Current Biology

Article

Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images

Highlights

- An analysis of generative networks yields distinct codes for seen and mental images
- An encoding model for mental images predicts human brain activity during imagery
- Low-level visual areas encode mental images with less precision than seen images
- The overlap of imagery and vision is limited by high-level representations

Authors

Jesse L. Breedlove, Ghislain St-Yves, Cheryl A. Olman, Thomas Naselaris

Correspondence

tnaselar@musc.edu

In Brief

Breedlove, St-Yves, et al. analyze brain activity of human subjects generating hundreds of mental images. Low-level visual cortical areas encode mental images much like high-level areas encode seen images. This coding strategy emerges in a network model of visual cortex that generates images by feeding representations from higher to lower levels.

Breedlove et al., 2020, Current Biology *30*, 2211–2224
June 22, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cub.2020.04.014

Current Biology

CellPress

Article

Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images

Jesse L. Breedlove,[1,4] Ghislain St-Yves,[1,4] Cheryl A. Olman,[2,3] and Thomas Naselaris[1,5,*]
[1]Medical University of South Carolina, Charleston, SC 29403, USA
[2]Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN 55455, USA
[3]Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: tnaselar@musc.edu
https://doi.org/10.1016/j.cub.2020.04.014

SUMMARY

The relationship between mental imagery and vision is a long-standing problem in neuroscience. Currently, it is not known whether differences between the activity evoked during vision and reinstated during imagery reflect different codes for seen and mental images. To address this problem, we modeled mental imagery in the human brain as feedback in a hierarchical generative network. Such networks synthesize images by feeding abstract representations from higher to lower levels of the network hierarchy. When higher processing levels are less sensitive to stimulus variation than lower processing levels, as in the human brain, activity in low-level visual areas should encode variation in mental images with less precision than seen images. To test this prediction, we conducted an fMRI experiment in which subjects imagined and then viewed hundreds of spatially varying naturalistic stimuli. To analyze these data, we developed imagery-encoding models. These models accurately predicted brain responses to imagined stimuli and enabled accurate decoding of their position and content. They also allowed us to compare, for every voxel, tuning to seen and imagined spatial frequencies, as well as the location and size of receptive fields in visual and imagined space. We confirmed our prediction, showing that, in low-level visual areas, imagined spatial frequencies in individual voxels are reduced relative to seen spatial frequencies and that receptive fields in imagined space are larger than in visual space. These findings reveal distinct codes for seen and mental images and link mental imagery to the computational abilities of generative networks.

INTRODUCTION

Why do we have mental images, and what is their relationship to images we see with our eyes? Neuroimaging studies have demonstrated significant brain activation in visual areas during imagery [1, 2] that is specific to content [3]. Multivoxel activity patterns during imagery and vision are correlated [4–9], and visual features encoded in activity patterns during vision can be used to predict [10] or decode [11–13] activity patterns during mental imagery. However, the correlation between activity patterns during imagery and vision varies considerably across visual areas [6, 7, 14] and over time [15, 16]. Although variation may simply reflect known area-to-area differences in signal to noise during imagery [17], it may also indicate that at least some visual areas maintain different codes for seen and mental images. Distinguishing between these possibilities with existing theories and experimental methods has been challenging.

One obstacle to uncovering differences in the codes for seen and mental images is that the discriminative artificial neural networks that undergird many current predictive models of visual representation [18–22] lack a mechanism for image generation and therefore provide no insight. Generative networks [23–25] offer a compelling alternative model, because they incorporate a mechanism for generating images. Although we caution that mental images should not be taken literally as pictures sampled from a generative network, the mechanisms that generate images in these networks may be similar to those deployed in the brain.

Generative networks embody a model of the visual world that links visual stimuli to abstract visual features, often interpreted as "causes" of the stimuli. One influential theory of vision holds that the brain uses such a world model to infer the probability of abstract features in the environment, given a visual stimulus [26, 27]. We generalize this theory by interpreting imagery as inference about visual stimuli that one might see, given an assumed abstract feature. This inference is inherently useful and so provides a potential answer to the question "why do we have mental images?". Conceptually, performing this inference would require the network to run "vision in reverse" [28], i.e., treat a clamped abstract feature represented at a high level as input and then feed activity back toward the sensor level. If such generative feedback is allowed to carry all the way down to the sensor level (which can happen in artificial networks, but not in the human visual system), the network will output a synthesized picture.

Under this interpretation of mental imagery, any potential difference between the codes for seen and mental images will depend upon the relationship between stimuli and the abstract features in the network. In practice, generative networks developed for artificial intelligence (AI) applications often learn a shallow hierarchy of abstract features that offers effective image compression while maintaining a co-variant relationship to fine details of visual stimuli [25, 29]. In contrast, in the human brain, abstract features are encoded across a relatively deep hierarchy of functionally distinct visual areas. Basic coding properties vary systematically with ascension of this hierarchy, with representations becoming less spatially precise and more abstract. For example, units in high-level visual areas have larger receptive fields (for a fixed eccentricity) [18, 30–32] and tuning functions with lower spatial frequency preference [33, 34] than units in low-level visual areas. If high-level visual areas serve as the effective input during imagery, then these coding properties of seen stimuli in higher visual areas should be imparted to (or inherited by) lower visual areas during imagery. We therefore hypothesized that the codes for seen and mental images should differ substantially in low-level visual areas. Vision and mental imagery should therefore not only be marked by differences in relative signal-to-noise ratio (SNR) but by differential tuning of individual units to seen and imagined features.

