quantified by calculating the variance-
281 to-mean ratio (Fano factor) for each unit, computed by fitting the relationship between the mean
282 and variance of spike count across the 80 conditions (Fano factor < 2.5). Finally, units needed
283 to pass a loose criterion on unit recording isolation, quantified by calculating the signal-to-noise
284 ratio (SNR) of the waveform as the difference between the maximum and minimum points of the
285 average waveform, divided by twice the standard deviation across the differences between each
286 waveform and the mean waveform (SNR > 2). In IT, this yielded a pseudopopulation of 193
287 units (of 563 possible units), including 98 units from monkey 1 and 95 units from monkey 2. In
288 V4, this yielded a pseudopopulation of 598 units (of 970 possible units), including 345 units from
289 monkey 1 and 253 units from monkey 2.

290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336

V4 receptive field mapping

To measure the location and extent of V4 receptive fields, bars were presented for 500 ms, one per trial, centered on a 5 x 5 invisible grid. Bar orientation, length, and width as well as the grid center and extent were adjusted for each recording session based on preliminary hand mapping. On each trial, the monkey was required to maintain fixation on a small response dot (0.125°) to receive a reward. The responses to at least five repeats were collected at each position for each recording session. Only those units that produced clear visually evoked responses at a minimum of one position were considered for receptive field position analysis. The center of the receptive field was estimated by the maximum of the response across the 5x5 grid of oriented bar stimuli and confirmed by visual inspection.

Quantifying single-unit modulations

To quantify the degree to which individual V4 and IT units were modulated by task-relevant variables (Figs 4, 7, 8), such as changes in visual and target identity, we applied a bias-corrected, ANOVA-like procedure described in detail by (Pagan and Rust 2014) and summarized here. As an overview, this procedure is designed to parse each unit's total response variance into variance that can be attributed to each type of experimental parameter as well as variance that can be attributed to trial variability. Total variance is computed across the spike count responses for each unit across 16 conditions (4 images * 4 targets for each transformation) and 10 trials. Variances are then transformed into measures of spike count modulation (in the units of standard deviation around each unit's grand mean spike count) via a procedure that includes bias correction for over-estimates in modulation due to noise.

To capture all types of modulation with intuitive groupings, the procedure begins by developing an orthonormal basis of 16 vectors. The number of basis vectors for each type of modulation is imposed by the experimental design. In particular, this basis \mathbf{b} included vectors \mathbf{b}_i that reflected 1) the grand mean spike count across all conditions, 2) whether the object in view was a target or a distractor ('target match'), 3) visual image identity ('visual'), 4) target object identity ('target identity'), and 5) nonlinear interactions between target and object identity not captured by target match modulation ('residual'). The initially designed set of vectors is converted into an orthonormal basis via a Gram-Schmidt orthogonalization process.

The resulting basis spans the space of all possible responses for our task. Consequently, we can re-express each trial-averaged vector of spike count responses to the 16 experimental conditions for each transformation, \mathbf{R} , as a weighted sum of these basis vectors. The weight corresponding to a basis vector for each unit reflect modulation of that unit's responses by that experimental parameter. To quantify the amounts of each type of modulation reflected by each unit, we began by computing the squared projection of each basis vector \mathbf{b}_i and \mathbf{R} . To correct for bias caused by over-estimates in modulation due to noise, an analytical bias correction, described and verified in (Pagan and Rust 2014), was then subtracted from this value. The squared weight for each basis vector \mathbf{b}_i is calculated as:

$$(1) w_i^2 = (\mathbf{R} \cdot \mathbf{b}_i^T)^2 - \frac{\sigma_{\epsilon}^2 \cdot (\mathbf{b}_i^T)^2}{m}$$

337 where σ_t^2 indicates the trial variance, averaged across conditions (n=16), and m indicates the
338 number of trials (m=10). If more than one dimension existed for a type of modulation, we
339 summed values of the same type (eq. 2). Next, we applied a normalization factor (1/(n-1)) where
340 n=16) to convert these summed values into variances. As a final step, we computed the square
341 root of these quantities to convert them into modulation measures that reflected the number of
342 spike count standard deviations around each unit's grand mean spike count. Modulation for
343 each parameter type X was thus computed as:

$$344 \quad (2) \sigma_X = \sqrt{\frac{1}{n-1} \cdot \sum_{i=j}^k w_i^2}$$

345 for the weights w_j through w_k corresponding to basis vectors \mathbf{b}_j through \mathbf{b}_k for that parameter
346 type, where the number of basis vectors corresponding to each parameter type were: target
347 match = 1; visual = 3; target identity = 3; residual = 8.

348 When estimating modulation for individual units, (Fig 4), the bias-corrected squared values were
349 rectified for each unit before taking the square root. When estimating modulation population
350 means (Fig 7b-e, Fig 8), the bias-corrected squared values were averaged across units before
351 taking the square root. Because these measures were not normally distributed, standard error
352 about the mean was computed via a bootstrap procedure. On each iteration of the bootstrap
353 (across 1000 iterations), we randomly sampled values from the modulation values for each unit
354 in the population, with replacement. Standard error was computed as the standard deviation
355 across the means of these resampled populations.

356 **Population performance: Visual object invariance**

357 To determine performance of the V4 and IT populations at classifying visual object identity (Fig
358 5), we computed 4-way object discrimination performance. As an overview, we formulated the
359 problem as four one-versus-rest linear classifications, and then took the maximum of these
360 classifications as a population's decision (Hung et al. 2005). Here we begin by describing the
361 general form of linear classifier that we used, a Fisher Linear Discriminant (FLD), and we then
362 describe the training and testing scheme for measuring cross-validated performance.

363 The general form of a linear decoding axis is:

$$364 \quad (3) f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b,$$

365 where \mathbf{w} is an N-dimensional vector containing the linear weights applied to each of N units, and
366 b is a scalar value. We fit these parameters using an FLD, where the vector of linear weights
367 was calculated as:

$$368 \quad (4) \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$$

369 and b was calculated as:

$$370 \quad (5) b = \mathbf{w} \cdot \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2$$

371 Here μ_1 and μ_2 are the means of two classes (e.g. two object classes, respectively) and the
372 mean covariance matrix is calculated as:

376 (6) $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$

377 where Σ_1 and Σ_2 are the regularized covariance matrices of the two classes. These covariance
378 matrices were computed using a regularized estimate equal to a linear combination of the
379 sample covariance and the identity matrix I (Pagan and Rust 2014):

380 (7) $\Sigma_i = \gamma \Sigma_i + (1 - \gamma) \cdot I$

381 We determined γ by exploring a range of values from 0.01 to 0.99, and we selected the value
382 that maximized average performance across all iterations, measured with the cross-validation
383 “regularization” trials set aside for this purpose (see below). We then computed performance for
384 that value of γ with separately measured “test” trials, to ensure a fully cross-validated measure.
385 Because this calculation of the FLD parameters incorporates the off-diagonal terms of the
386 covariance matrix, FLD weights are optimized for both the information conveyed by individual
387 units as well as their pairwise interactions.

388 To classify which of four objects was in view, we used a standard “one-versus-rest”
389 classification scheme. Specifically, one linear classifier was determined for each object based
390 on the training data. To determine the population decision about which object was presented, a
391 response vector \mathbf{x} , corresponding to the population response of one of the four objects, was
392 then applied to each of the classifiers, and the classifier with the largest output (the classifier
393 with the largest, positive $f(\mathbf{x})$) was taken as the population decision. To train the classifiers, we
394 used an iterative resampling procedure. On each iteration of the resampling, we randomly
395 shuffled the trials for each condition and for each unit, and (for numbers of units less than the
396 full population size) randomly selected units. On each iteration, 8 trials from each condition were
397 used for training the decoder, 1 trial from each condition was used to determine a value for
398 regularization, and 1 trial from each condition was used for cross-validated measurement of
399 performance.

400
401 We compared classifier performance for the “reference” cases (when cross-validated test trials
402 were selected from the same transformation used to train the classifier; Fig 5a-b, black) versus
403 the “generalization” cases (when test trials were selected from transformations different than the
404 one used for training, Fig 5a-b, cyan). To summarize the results for a given transformation,
405 reference and generalization performance was compared for the same test data: e.g. In the
406 case of the transformation “Up”, reference performance was computed by training and cross-
407 validated testing on “Up” and generalization performance was computed as the average of
408 training on all other transformations and testing on “Up”.

409
410 To ensure that visual classification performance was not biased by the target match signal, we
411 computed performance for targets and distractors separately and averaged their results.
412 Specifically, we computed visual classification performance for the four objects presented as
413 target matches or for different combinations of the four objects presented as distractors. Each
414 set of 4 distractors was selected to span all possible combinations of mismatched object and
415 target identities (e.g. objects 1, 2, 3, 4 paired with targets 4, 3, 2, 1), of which there are 9
416 possible sets. As a final measure of visual classification performance, we averaged across 10
417 performance values (1 target match and 9 distractor combinations) as well as, when relevant,
418 multiple transformations. One performance value was computed on each iteration of the
419 resampling procedure, and mean and standard error of performance was computed as the
420 mean and standard deviation of performance across 1000 resampling iterations. Standard error

421 thus reflected the variability due to the specific trials assigned to training and testing and, for
422 populations smaller than the full size, the specific units chosen. Finally, generalization capacity
423 was computed on each resampling iteration by taking the ratio of the chance-subtracted
424 reference performance and the chance-subtracted generalization performance (where chance =
425 25%).

426 **Population performance: Target match information**

428
429 To determine the ability of the V4 and IT populations to classify target matches versus
430 distractors (Figs 9 & 10), we applied two types of decoders: a linear classifier (an FLD,
431 described above) and a Maximum Likelihood decoder (a decoder that can classify based on
432 linear as well as nonlinearly formatted target match information). Both decoders were cross-
433 validated with the same resampling procedure. On each iteration of the resampling, we
434 randomly shuffled the trials for each condition and for each unit, and (for numbers of units less
435 than the full population size) randomly selected units (with the exception of Fig 9c, cyan, where
436 we selected the ‘best’ units, as described below). On each iteration, 8 trials from each condition
437 were used for training the decoder, 1 trial from each condition was used to determine a value for
438 regularization of the FLD linear classifier (see below) and 1 trial from each condition was used
439 for a cross-validated measurement of performance.

440
441 To circumvent issues related to the format of visual information, classifier analyses were
442 performed per transformation (“Big”, “Up”, “Left” and “Small”). The data for each transformation
443 consisted of 16 conditions (4 visual objects viewed under 4 different target contexts). To ensure
444 that decoder performance relied only on target match information and not on other factors, such
445 as differences in the numbers of each class, each classification was computed for 4 target
446 matches versus 4 (of 12 possible) distractors. Each set of 4 distractors was selected to span all
447 possible combinations of mismatched object and target identities (e.g. objects 1, 2, 3, 4 paired
448 with targets 4, 3, 2, 1), of which there are 9 possible sets. Performance was computed on each
449 resampling iteration by averaging the binary performance outcomes across the 9 possible sets
450 of target matches and distractors, each which contained 8 cross-validated test trials, and across
451 the four transformations used. For both types of classifiers, mean and standard error of
452 performance was computed as the mean and standard deviation of performance across 1000
453 resampling iterations. Standard error thus reflected the variability due to the specific trials
454 assigned to training and testing and, for populations smaller than the full size, the specific units
455 chosen.

456
457 To compute linear classifier performance (Fig 9), we used a 2-way Fisher Linear Discriminant,
458 described as in the general form above. In this case, the classes described in eqs. 4-6
459 correspond to target matches and distractors. To compute neural population performance, we
460 began by computing the dot product of the test data and the linear weights \mathbf{w} , adjusted by b (Eq.
461 5). Each test trial was then assigned to one class, and proportion correct was then computed as
462 the fraction of test trials that were correctly assigned, according to their true labels. To compute
463 linear classifier performance for the best V4 units (Fig 9c, cyan), we ranked units by their d'
464 based on the training data and sub-selected top-ranked units to measure cross-validated
465 performance. Unit d' was computed as:

466
467
$$(8) \ d' = \frac{|\mu_{Match} - \mu_{Distractor}|}{\sigma_{pooled}},$$

468 where μ_{Match} and $\mu_{Distractor}$ correspond to the mean across the set of target match and
469 distractors, $\sigma_{pooled} = \sqrt{\frac{\sigma_{Match}^2 + \sigma_{Distractor}^2}{2}}$, and σ_{Match} and $\sigma_{Distractor}$ correspond to the standard
470 deviation across the set of target matches and distractors, respectively.