To convert this hypothesis into specific and empirically testable predictions about tuning to seen and imagined features, we conducted *in silico* analyses of the activation of units in a hierarchical generative network. Specifically, we studied tuning of units in the network to imagined spatial frequency as well as receptive fields in imagined space. We induced this tuning by clamping the representation of seen stimuli at or near the top of a processing hierarchy and then implementing an inference operation that roughly corresponds to "vision in reverse." As expected, we found that units at the clamped level imparted their tuning properties to units in levels below. Based on these analyses, we made the specific prediction that imagined spatial frequencies encoded by human brain activity in low-level visual areas would be reduced relative to seen spatial frequencies, that the size of the reduction would be determined by tuning to seen spatial frequencies in a higher-level visual area, and that low-level areas would integrate information over larger regions of imagined space than visual space.

To test our predictions, we conducted an fMRI experiment in which subjects viewed and then imagined hundreds of spatially varying naturalistic stimuli. We then used brain activity to estimate independent visual- and imagery-encoding models. These encoding models accurately predicted brain responses to both seen and imagined stimuli and enabled accurate decoding of their position and content. They also revealed, for every voxel, tuning to seen and imagined spatial frequencies, as well as the location and size of receptive fields in visual and imagined space. Voxels in visual areas V4 and above (e.g., in the intraparietal sulcus) had relatively low spatial frequency preferences and large, foveally concentrated receptive fields during both imagery and vision. Remarkably, voxels in visual areas V1–V3 appeared to inherit these encoding properties during imagery, showing preference for lower imagined than seen spatial frequency and larger and more foveal receptive fields in imagined than in visual space. These results confirmed our predicted reduction in spatial frequency and enlargement of receptive fields, thus linking mental imagery in the human brain to the powerful computational abilities of generative networks.

RESULTS

A Deep Generative Network Exhibits Distinct Codes for Seen and Mental Images

We considered activity patterns in a deep generative network specified by a hierarchy of L processing levels with feedforward and feedback connections. The lowest level (level 0) of the network analogized the retina. Levels above analogized functionally distinct visual areas (e.g., V1, V2, V3, and so on; see nodes in Figure 1A). We modeled vision and mental imagery as distinct input configurations of this network and used activity patterns in the network to derive qualitative predictions about activity patterns in the brain.

During vision, the retina is activated by a visual stimulus s. Let r_k be an activity pattern at a level k and μ_k be the *expected* activity pattern at that level. We modeled vision in the generative network by clamping the activity pattern at the lowest level to the visual stimulus, $r_0^{\text{vis}} = s$ ("vis" denotes vision). Because, in this case, s is the only source of variation, the resulting expected activity pattern at level $l > 0$ is specified by a forward transformation $\mu_l^{\text{vis}} = T_l^0[s]$ that maps from level 0 (superscript) to level l (subscript). (Note that we refer to T_l^0 as a forward transform only because it yields an expected activity pattern μ_l as a function of the stimulus s. The brain is of course dynamic, and we interpret μ_l as the expected activity pattern at steady state and T_l^0 as incorporating the effects of both feedforward and feedback connections between levels in the network. See Method S1.1.4 for explicit analysis.)

During imagery experiments, the retinal stimulus is uninformative (e.g., a blank screen). We therefore modeled the visual input during imagery by clamping activity at the sensor level to a null pattern $r_0^{img} = 0$. We assumed that subjects imagine by reinstating in a high-level brain area the expected activity pattern that would have been evoked by seeing the imagined stimulus. We modeled this mechanism by clamping the activity in the network at a level higher than l, say $l + d$, to its expected activity pattern during vision, $r_{l+d}^{\text{img}} = \mu_{l+d}^{\text{vis}}$ ("img" denotes imagery). Now, the clamped activity pattern in level $l + d$ is the only source of variation, so the resulting expected activity pattern at level l is specified by a feedback transformation $\mu_l^{\text{img}} = \overline{T}_l^{l+d}\left[\mu_{l+d}^{\text{vis}}\right]$ that maps activity at level $l + d$ (superscript) to activity at level l (subscript). This feedback transformation is the operation in our model that most closely resembles "vision in reverse."