471
472 As a measure of total target match information (Fig 10; combined linear and nonlinear), we
473 implemented a maximum likelihood decoder (Pagan et al. 2013; Pagan et al. 2016). We began
474 by using the set of training trials to compute the average response r_{uc} of each unit u to each of
475 the 2 conditions c (target matches versus distractors). We then computed the likelihood that a
476 test response k was generated from a particular condition as a Poisson-distributed variable:

477
$$(9) \text{lik}_{u,c}(k) = \frac{(r_{uc})^k \cdot e^{-r_{uc}}}{k!}$$

478 The likelihood that a population response vector was generated in response to each condition
479 was then computed as the product of the likelihoods of the individual units. We assigned the
480 population response to the category with the maximum likelihood, and we computed
481 performance as the fraction of trials in which the classification was correct based on the true
482 labels of the test data.

483
484

485 **Statistical analysis**

486
487 Because our measures were not normally distributed, we computed P values via resampling
488 procedures. When comparing the magnitudes of single unit modulation values between V4 and
489 IT (Fig 4, Fig 7b-e, Fig 8), a bootstrap procedure was applied in which values were randomly
490 sampled from the values for each unit, with replacement, across many iterations. We calculated
491 P values as the fraction of resampling iterations on which the difference was flipped in sign
492 relative to the actual difference between the means of the full data set (for example, if the mean
493 of visual modulation in V4 was larger than the mean of visual modulation in IT, the fraction of
494 iterations in which the mean of visual modulation in IT was larger than the mean of visual
495 modulation in V4).

496
497 When comparing generalization capacity between the V4 and IT populations (Fig 5d), we began
498 by computing generalization capacity for each of 1000 resampling iterations of the reference
499 and generalization classifiers. We calculated P values as the fraction of resampling iterations on
500 which the difference was flipped in sign relative to the actual difference between the means of
501 the full data set (for example, if the mean of generalization capacity in IT was larger than the
502 mean of generalization capacity in V4, the fraction of iterations in which the mean of
503 generalization capacity in V4 was larger than the mean of generalization capacity in IT).

504
505 When comparing population decoding measures (Figs 5a-c, 9c, & 10c), 1000 iterations of cross-
506 validated population performance were computed, and P values were calculated as the fraction
507 of classifier iterations on which the difference was flipped in sign relative to the actual difference
508 between the means across classifier iterations (for example, if the mean of decoding measure 1
509 was larger than the mean of decoding measure 2, the fraction of iterations in which the mean of
510 measure 2 was larger than the mean of measure 1). When evaluating whether a population
511 decoding measure was different from chance (Figs 9-10), P values were calculated as the

512 fraction of classifier iterations on which performance was greater than chance performance
513 (50%).
514
515
516

517 RESULTS

518
519 To compare responses in V4 and IT, we trained two monkeys to perform an “invariant delayed-
520 match-to-sample” (IDMS) task that required them to report when target objects appeared across
521 variation in the objects’ positions, sizes and background contexts. Some of the data presented
522 here were also included in an earlier publication (Roth and Rust 2018). There, we reported that
523 during IDMS, neural signals in IT reflected behavioral confusions on the trials in which the
524 monkeys made errors, and IT target match signals were configured in a manner that minimized
525 their interference with IT visual representations. The focus of the current report is a
526 determination of how these signals arrive in IT via a systematic comparison between IT and its
527 input brain area, V4.

528
529

530 The invariant delayed-match-to-sample task (IDMS)

531
532 Monkeys performed an invariant delayed-match-to-sample (IDMS) task in short blocks of trials
533 (~3 minutes on average) with a fixed target object. Each block began with a cue trial that
534 indicated the target for that block (Fig 2a ‘Cue Trial’). The remainder of the block was comprised
535 primarily of test trials (Fig 2a, ‘Test trial’). Test trials began with the presentation of a distractor
536 and on most trials, this was followed by 0-5 additional distractors (for a total of 1-6 distractor
537 images) and then an image containing the target match. The monkeys’ task required them to
538 maintain fixation during the presentation of distractors and make a saccade in response to the
539 appearance of a target match to receive a juice reward. To minimize the possibility that
540 monkeys would predict the target match, on a small fraction of the trials the target match did not
541 appear and the monkeys were rewarded for maintaining fixation through 7 distractors. Unlike
542 other classic DMS tasks (Eskandar et al. 1992; Chelazzi et al. 1993; Leuschow et al. 1994;
543 Miller and Desimone 1994; Pagan et al. 2013) our experimental design does not incorporate a
544 cue at the beginning of each test trial, to better mimic real-world object search, where target
545 matches are not repeats of the same image presented shortly before.

546
547
548

549 **Figure 2.** *The invariant delayed-match-to-sample task (IDMS).* **a)** Monkeys initiated trials by
550 fixating on a small dot. Each block (~3 minutes in duration) began with a cue trial which
551 indicated the target object. On subsequent trials, a random number (1-7) of distractors were
552 presented, and on most trials, this was followed by the target match. Monkeys were required to
553 maintain fixation throughout the distractors and make a saccade to a response dot within a
554 window 75 - 600 ms following the onset of the target match to receive a reward. In cases where
555 the target match was presented for 400 ms and the monkey had still not broken fixation, a
556 distractor stimulus was immediately presented. **b)** A schematic of the full experimental design,
557 which included 80 conditions: looking “at” each of 4 objects, each presented at 5 identity-
558 preserving transformations (for 20 images in total), viewed in the context of looking “for” each
559 object as a target. In this design, target matches (gray) fall along the diagonal of each “looking
560 at” / “looking for” transformation slice whereas distractors (white) fall off the diagonal. **c)** Images
561 used in the task: 4 objects were presented at each of 5 identity-preserving transformations (“up”,
562 “left”, “right”, “big”, “small”), for 20 images in total. In any given block, 5 of the images were
563 presented as target matches and 15 were distractors. **d)** Percent correct for each monkey,
564 calculated based on both misses and false alarms (but disregarding fixation breaks), shown as

565 a function of the number of distractors preceding the target match. Error bars indicate standard
566 error across experimental sessions. **e)** Histograms of reaction times during correct trials (ms
567 after stimulus onset), with means labeled.

568
569
570 Our experimental stimuli consisted of a fixed set of 20 images: 4 objects presented at each of 5
571 transformations (Fig 2b). These specific images were selected in order to make the task of
572 classifying object identity challenging for the IT population and these specific transformations
573 were selected based on findings from our previous work (Rust and DiCarlo 2010). In a given
574 target block (e.g. a ‘banana block’), a subset of 5 of the images were target matches and the
575 remaining 15 were distractors (Fig 2c). The full experimental design amounted to 20 images (4
576 objects presented at 5 identity-preserving transformations), all viewed in the context of each of
577 the 4 objects as a target, resulting in 80 experimental conditions (Fig 2b). In this design, “target
578 matches” fall along the diagonal of each looking at / looking for matrix slice (where a matrix
579 “slice” corresponds to the conditions at one fixed transformation; Fig 2b, gray). For each of the
580 80 conditions, we collected at least 10 repeats on correct trials. Behavioral performance was
581 high overall (Fig 2d). The monkeys’ mean reaction times (computed as the time their eyes left
582 the fixation window relative to the target match stimulus onset) were 311 ms and 363 ms for
583 monkey 1 and 2, respectively (Fig 2e).

584
585 To systematically compare the responses of V4 and IT during this task, we applied a population-
586 based approach in which we fixed the images and their placement in the visual field across all
587 the units that we studied, and we sampled from units whose receptive fields overlapped the
588 stimuli. Specifically, we presented images at the center of gaze, with a diameter of 5 degrees.
589 Neurons in IT typically have receptive fields that extend beyond 5 degrees and extend into all
590 four quadrants (Fig 3a top; Op De Beeck and Vogels 2000). In contrast, V4 receptive fields are
591 smaller, retinotopically organized, and confined to the contralateral hemifield (Fig 3a bottom;
592 Desimone and Schein 1987; Gattass et al. 1988). To compare these two brain areas, we
593 applied extensions of approaches developed in our earlier work in which we compared the
594 responses of a set of randomly sampled IT units with a population of V4 units whose receptive
595 fields tiled the image (Rust and DiCarlo 2010). This required sampling V4 units with receptive
596 fields in both upper and lower visual fields, which we achieved through recording at different
597 positions within and around the inferior occipital sulcus. This also required measuring units with
598 receptive fields on both sides of the vertical meridian, which we approximated by isolating our
599 recordings to one hemisphere but reflecting the images along the vertical axis in approximately
600 half the sessions.

601
602
603
604 **Figure 3. V4 and IT receptive field locations.** Images were displayed at the center of gaze and
605 were 5 degrees in diameter. Red circles indicate the location and size of the images. **a)**
606 Schematic of expected receptive field locations and sizes for neurons in IT (top; Op De Beeck
607 and Vogels 2000) V4 (bottom; Desimone and Schein 1987; Gattass et al. 1988). **b)** We
608 targeted V4 units with receptive fields that tiled the images. After approximate receptive field
609 localization with hand mapping, receptive field locations were determined with oriented bar
610 stimuli presented in a 5 x 5 grid of different positions (see Methods). Shown are the receptive
611 field centers of a subset of recorded V4 units; one dot is shown for each unique receptive field

612 location recorded. On approximately half of the sessions, images were reflected across the
613 vertical axis, and for these sessions, the receptive field centers are plotted in the ipsilateral
614 visual field. Monkey 1: gray; Monkey 2: white.

615

616
617

618 Because V4 receptive fields in the region of the field that we recorded are small, one issue of
619 concern is the replicability of retinal image placement across trials. We quantified the stability of
620 monkeys' eye positions across repeated trials as the percent of eye positions that were within
621 windows corresponding to V4 receptive field sizes at the range of eccentricities we recorded
622 (Gattass et al. 1988). We found that 89% of eye positions fell within windows corresponding to
623 the average RF sizes at the fovea (average foveal receptive field size = 0.56 degrees), and 98%
624 of eye positions were within windows corresponding to RF sizes at an eccentricity of 2.5
625 degrees (average receptive field size at 2.5 degrees = 1.4 degrees). To achieve this in Monkey
626 2, fixational control was improved by aligning the images closer to the center of gaze at stimulus
627 onset (see Methods). These approaches were effective in producing similar distributions of trial-
628 by-trial variability between V4 and IT, as measured by the mean and standard deviation of the
629 variance-to-mean ratio (Fano factor) across units (mean +/- std, V4 = 1.41 +/- 0.3; IT = 1.35 +/-
630 0.33).

631

632 As two monkeys performed this task, we recorded neural activity from small populations using
633 24-channel probes that were acutely lowered into V4 or IT before each session. With the
634 rationale that V4 contains approximately 3-fold more units than IT near the fovea (DiCarlo et al.
635 2012), we aimed to collect 3-fold more units from V4. Following a screen for units based on their
636 stability, isolation, and task modulation (see Methods), our data included 598 V4 units and 193
637 IT units (Monkey 1: 345 units in V4 and 98 in IT; Monkey 2: 253 units in V4 and 95 in IT). The
638 data reported here were extracted from trials with correct responses. For all analyses except Fig
639 7, we counted spikes in equal length windows in V4 and IT but adjusted for the difference in
640 latency between the two brain areas (170 ms, V4: 40-210 ms; IT: 80-250 ms following stimulus
641 onset). These windows always preceded the monkeys' reaction times and thus corresponded to
642 periods of fixation.

643

644

645 **Visual modulation as a benchmark for verifying V4 and IT data:**

646

647 When making systematic comparisons between V4 and IT, there are important factors to
648 consider. For example, should the information contained in the V4 and IT populations be
649 compared with equal numbers of units? Similarly, what are appropriate benchmarks for
650 determining whether the samples recorded from each brain area are representative? As an
651 example, imagine a scenario in which the same information about whether an image is a target
652 match or a distractor is reflected in both V4 and IT to the same degree, but the V4 neurons
653 recorded in an experiment all have small, overlapping receptive fields confined to the same,
654 small region of the visual field. In contrast, IT neurons, by virtue of their large receptive fields,
655 would have access to much more of the visual field. From this data we might erroneously find
656 that the magnitude of total target match information is larger in IT than V4 by way of non-
657 representative sampling.

658

659 As a benchmark for assessing whether the data we recorded from each brain area were
660 representative, we compared the amount of visual modulation present in each brain area, at
661 each transformation, with the following rationale. First, all the visual information contained in IT
662 is thought to arrive there after first travelling through V4 (Felleman and Van Essen 1991), and
663 consequently, samples of V4 and IT are comparable only if visual information is equal or higher
664 in the V4 sample. Second, comparisons of visual information at each transformation
665 independently circumvent issues related to well-established differences in the format of visual
666 information between the two brain areas: object identity (across changes in object position, size
667 and background context) is more accessible to a linear read-out in IT whereas it is more
668 nonlinear in V4 (e.g. Rust and DiCarlo 2010).

669
670 To compare the amounts of visual information in our recorded V4 and IT populations, we
671 computed a single-unit measure of visual modulation that disentangles modulations due to
672 changes in visual identity from other factors, such as top-down target modulation. This measure
673 quantifies the modulation in a unit's spike count that can be attributed to changes in the identity
674 of the object in view, computed separately for each of the 5 transformations. Specifically, the
675 analysis employs a bias-corrected procedure that quantifies different types of modulation in
676 terms of the number of standard deviations around each unit's grand mean spike count (Pagan
677 and Rust 2014). For three of the five transformations ('left', 'small', 'up'), mean visual modulation
678 was statistically indistinguishable between V4 and IT (Fig 4a-c). For one transformation ('big';
679 Fig 4d) mean visual modulation was larger in V4, but we retained this transformation for
680 subsequent analyses because its incorporation reflected a worst-case scenario against the
681 sampling problem of concern (i.e. one in which V4 has been inadequately sampled). In contrast,
682 for the final transformation ('right'; Fig 4f), the V4 population had significantly lower performance
683 than IT ($p < 1e10^{-5}$), and investigation of the recorded receptive field locations (Fig 3b) revealed
684 that this was likely due to incomplete sampling at that location. As such, we disregarded this
685 transformation from further analyses. Subsequent analyses are focused on the 4 of 5
686 transformations in which visual modulation, averaged across transformations, was not
687 statistically distinguishable in V4 as compared to IT, either in the pooled data or in either
688 monkey (Fig 4f; Monkey 1: V4 mean = 0.26, IT mean = 0.21, $p = 0.08$; Monkey 2: V4 mean =
689 0.16, IT mean = 0.17, $p = 0.53$). The fact that visual modulation is matched between V4 and IT
690 across these four transformations suggests that the two populations can and should be
691 compared with approximately matched numbers of units, consistent with previous reports (Rust
692 and DiCarlo 2010).

693
694

695 **Figure 4. Comparison of visual modulation in V4 and IT.** Shown are distributions of visual
696 modulation magnitudes across units, parsed by transformation for V4 (open bars, $n = 598$ units)
697 and IT (gray, $n = 193$ units) and plotted on a log axis. Following a bias correction to remove the
698 impact of trial variability, visual modulation was computed in units of standard deviation around
699 each unit's grand mean spike count. The first bin includes units with negligible visual modulation
700 (modulation < 0.001) and the broken axis indicates that these bars should extend to the
701 proportions labeled just above. Means of each distribution, including units with negligible visual
702 modulation, are indicated by arrows and values are indicated at the bottom of each panel. The
703 p-values at the top of each panel were computed via a bootstrap significance test evaluating the
704 probability that differences in the means between V4 and IT can be attributed to chance. **a-e)**
705 Distributions parsed by transformation. Visual modulation corresponding to the transformation
706 'right' was higher in IT as compared to V4, due to incomplete sampling of receptive fields at this

707 location (Fig 3b), and was thus disregarded from further analyses. **f)** Distributions of visual
708 modulation, averaged for each unit across the transformations ‘left’, ‘small’, ‘up’, and ‘big’.

709
710

711 **A comparison of visual object invariance in V4 and IT**

712
713 Information about object identity, across changes in identity-preserving transformations, is
714 reported to be more accessible to a linear read-out in IT as compared to V4 (Rust and DiCarlo
715 2010). To determine whether this difference between V4 and IT was reflected during the IDMS
716 task, we measured the ability of a 4-way linear object identity classifier, trained at each
717 transformation, to generalize to other transformations. Specifically, “reference performance” was
718 measured as cross-validated classifier performance when the training and testing trials came
719 from the same transformation. “Generalization performance” was measured as cross-validated
720 classifier performance when the testing trials came from the three transformations that were not
721 used for training. To avoid confounding visual and target match modulation, each type of
722 performance was computed separately for target matches and distractors (in all possible
723 combinations) and then averaged (see Methods). Finally, “generalization capacity” was
724 measured as the ratio of generalization over reference performance after subtracting the value
725 expected by chance (where chance = 25%).

726
727 Fig 5a depicts how reference and generalization performance grew as a function of population
728 size in each brain area. In V4, generalization performance remained modest across all
729 population sizes whereas V4 reference performance grew at a faster rate. In IT, both reference
730 and generalization performance grew at non-negligible rates. Fig 5b summarizes the results in
731 the two brain areas by plotting the endpoints of the plots in Fig 5a. Generalization capacity,
732 computed as the ratio of generalization over reference performance, was higher in IT as
733 compared to V4 (V4 = 0.16; IT = 0.47; $p < 0.001$), consistent with IT reflecting a more linearly-
734 separable object representation. This plot also reveals slightly lower reference performance in
735 V4 for matched numbers of units (Fig 5b) despite the two populations reflecting matched
736 average single-unit visual modulation (Fig 4f). We have determined that this small difference
737 can be attributed to the slightly higher variance-to-mean ratio in V4 as compared to IT (reported
738 above, mean Fano factor V4 = 1.41; mean Fano factor IT = 1.35), as opposed to other factors
739 such how the information is tiled across the stimulus space or differences in task-relevant
740 modulation (not shown). To confirm that IT generalization capacity remained higher even under
741 conditions in which more total visual information was available in V4, we also computed
742 generalization capacity for the full V4 population ($n = 598$ units). As shown in Figure 5c,
743 generalization capacity remained higher in IT even under these conditions (mean V4 = 0.20;
744 mean IT = 0.47; $p < 0.001$). Higher generalization capacity also held for each of the
745 transformations individually (Fig 5d; ‘Big’ $p < 0.001$; ‘Left’ $p < 0.001$; ‘Small’ $p = 0.046$; ‘Up’ $p =$
746 0.001).

747
748

749 **Figure 5. Comparison of visual object invariance across identity-preserving transformations in**
750 **V4 versus IT. a)** Performance of V4 and IT on a 4-way linear read-out of object identity,
751 assessed either with cross-validated trials of the same transformation (“Reference”) or when
752 asked to generalize to transformations not used for training (“Generalization”; see text). **b)**
753 Reference and generalization performance for matched numbers of units ($n = 193$ for both V4

754 and IT populations), replotted from the endpoints in panel a. Generalization capacity was
755 computed as the ratio of generalization over reference performance after subtracting the value
756 expected by chance (where chance = 25%). **c)** Reference and generalization performance for
757 the full recorded V4 population (n = 598 units) as compared to the full recorded IT population
758 replotted from panel b (n = 193 units). **d)** Generalization capacity computed for matched
759 numbers of units in V4 and IT (n = 193 units), applied to each transformation separately. Single
760 asterisks denote $p < 0.05$; double asterisks denote $p < 0.01$; triple asterisks denote $p < 0.001$. In
761 all panels, error bars (standard error) reflect the variability that can be attributed to the specific
762 subset of trials chosen for training and testing and, for subsets of units smaller than the full
763 population, the specific subset of units chosen.

764
765
766

767 In sum, the results presented thus far demonstrate that, consistent with earlier reports, V4 and
768 IT can be compared with approximately matched numbers of units, and that visual
769 representations of object identity are more accessible to a linear population read-out in IT during
770 IDMS.

771
772
773

774 **Conceptualizing IDMS target match computation**

775

776 To interpret the different types of signals that might be reflected in V4 and IT during IDMS, it is
777 useful to conceptualize how target match signals – which reflect the solution to IDMS – might be
778 computed. When considered in terms of a single 4x4 “looking at” vs. “looking for” slice of the
779 experimental design matrix (Fig 2b), target match signals are reflected as diagonal structure (Fig
780 6a, right, ‘Target match (four object)’). In the most straightforward description of target match
781 computation, congruent ‘visual’ information (vertical structure) and ‘target identity’ information
782 (horizontal structure) combine in a nonlinear fashion to compute target match detectors that are
783 selective for one object presented as a target match (‘Target match (one object)’). Finally, these
784 are pooled across the four different objects to create ‘Four object target match detectors’ that
785 respond whenever a target is in view (Fig 6a). Consequently, the class of proposals presented
786 in Fig 1a, where top-down modulation is integrated exclusively in V4, has at least two variants.
787 In the first, target match signals exist in V4 and arrive in IT via a feed-forward process (Fig 6b),
788 possibly with some linear pooling to produce target match invariance (across object identity). In
789 the second, target identity signals (as opposed to target match signals) are reflected in V4, and
790 IT target match signals are computed in IT via the nonlinear combination of these inputs (Fig
791 6c).

792
793

794 **Figure 6.** *Conceptualizing IDMS target match computation.* **a)** An idealized depiction of how
795 target match signals, which reflect the solution to the IDMS task, might be computed. For
796 simplicity, the computation is described for one 4x4 slice of the experimental design matrix,
797 which corresponds to viewing each of four objects (‘Looking AT’) in the context of each of four
798 objects as a target (‘Looking FOR’) at one transformation. In the first stage of this idealization of
799 target match computation, a unit reflecting visual information and a unit reflecting persistent
800 target identity information (i.e. working memory) are combined, and the result is passed through

801 a threshold. The resulting unit reflects target match information for one object. Next, four of
802 these units (each with a different object preference) are linearly combined to produce a unit that
803 signals whether a target match is present, regardless of the identity of the object. **b)** A variant of
804 the class of “IT: Inherited” proposals (Fig 1a) in which target match information is computed in
805 V4 and then fed forward to IT. **c)** A variant of the class of “IT: Inherited” proposals in which
806 visual and target identity information are both present in V4 and then fed forward to IT, where
807 they are combined to compute the target match signal. **d-e)** The response matrices
808 corresponding to 3 example units from V4 and IT. Response matrices were plotted as the
809 average firing rates across trials, and rescaled from the minimum (black) to maximum (white)
810 response across all experimental conditions.

811

812

813 We found examples of nearly all of these types of idealized units in V4 and/or IT (Figures 6d-e).
814 In both areas, we found ‘purely visual’ units that responded selectively to images but were not
815 modulated by other factors, such as target identity or whether an image was presented as a
816 target match (Fig 6d-e, ‘Purely visual’). In contrast, one notable difference between V4 and IT
817 was the existence of a handful of IT units (~10/193) that reflected the remarkable property of
818 responding to nearly every image presented as a target match (every object at every
819 transformation) but not when those same images were presented as distractors (Fig 6e, ‘Target
820 match (four object)’). We did not find any such units in V4. However, in both V4 and IT, we
821 found units that responded preferentially to individual objects presented as target matches as
822 compared to distractors (Fig 6d-e, ‘Target match (one object)’). We note that while these
823 illustrative examples were chosen because they reflect intuitive forms of pure selectivity, many
824 (if not most) units tended to reflect less intuitive mixtures of visual and task-relevant modulation.
825

826 To more quantitatively compare the types of signals reflected in V4 and IT, we extended the
827 procedure presented in Fig 4 to not only quantify ‘visual’ modulation (i.e. modulation that can be
828 attributed to changes in the identity of the visual image), but also other types of non-overlapping
829 modulations that could be attributed to: ‘target identity’ modulation - changes in the identity of a
830 sought target; ‘target match’ modulation - changes in whether an image was a target match or a
831 distractor; and ‘residual’ modulation - nonlinear interactions between visual and target identity
832 that are not target match modulation (e.g. an enhanced response to a particular distractor
833 condition). When considered in terms of a single 4x4 “looking at” vs. “looking for” slice of the
834 experimental design matrix (Fig 2c), these modulations produce vertical, horizontal, diagonal,
835 and off-diagonal structure, respectively (Fig 7a). Notably, this analysis defines target match
836 modulation as a differential response to the same images presented as target matches versus
837 distractors, or equivalently, diagonal structure in the transformation slices presented in Fig 7a.
838 Consequently, units similar to both the ‘target match (one object)’ unit as well as the ‘target
839 match (four object)’ unit (Fig 6d-e) reflect target match modulation, as both units have a
840 diagonal component to their responses. What differentiates these two types of units is that the
841 ‘Target match (one object)’ unit also reflects selectivity for image and target identity, which is
842 reflected in this analysis as a mixture of target match, visual, and target identity modulation.
843

844

845 **Figure 7. Evolution of different types of single unit modulations in V4 and IT. a)** To illustrate the
846 different types of task-relevant signals that could be present in V4 and IT, shown is a slice

847 through the IDMS experimental design (Figure 2c), corresponding to one transformation. Shown
848 are visual modulations, which differentiate between different objects in view (vertical structure);
849 target identity modulations, which differentiate between different target objects (horizontal
850 structure); target match modulations, which differentiate between whether objects appear as a
851 target match versus a distractor (diagonal structure); and residual modulations, which
852 differentiate between any other types of conditions (e.g. a response to a particular distractor
853 condition such as looking for object 4 when looking at object 2). **b-e)** Modulations were
854 computed for each type of experimental parameter in units of the standard deviations around
855 each unit's grand mean spike count (see Results). In each panel, average modulation
856 magnitudes across units in V4 ($n = 598$) and IT ($n = 193$) shown on the left as a function of time
857 (ms after stimulus onset). Modulation magnitudes, computed in spike count bins 50 ms wide and
858 shifted by 10 ms, are plotted corresponding to the midpoint of each bin. The bar plots show
859 average signal magnitudes quantified within broader spike counting windows indicated by the
860 rectangles on the left (V4: 40-210 ms, red rectangle; IT: 80-250 ms, gray rectangle). Triple
861 asterisks denote $p < 0.001$; 'ns' indicates $p > 0.05$. Error bars reflect the standard error of
862 modulation across units, computed via a bootstrap procedure.

863
864 To compare these different types of task-relevant signals between V4 and IT, we applied the
865 analysis to spike count windows positioned at sliding locations relative to stimulus onset, as well
866 as the same counting windows described for Fig 4 (170 ms; V4: 40-210 ms; IT: 80-250 ms; Fig
867 7b-e). As expected, visual modulation did not exist before stimulus onset, and visual signals
868 arrived in V4 ~ 40 ms earlier than in IT in both animals (Fig 7b). In contrast, modulations
869 reflecting information about whether an image was a target match or a distractor ('target match'
870 modulation) were considerably smaller in V4 as compared to IT in both animals (Fig 7c; monkey
871 1 $p < 0.001$; monkey 2 $p < 0.001$). In monkey 1, V4 target match modulations increased
872 throughout the viewing period, and reached levels that were similar to those found in IT, but this
873 rise occurred with a delay in V4 relative to IT. This was not replicated in monkey 2, where target
874 match modulations were small throughout the viewing period.

875
876 Modulations reflecting information about the identity of the target ('target identity' modulation)
877 were present in both V4 and IT before stimulus onset (Fig 7c), consistent with persistent
878 working memory signals in both brain areas. These persistent signals were stronger in IT as
879 compared to V4 in monkey 1 ($p < 0.001$) but comparable in size between V4 and IT in monkey 2
880 ($p = 0.23$). Lastly, we found that in both V4 and IT, residual modulation was small relative to the
881 other types of modulations (Fig 7e). Residual modulation was comparable in size between V4
882 and IT in monkey 1 ($p = 0.46$) and larger in IT than V4 in monkey 2 ($p < 0.001$). To summarize
883 these results, we found that in both monkeys, visual modulation was matched between V4 and
884 IT whereas target match signals were weaker in V4. We also found persistent target identity
885 signals that were reflected in both areas before and throughout the stimulus-evoked period.

886
887 As a complementary analysis, we also quantified the total amount of non-visual, 'cognitive'
888 modulation (combined target match, target identity, and residual modulation), and compared it to
889 the evolution of the visual modulation (Fig 8). In both brain areas, total cognitive modulation was
890 considerable throughout the analysis window. During the latency-corrected stimulus-evoked
891 period, cognitive modulations were 41% and 81% the size of the visual modulations in V4 and
892 IT, respectively. These results demonstrate that considerable non-visual, task-relevant
893 modulations exist in both brain areas, and they also suggest that these are smaller in V4 as
894 compared to IT.

895
896
897 **Figure 8. Single unit cognitive modulations in V4 and IT cortex. a-b)** Cognitive modulations (V4:
898 light red; IT: light gray) were computed as the sum of target identity, target match, and residual
899 modulations, and are shown alongside visual modulations (V4: dark red; IT: dark gray). Mean
900 modulation magnitudes are computed in the same manner and shown with the same
901 conventions as Fig 7. Labels in the bar plots above the cognitive modulation magnitudes
902 indicate the proportional size of cognitive relative to visual modulations in each brain area.

903
904 Below we focus on how these data constrain descriptions of how top-down task-relevant signals
905 combine with feed-forward visual information during IDMS (e.g. Fig 1; Fig 6b,c). As an overview,
906 we begin by evaluating the variant of the “IT: Inherited” class in which IT target match signals
907 are inherited directly from V4 (Fig 6b), both under the assumption that IT uniformly samples V4
908 units, as well as when IT is allowed to preferentially sample the “best” V4 units. Next, we
909 evaluate the variant of the “IT: Inherited” class that allows for IT nonlinear computation applied
910 to input arriving from V4 (Fig 6c). After ruling out both of these proposals, we conclude that
911 during the IDMS task, top-down signals must be integrated directly within IT (Fig 1b).

912
913
914 **Could target match signals arrive in IT via input from the “best” V4 neurons?**

915
916 The results presented in Fig 7c demonstrate that target match signals are, on average, larger in
917 IT than V4. This suggests that target match signals are unlikely to arrive in IT from V4 via a
918 simple feed-forward process, under the assumption that IT uniformly samples V4 neurons.
919 However, evidence from other studies suggests that the brain can learn to preferentially read-
920 out the subset of neurons that carry the most task-relevant information with extensive training
921 (Law and Gold 2009) and the monkeys involved in these experiments were trained extensively.
922 Could a version of the feed-forward proposal in which IT preferentially samples the “best” V4
923 neurons account for our data? To allow us to address this question, we sampled 3-fold more
924 units in V4 as compared to IT, consistent with anatomical estimates of the ratios of neurons
925 between the two brain areas (DiCarlo et al. 2012). This allowed us to compare V4 and IT under
926 different assumptions, including that IT sampled V4 units “uniformly” versus the “best” subset
927 with regard to the amount of IDMS information reflected in their responses.

928
929 Target match signals, reflected as diagonal matrix structure (e.g. target match units for one
930 object or across multiple objects; Fig 6a) translate into a linearly separable representation of the
931 same images presented as target matches as compared to distractors (Fig 9a). To quantify the
932 amount of linearly separable target match information in V4 and IT, we computed the cross-
933 validated performance of a linear classifier to perform this 2-way classification at each
934 transformation separately and then averaged over transformations (Fig 9b, see Methods). To
935 verify that uniform sampling of V4 could not account for target match information in IT, we
936 randomly selected IT units up to the total numbers of units that we recorded (Fig 9c, gray), and
937 compared this to a random selection of V4 units for matched sized populations (and thus always
938 a subset of the V4 data Fig 9c, red). As expected based on the results presented in Fig 7c,
939 cross-validated population performance was higher than chance in V4, but was significantly
940 higher in IT as compared to V4 (Fig 9c, gray versus red; in both monkeys, compared at $n = 98$ in
941 monkey 1 and $n = 95$ in monkey 2, $p < 0.001$). These results verify that IT target match

942 information is not directly inherited from V4 under the assumption of a uniform sampling of V4
943 by IT.

944
945 To assess whether a “best” sampling description of V4 by IT could account for our data, we
946 recomputed performance for V4 and IT populations that were matched in size, but when only
947 the top-ranked V4 units were included. In this analysis, units were ranked based on the training
948 data before computing cross-validated performance. We found that V4 performance was slightly
949 higher for the best units as compared to randomly selected units (Fig 9c, cyan vs. red),
950 however, performance for the best V4 units remained lower than IT performance in both
951 monkeys (Fig 9c, cyan vs. gray, $p < 0.001$). These results suggest that during IDMS, IT target
952 match modulation cannot be accounted for via feed-forward propagation of this modulation from
953 V4, even if IT were to sample from the “best” V4 subset (Fig 6b).

954
955

956 **Figure 9.** *A comparison of linearly separable target match information in V4 and IT. a-b)* The
957 IDMS task can be envisioned as a two-way classification of the same images presented as
958 target matches versus as distractors. Shown are cartoon depictions where each point depicts a
959 hypothetical population response for a population of two neurons on a single trial, and clusters
960 of points depict the dispersion of responses across repeated trials for the same condition.
961 Included are the hypothetical responses to the same images presented as target matches
962 (black) and as distractors (gray). The dotted line depicts a hypothetical linear decision boundary.
963 **a)** A schematic of two neurons that each respond to one object as a target match. In this
964 scenario, target matches and distractors are linearly separable. **b)** A schematic of the IDMS
965 task, where four images must be classified as target matches as compared to distractors,
966 applied to a linearly separable representation. **c)** Performance of a linear classifier trained to
967 classify whether an object was a target match or a distractor, invariant of the object’s identity (at
968 one transformation). Performance was assessed at each identity-preserving transformation
969 (‘Big’, ‘Left’, ‘Small’, ‘Up’), and then averaged. Performance was higher in IT (gray) than in V4,
970 both when V4 units were sampled uniformly from the full population (red) and when V4 units
971 were sampled by choosing the best possible V4 units based on the training data (cyan). Monkey
972 1: $n = 98$ units, Monkey 2: $n = 95$ units. Error bars (standard error) reflect the variability that can
973 be attributed to the specific subset of trials chosen for training and testing, and, for subsets of
974 units smaller than the full population, the specific subset of units chosen. Dashed line indicates
975 chance performance.

976
977

978 **Could target match signals arrive in IT via nonlinear combinations of input from V4?**

979

980 To evaluate the variant of the matched proposal in which IT target match signals are computed
981 via nonlinear combinations of inputs arriving from V4 (Fig 6c), we quantified the “total” target
982 match information in each brain area, regardless of its format. Specifically, combinations of
983 visual and target identity signals (reflected in different units) map to target match information
984 present in a nonlinearly separable format (Fig 10a) whereas target match signals map to target
985 match information that is linear (Fig 9a) and a measure of total target match information
986 quantifies information regardless of its format.

987

988 Total target match information was measured as cross-validated performance on the same 2-
989 way classification as presented in Fig 9, but for a maximum likelihood (as opposed to linear)
990 classifier (see Methods). This nonlinear classifier assesses the total amount of target match
991 information regardless of its format (combined linear and nonlinear). Cross-validated population
992 performance was higher than chance in V4 (Fig 10b, filled red points; in both monkeys,
993 compared at $n = 98$ in monkey 1 and $n = 95$ in monkey 2, $p < 0.001$), but was also higher in IT as
994 compared to V4 (Fig 10b, gray; in both monkeys, compared at $n = 98$ in monkey 1 and $n = 95$ in
995 monkey 2, $p < 0.001$). These results suggest that IT target match information is not exclusively
996 inherited via feed-forward projections arriving from V4 (Fig 1a; Fig 6b-c), but rather, integrated
997 directly in IT itself (Fig 1b).

998
999

1000

1001 **Figure 10.** A comparison of total (linear and nonlinear) target match information in V4 and IT. **a)**

1002 A schematic of two neurons, one 'visual' neuron and one 'target identity' neuron. In this
1003 scenario, target match information exists, but is present in a non-linearly separable format. **b)** A
1004 schematic of the IDMS task where four images must be classified as target matches versus
1005 distractors, applied to a non-linearly separable representation. **c)** Performance of a nonlinear,
1006 maximum likelihood classifier trained to classify between whether an object was a target match
1007 or a distractor, invariant of object identity. Performance was assessed at each identity-
1008 preserving transformation, and then averaged. Error bars (standard error) reflect the variability
1009 that can be attributed to the specific subset of trials chosen for training and testing, and, for
1010 subsets of units smaller than the full population, the specific subset of units chosen. Dashed line
1011 indicates chance performance.

1012

1013

1014

1015

1016 DISCUSSION

1017
1018 Finding sought objects requires the brain to compare visual information about the objects in
1019 view with information about the currently sought target to compute a signal that reports when a
1020 target match has been found. During object search, information about the identity of a sought
1021 target and/or whether it is a target match is thought to be fed-back to mid to higher stages of the
1022 ventral visual pathway, including V4 and IT, but the specific path this information takes is
1023 unclear. In this study, we sought to differentiate between scenarios in which top-down
1024 information is integrated directly in IT (Fig 1b) versus those in which it is integrated in V4 and
1025 arrives in IT via feed-forward propagation (Fig 1a). We evaluated a number of feed-forward
1026 descriptions between V4 and IT, and found none of them could account for the amount of non-
1027 visual, task-relevant information present in IT. These included a model in which IT uniformly
1028 samples target match signals from V4 (Fig 9, red), a model in which IT preferentially samples
1029 target match signals from the best V4 units (Fig 9, cyan), and a model that allowed for IT
1030 nonlinear processing of inputs arriving from V4 (Fig 10). Together, these results suggest that
1031 during IDMS, top-down, task-specific signals in IT are not exclusively inherited from V4 but
1032 rather are integrated within IT, at least in part.

1033
1034 We found non-visual, task-specific signals to be sizeable in V4 (~40% of the size of visual
1035 modulation), consistent with many other reports (Moran and Desimone 1985; Haenny et al.
1036 1988; Motter 1994; Motter 1994; Luck et al. 1997; McAdams and Maunsell 1999; McAdams and
1037 Maunsell 2000; Chelazzi et al. 2001; Ogawa and Komatsu 2004; Bichot et al. 2005; Hayden and
1038 Gallant 2005; Mirabella et al. 2007; Cohen and Maunsell 2009; Kosai et al. 2014). At the same
1039 time, we also found that non-visual, task-specific modulations to be even larger in IT (~80% the
1040 size of visual modulation). In a previous study, during a visual target search task in which
1041 monkeys made a saccade to a target match following the presentation of a sample image, non-
1042 visual, task-specific signals were reported to be more similar in V4 and IT (63% and 70% of the
1043 visually-evoked response in V4 and IT, respectively; Chelazzi et al. 1998; Chelazzi et al. 2001).
1044 One notable difference between our study and this earlier work is that our study compared V4
1045 and IT during a version of the delayed-match-to-sample task in which sought target objects
1046 could appear at different positions, sizes and background contexts. The fact that top-down, task-
1047 specific signals were considerably larger in IT versus V4 in our task may follow from the fact that
1048 IT contains a more explicit, linear representation of object identity across these transformations
1049 than V4 (reviewed by DiCarlo et al. 2012). Consequently, top-down modulation may be targeted
1050 directly to IT in situations that require an invariant object representation whereas the brain might
1051 target the pathway differently when tasks have different computational requirements. For
1052 example, because V4 receptive fields are smaller and retinotopically organized, V4 might serve
1053 as the primary locus for the integration of top-down signals for tasks that require spatial
1054 specificity, such as covert spatial attention tasks, and in these tasks little top-down integration
1055 might occur in IT (Moran and Desimone 1985). Only one earlier study has reported on the
1056 responses of IT neurons in the context of a DMS task in which, objects could appear at different
1057 identity-preserving transformations (Leuschow et al. 1994), but this study did not measure
1058 signals in V4.

1059
1060 Our results, which demonstrate larger non-visual, task-relevant modulations in IT as compared
1061 to V4, are consistent with more general interpretations that the magnitudes of top-down
1062 modulation exist in a gradient-like fashion hierarchically along the ventral visual pathway
1063 (reviewed by Noudoost et al. 2010). As described above, such gradients are consistent both
1064 with the integration of top-down modulation at multiple stages of the pathway (Fig 1a) as well as

1065 integration at a single locus, followed by feedback within the pathway itself (Fig 1b, right). One
1066 study (Buffalo et al. 2010) provided evidence supporting the latter description in V1, V2 and V4
1067 in the form of noting that not only the magnitude of modulation was greater in higher visual
1068 areas, but it also arrived earlier, consistent with a feed-back description. In our data, this issue
1069 was ambiguous: we found that in one monkey, the arrival of the target match signal appeared to
1070 be delayed in V4 as compared to IT (Fig 7c, Monkey 1) whereas in the other animal, it appeared
1071 to arrive earlier (Figure 7c, Monkey 2).

1072
1073 In an earlier series of reports, we compared the responses of IT and its projection area,
1074 perirhinal cortex, during a more classic version of the delayed-match-to-sample task (that did
1075 not incorporate variation in the objects' transformations; Pagan et al. 2013; Pagan and Rust
1076 2014; Pagan et al. 2016). We found that the responses of perirhinal cortex were well-described
1077 by a model in which top-down, task-relevant signals were integrated within or before IT
1078 consistent with a feed-forward process between IT and perirhinal cortex. The results presented
1079 here extend this understanding to suggest that the locus of top-down integration during DMS
1080 search tasks is unlikely to exclusively be V4, and that some amount of top-down integration is
1081 likely to happen directly within IT itself.

1082
1083 Computing a target match signal requires the combination of the visual representation of the
1084 currently viewed scene with a remembered representation of the sought target (e.g. Fig 6a). In
1085 an analysis of the same IT data presented here, we found that the IT population misclassified
1086 trials on which the monkeys made errors, supporting notions that the IT target match signal is in
1087 fact related to the neural signals used to make target match behavioral judgments (Roth and
1088 Rust 2018). The additional target match information present in IT that is not also present in V4
1089 could reflect the implementation of this comparison in IT itself, or alternatively, the comparison
1090 might be implemented in a higher order brain area and fed-back to IT cortex. The timing of the
1091 arrival of this signal in IT (which peaks at ~150 ms; Fig 7c) relative to the monkeys' median
1092 reaction times (~335 ms; Fig 2e), does not rule out the former scenario, but with our data we
1093 cannot definitively distinguish between these alternatives. Additionally, in this study monkeys
1094 were trained extensively on the images used in these experiments and future experiments will
1095 be required to address the degree to which these results hold under more everyday conditions
1096 in which monkeys are viewing images and objects for the first time.

1097
1098
1099
1100

1101

1102

1103

1104

1105

1106

1107

References:

- 1108
1109
- 1110 Bichot, N. P., A. F. Rossi and R. Desimone (2005) Parallel and serial neural mechanisms for
1111 visual search in macaque area V4. *Science* 308(5721): 529-534.
- 1112 Buffalo, E. A., P. Fries, R. Landman, H. Liang and R. Desimone (2010) A backward progression
1113 of attentional effects in the ventral stream. *Proc Natl Acad Sci U S A* 107(1): 361-365.
- 1114 Chelazzi, L., J. Duncan, E. K. Miller and R. Desimone (1998) Responses of neurons in inferior
1115 temporal cortex during memory-guided visual search. *J. Neurophysiology* 80: 2918-2940.
- 1116 Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (1993) A neural basis for visual search in
1117 inferior temporal cortex. *Nature* 363: 345-347.
- 1118 Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (2001) Responses of neurons in macaque
1119 area V4 during memory-guided visual search. *Cereb Cortex* 11(8): 761-772.
- 1120 Cohen, M. R. and J. H. Maunsell (2009) Attention improves performance primarily by reducing
1121 interneuronal correlations. *Nat Neurosci* 12(12): 1594-1600.
- 1122 Desimone, R. and S. J. Schein (1987) Visual properties of neurons in area V4 of the macaque:
1123 sensitivity to stimulus form. *J Neurophysiol* 57(3): 835-868.
- 1124 DiCarlo, J. J., D. Zoccolan and N. C. Rust (2012) How does the brain solve visual object
1125 recognition? *Neuron* 73(3): 415-434.
- 1126 Eskandar, E. N., B. J. Richmond and L. M. Optican (1992) Role of inferior temporal neurons in
1127 visual memory I. Temporal encoding of information about visual images, recalled images, and
1128 behavioral context. *Journal of Neurophysiology* 68: 1277-1295.
- 1129 Felleman, D. J. and D. C. Van Essen (1991) Distributed hierarchical processing in the primate
1130 cerebral cortex. *Cereb Cortex* 1(1): 1-47.
- 1131 Gattass, R., A. P. Sousa and C. G. Gross (1988) Visuotopic organization and extent of V3 and
1132 V4 of the macaque. *J Neurosci* 8(6): 1831-1845.
- 1133 Gibson, J. R. and J. H. R. Maunsell (1997) The sensory modality specificity of neural activity
1134 related to memory in visual cortex. *Journal of Neurophysiology* 78: 1263-1275.

- 1135 Haenny, P. E., J. H. Maunsell and P. H. Schiller (1988) State dependent activity in monkey
1136 visual cortex. II. Retinal and extraretinal factors in V4. *Exp Brain Res* 69(2): 245-259.
- 1137 Hayden, B. Y. and J. L. Gallant (2005) Time course of attention reveals different mechanisms
1138 for spatial and feature-based attention in area V4. *Neuron* 47(5): 637-643.
- 1139 Hung, C. P., G. Kreiman, T. Poggio and J. J. DiCarlo (2005) Fast readout of object identity from
1140 macaque inferior temporal cortex. *Science* 310(5749): 863-866.
- 1141 Kosai, Y., Y. El-Shamayleh, A. M. Fyall and A. Pasupathy (2014) The role of visual area V4 in
1142 the discrimination of partially occluded shapes. *J Neurosci* 34(25): 8570-8584.
- 1143 Law, C. T. and J. I. Gold (2009) Reinforcement learning can account for associative and
1144 perceptual learning on a visual-decision task. *Nat Neurosci* 12(5): 655-663.
- 1145 Leuschow, A., E. K. Miller and R. Desimone (1994) Inferior temporal mechanisms for invariant
1146 object recognition. *Cerebral Cortex* 5: 523-531.
- 1147 Luck, S. J., L. Chelazzi, S. A. Hillyard and R. Desimone (1997) Neural mechanisms of spatial
1148 selective attention in areas V1, V2, and V4 of macaque visual cortex. *J Neurophysiol* 77(1): 24-
1149 42.
- 1150 Maunsell, J. H., G. Sclar, T. A. Nealey and D. D. DePriest (1991) Extraretinal representations in
1151 area V4 in the macaque monkey. *Vis Neurosci* 7(6): 561-573.
- 1152 McAdams, C. J. and J. H. Maunsell (1999) Effects of attention on orientation-tuning functions of
1153 single neurons in macaque cortical area V4. *J Neurosci* 19(1): 431-441.
- 1154 McAdams, C. J. and J. H. Maunsell (2000) Attention to both space and feature modulates
1155 neuronal responses in macaque area V4. *J Neurophysiol* 83(3): 1751-1755.
- 1156 Miller, E. K. and R. Desimone (1994) Parallel neuronal mechanisms for short-term memory.
1157 *Science* 263(5146): 520-522.
- 1158 Mirabella, G., G. Bertini, I. Samengo, B. E. Kilavik, D. Frilli, C. Della Libera and L. Chelazzi
1159 (2007) Neurons in area V4 of the macaque translate attended visual features into behaviorally
1160 relevant categories. *Neuron* 54(2): 303-318.

- 1161 Moran, J. and R. Desimone (1985) Selective attention gates visual processing in the extrastriate
1162 cortex. *Science* 229(4715): 782-784.
- 1163 Motter, B. C. (1994) Neural correlates of attentive selection for color or luminance in extrastriate
1164 area V4. *J Neurosci* 14(4): 2178-2189.
- 1165 Motter, B. C. (1994) Neural correlates of feature selective memory and pop-out in extrastriate
1166 area V4. *J Neurosci* 14(4): 2190-2199.
- 1167 Noudoost, B., M. H. Chang, N. A. Steinmetz and T. Moore (2010) Top-down control of visual
1168 attention. *Curr Opin Neurobiol* 20(2): 183-190.
- 1169 Ogawa, T. and H. Komatsu (2004) Target selection in area V4 during a multidimensional visual
1170 search task. *J Neurosci* 24(28): 6371-6382.
- 1171 Op De Beeck, H. and R. Vogels (2000) Spatial sensitivity of macaque inferior temporal neurons.
1172 *J Comp Neurol* 426(4): 505-518.
- 1173 Pagan, M. and N. C. Rust (2014) Dynamic target match signals in perirhinal cortex can be
1174 explained by instantaneous computations that act on dynamic input from inferotemporal cortex.
1175 *J Neurosci* 34(33): 11067-11084.
- 1176 Pagan, M. and N. C. Rust (2014) Quantifying the signals contained in heterogeneous neural
1177 responses and determining their relationships with task performance. *J Neurophysiol* 112(6):
1178 1584-1598.
- 1179 Pagan, M., E. P. Simoncelli and N. C. Rust (2016) Neural Quadratic Discriminant Analysis:
1180 Nonlinear Decoding with V1-Like Computation. *Neural Comput*: 1-29.
- 1181 Pagan, M., L. S. Urban, M. P. Wohl and N. C. Rust (2013) Signals in inferotemporal and
1182 perirhinal cortex suggest an untangling of visual target information. *Nat Neurosci* 16(8): 1132-
1183 1139.
- 1184 Roth, N. and N. C. Rust (2018) Inferotemporal cortex multiplexes behaviorally-relevant target
1185 match signals and visual representations in a manner that minimizes their interference. *PLoS*
1186 *ONE* In press.
- 1187 Rust, N. C. and J. J. DiCarlo (2010) Selectivity and tolerance ("invariance") both increase as
1188 visual information propagates from cortical area V4 to IT. *J Neurosci* 30(39): 12978-12995.

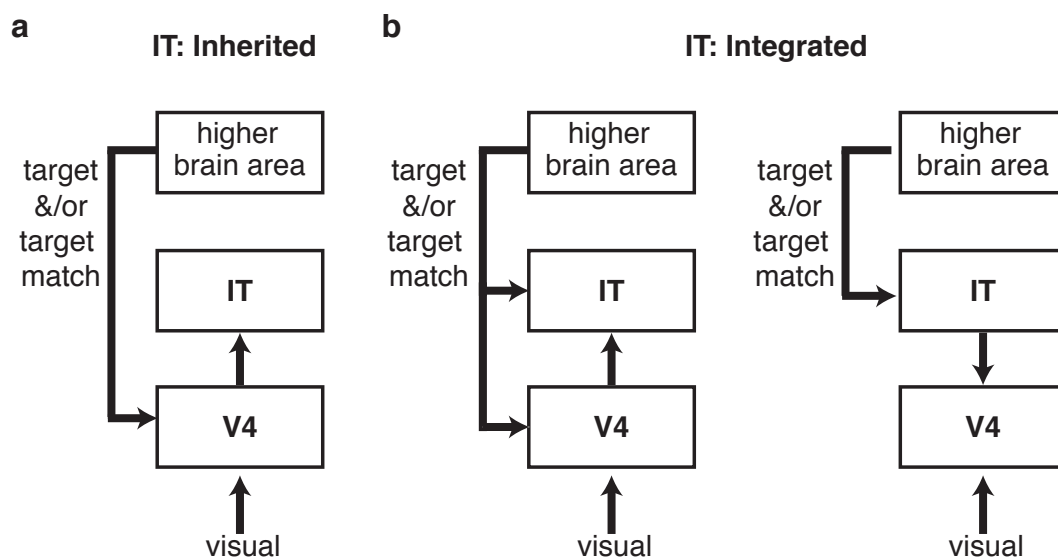


Figure 1

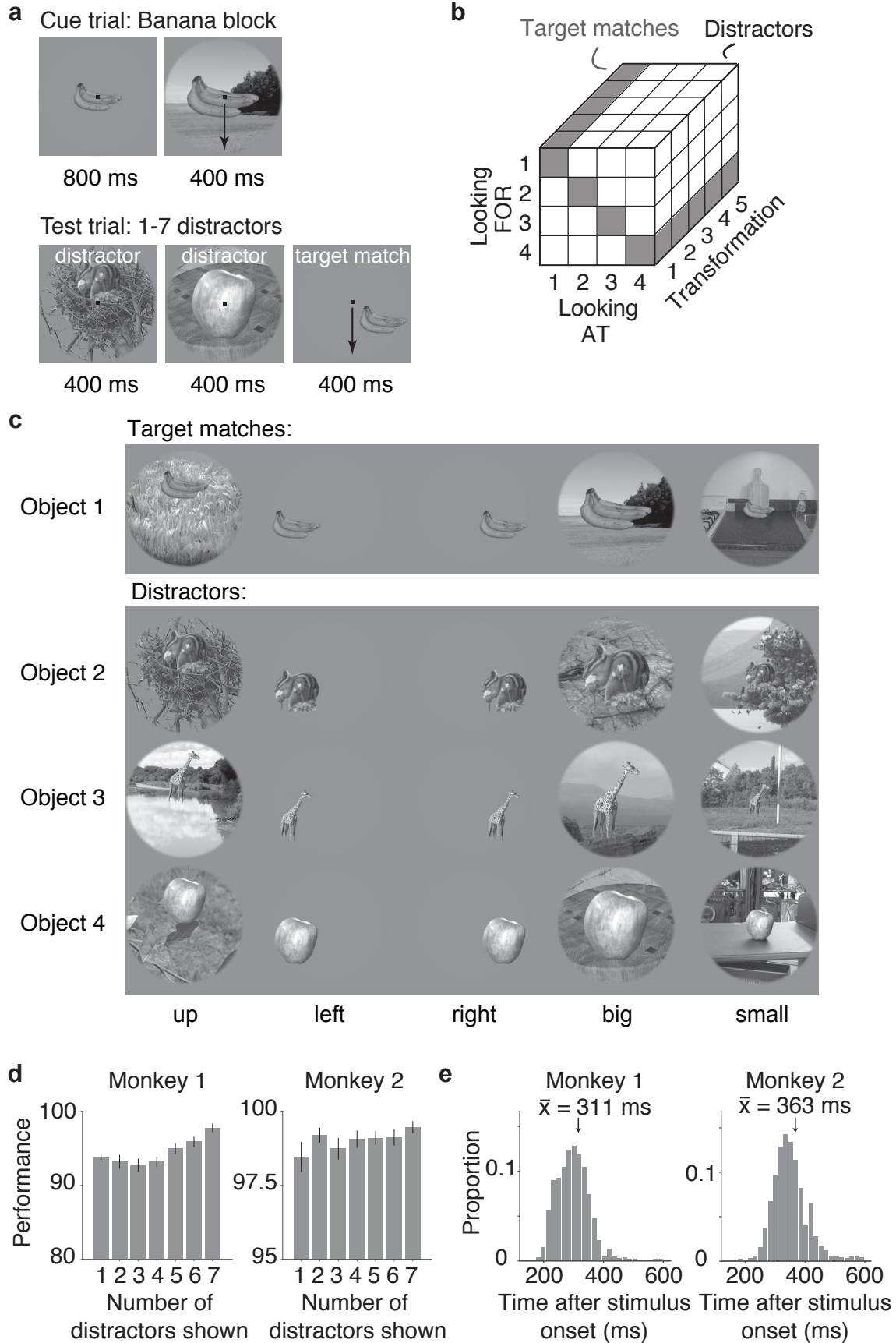


Figure 2

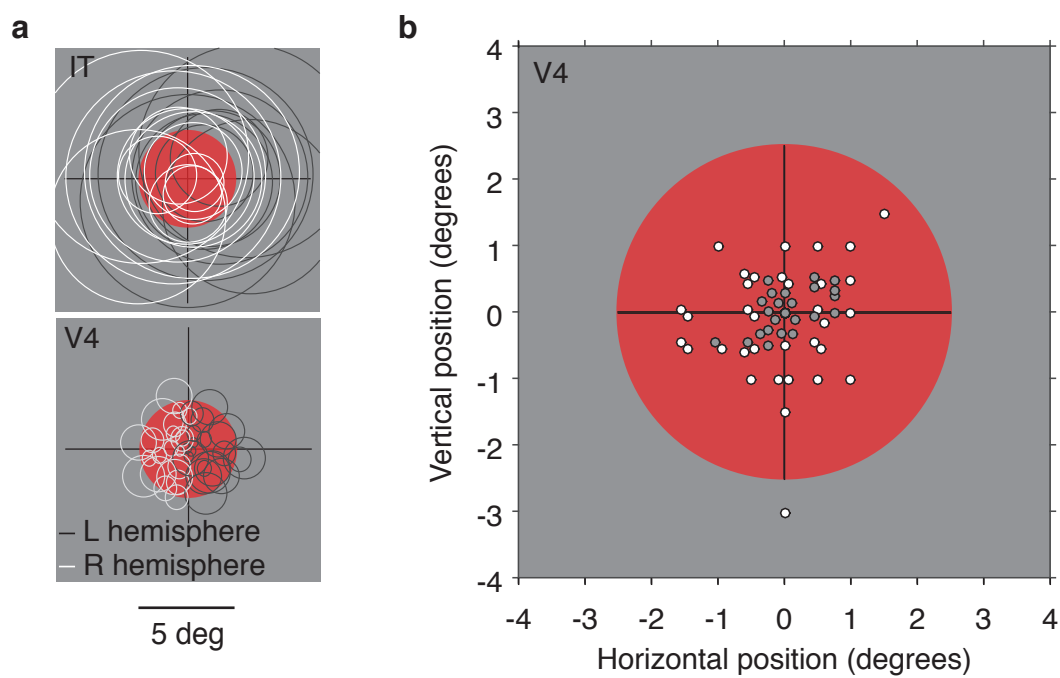


Figure 3

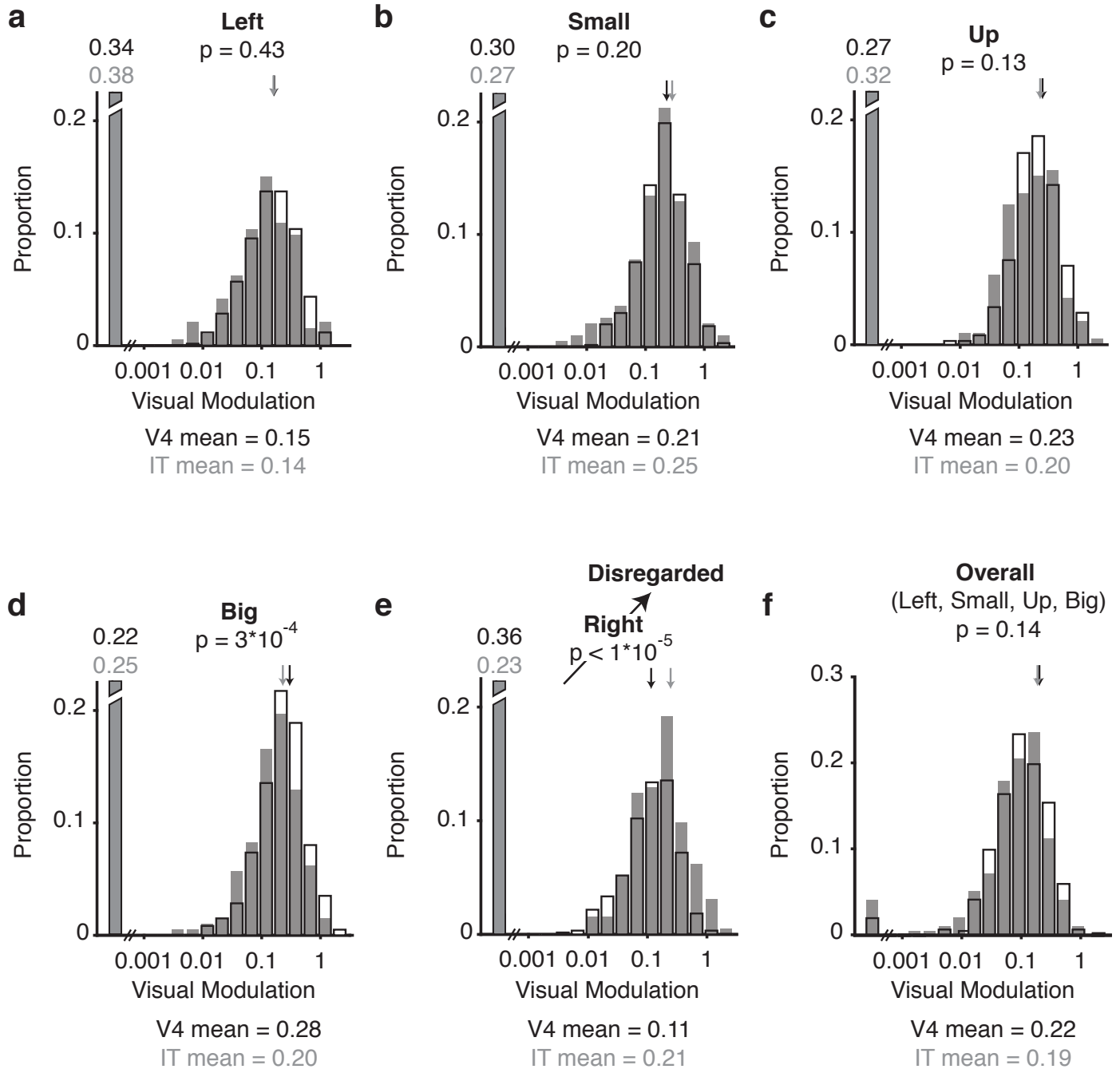


Figure 4

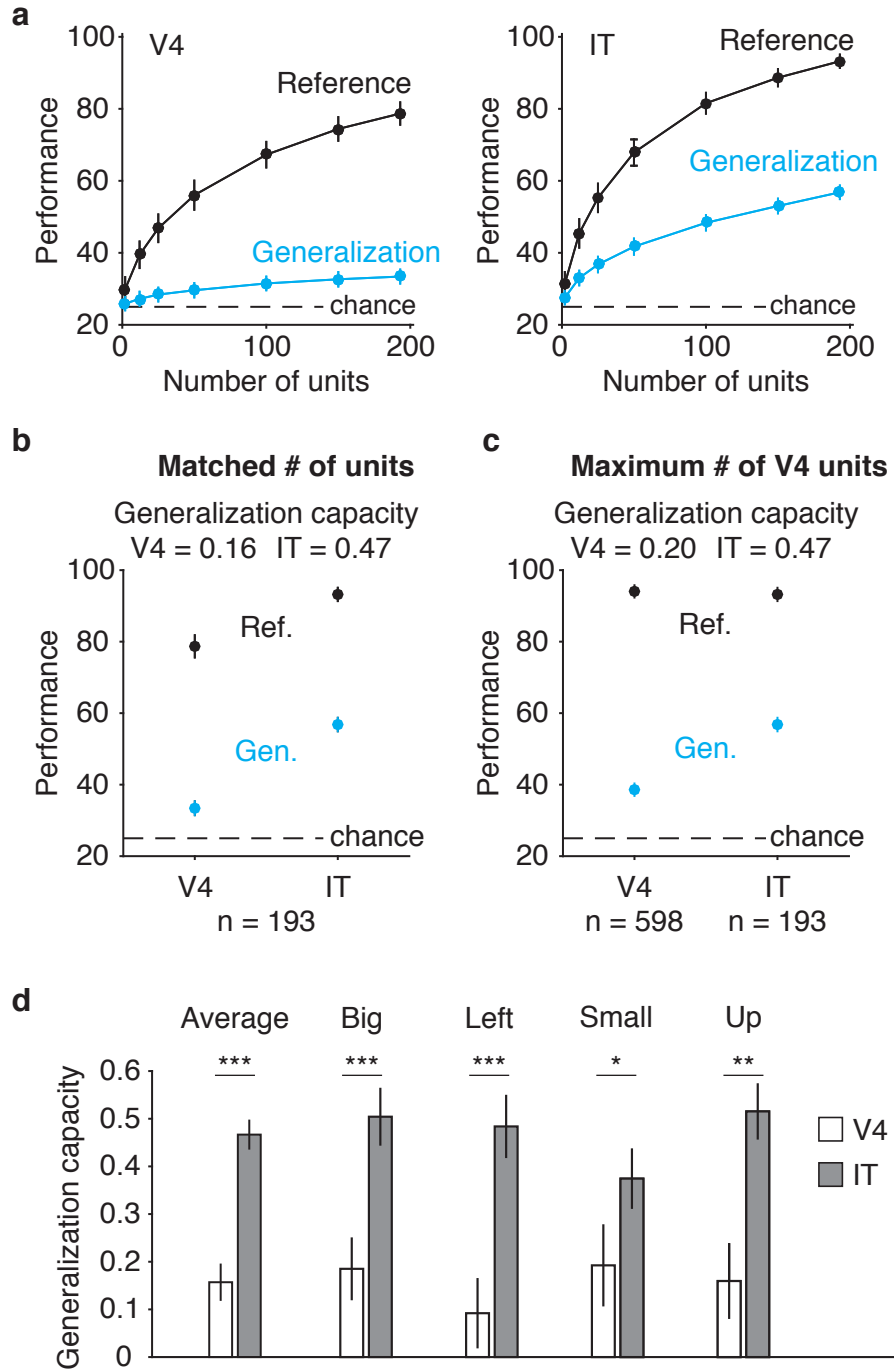


Figure 5

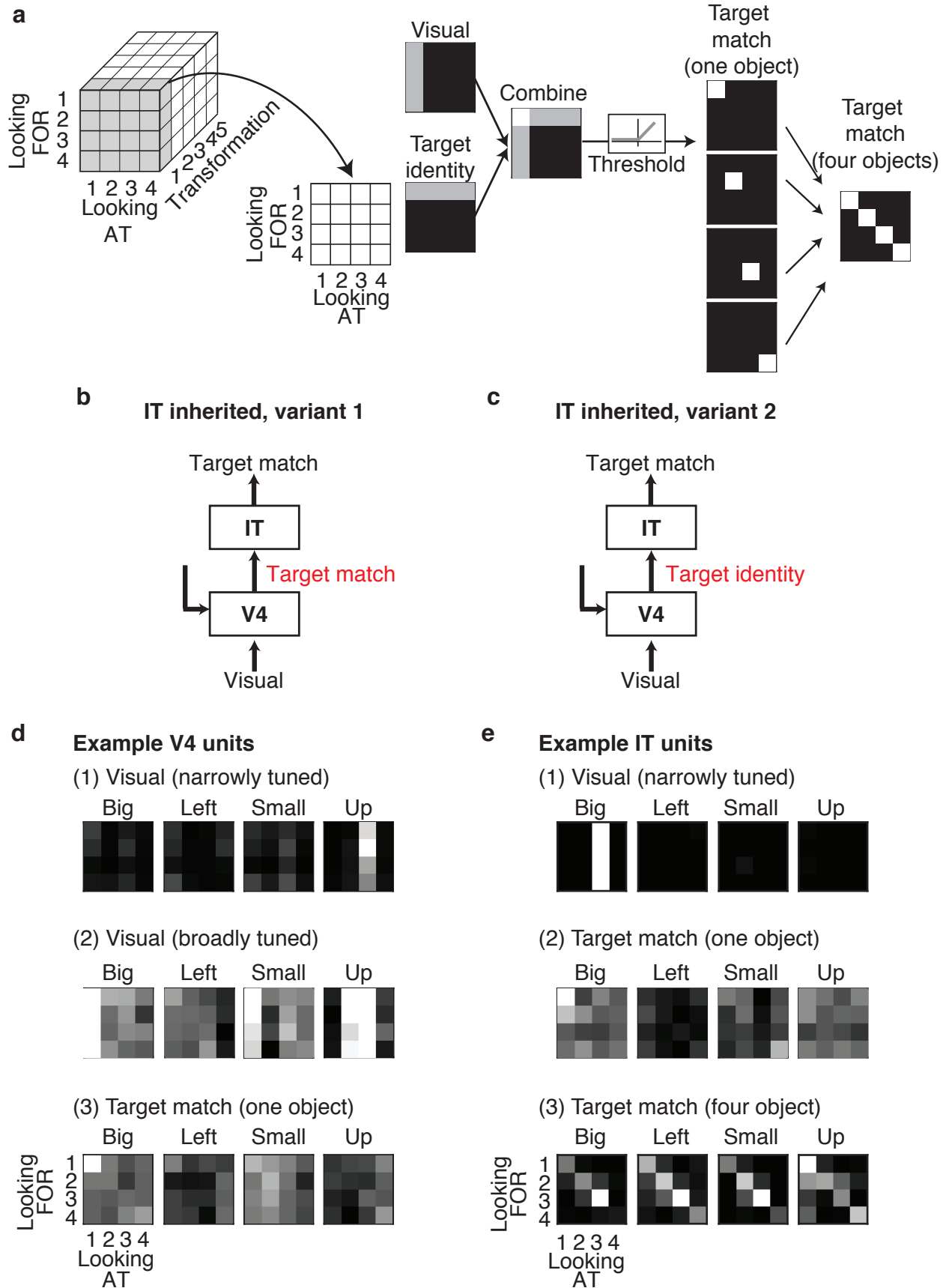


Figure 6

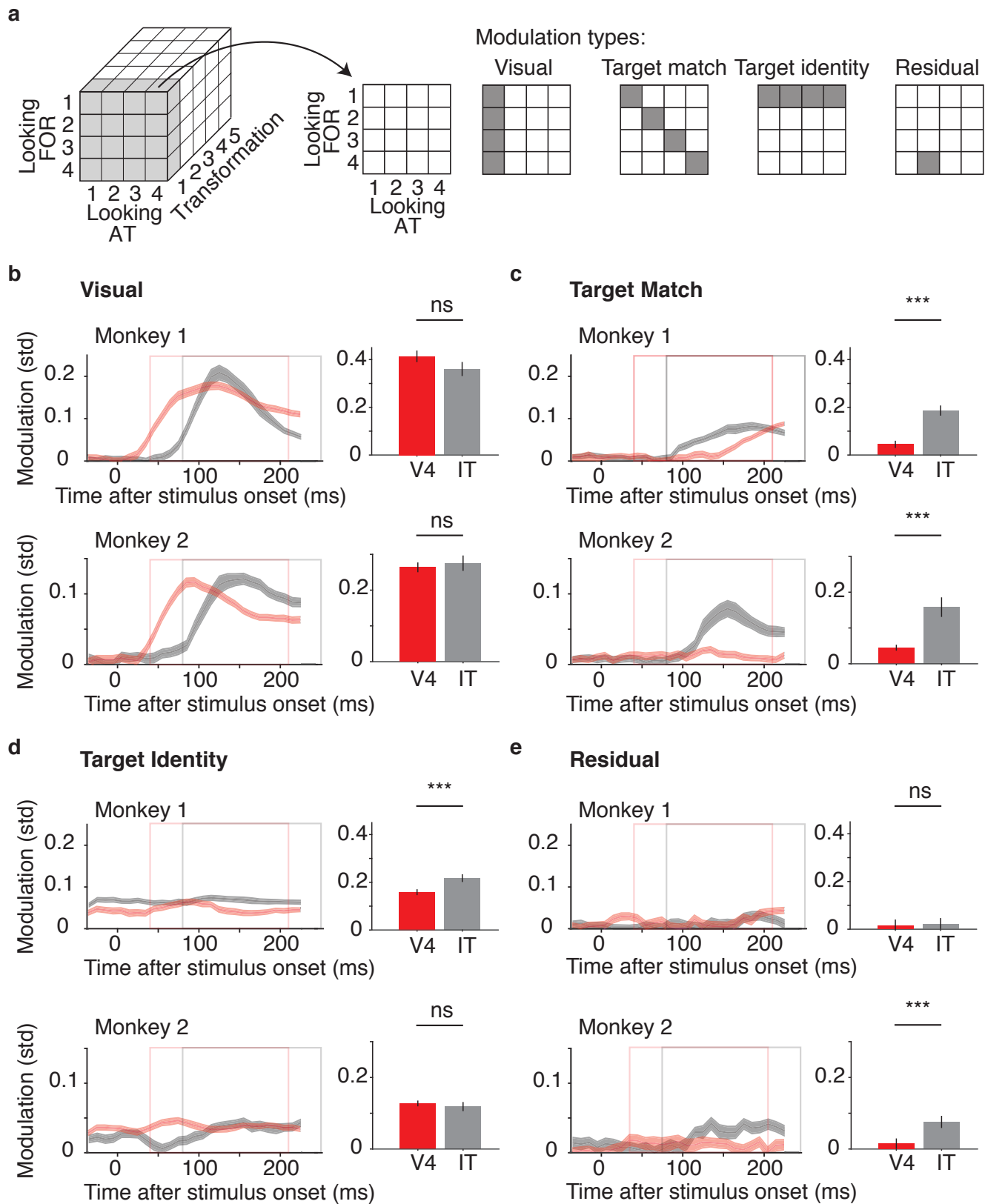


Figure 7

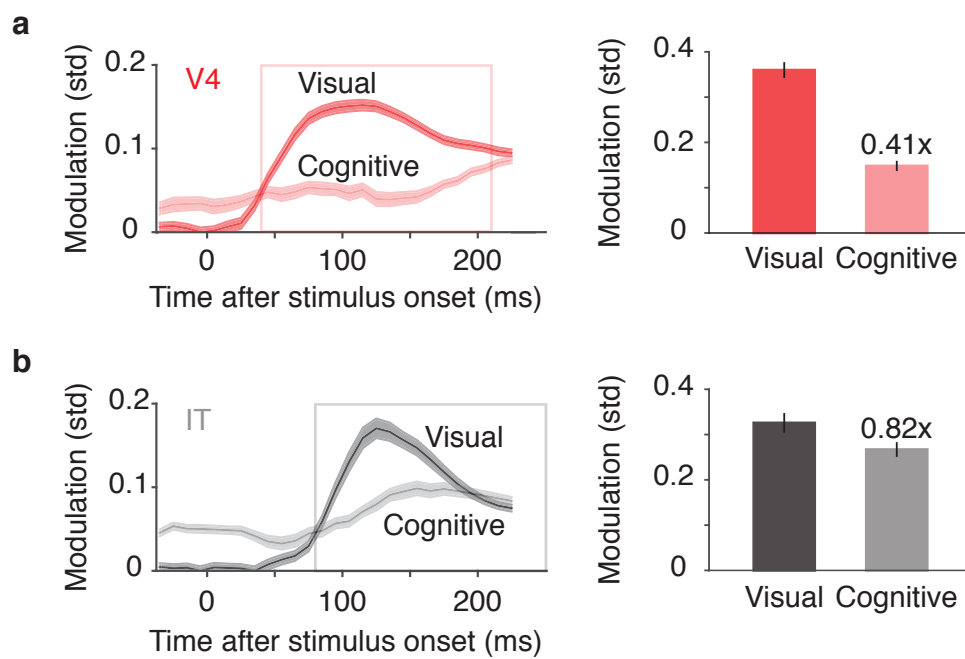


Figure 8

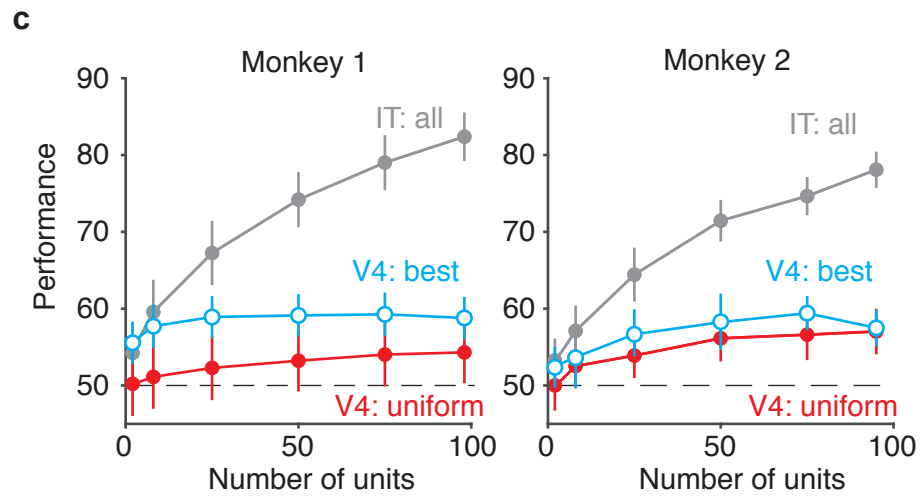
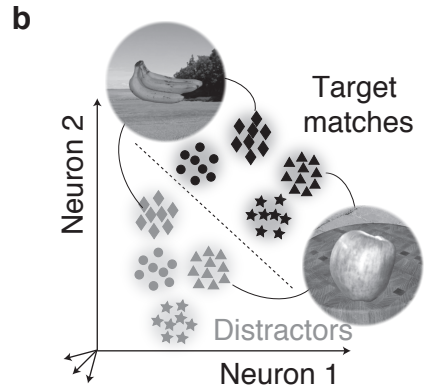
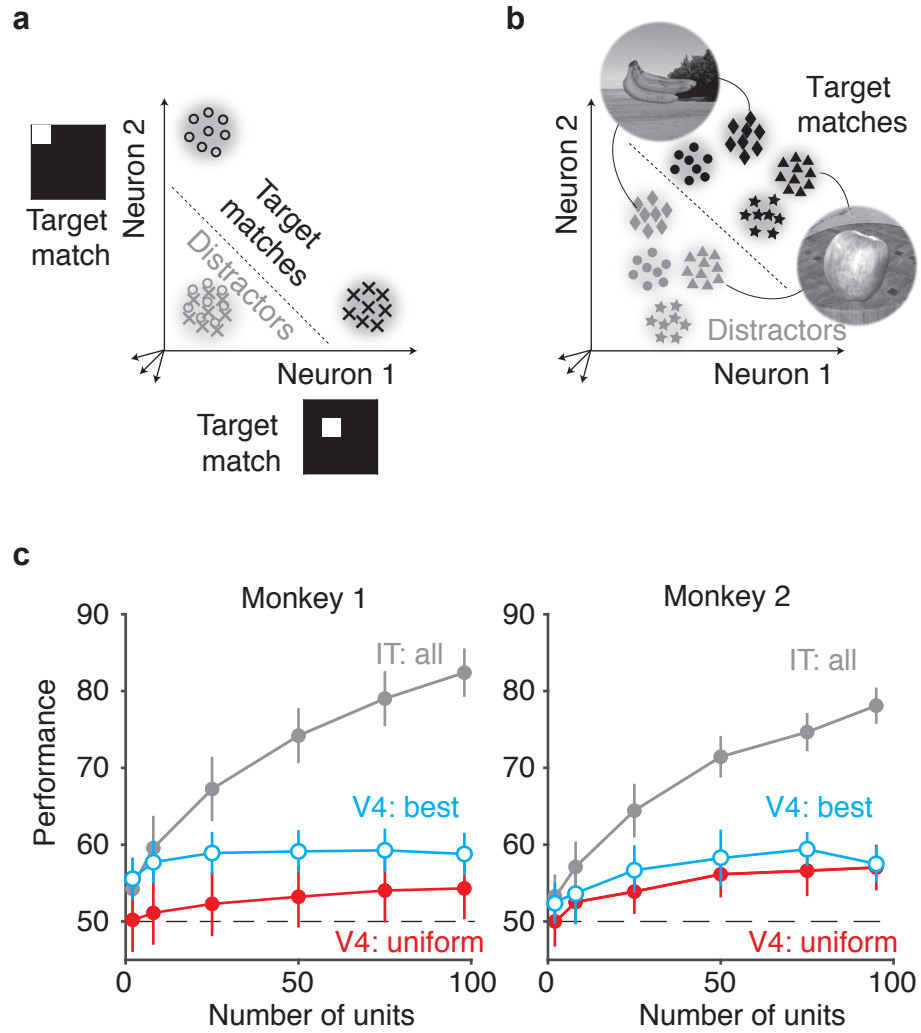


Figure 9

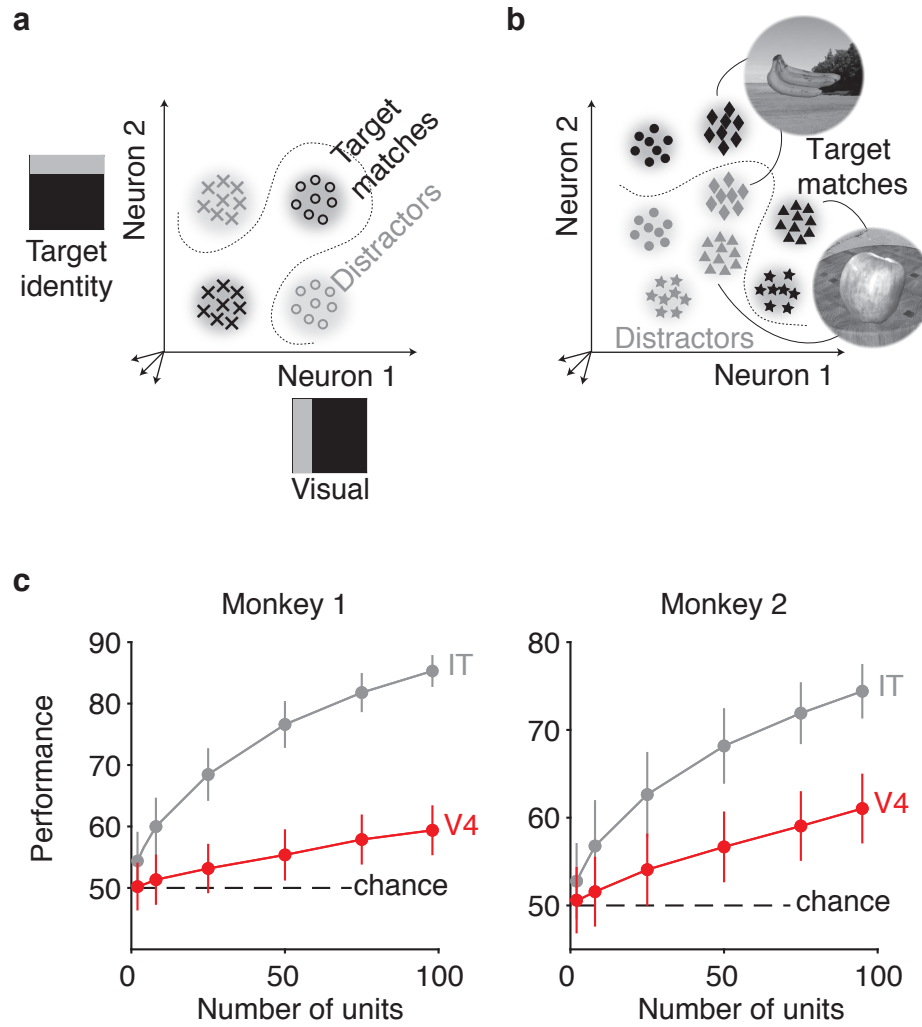


Figure 10