Because of the hierarchical structure of the network, the imagery activity pattern μ_l^{img} can be re-written as an explicit function of the stimulus s (see Method S1.1.2 for derivation):

$$\mu_l^{\text{img}} = \underbrace{\overline{T}_l^{l+d} \circ T_{l+d}^{l}}_{\Omega_{l,\ l+d}} \circ T_l^0[s]. \qquad \text{(Equation 1)}$$

This expression reveals how the expected imagery activity pattern μ_l^{img} differs from the expected visual activity pattern $\mu_l^{\text{vis}} = T_l^0[s]$. The difference is specified by the distortion $\Omega_{l,l+d}$, which can be construed as an "echo," because T_{l+d}^{l} maps the expected visual activity pattern at level l into the abstract representation at the clamped level $l + d$ and then $\overline{T}_l^{l+d}$ maps it back to level l. The

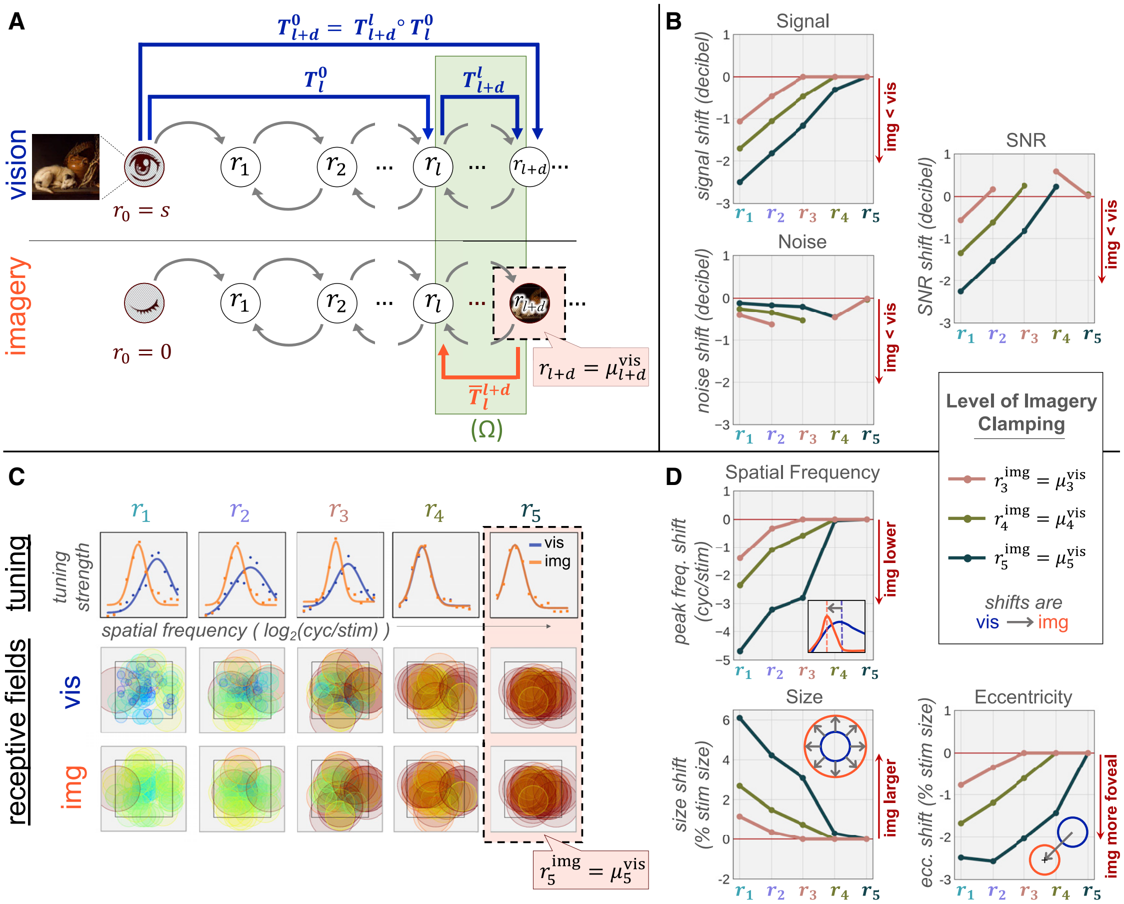

Figure 1. A Deep Generative Network Exhibits Distinct Codes for Seen and Mental Images

(A) The visual system as a deep generative network specified by a hierarchy of processing levels (circles; r_l denotes an activity pattern) and feedforward and feedback connections (gray arrows). During vision, the expected visual activity pattern at a processing level, say $l+d$, is determined by a transformation T_{l+d}^{0} (long blue arrow) of activity (denoted s for stimulus) at the sensor level 0 (eye). T_{l+d}^{0} is equivalent to a composition of transformations (shorter blue arrows) of activity patterns between intervening levels. During imagery, $s = 0$, but at least one processing level is clamped to its expected visual activity pattern ($r_{l+d} = \mu_{l+d}^{\text{vis}}$ in this example; red box). Expected imagery activity patterns beneath the clamped level (e.g., μ_l^{img}) differ from their visual activity patterns by a transformation, Ω, from the current to the clamped level (T_{l+d}^{l}, shortest blue arrow) and from the clamped level back ($\overline{T}_{l}^{l+d}$, orange arrow).

(B–D) *In silico* experiments on a deep generative network illustrate the predicted effects of Ω on brain activity patterns during mental imagery.

(B) Signal (top), noise (bottom), and signal to noise (right) during imagery relative to vision at each processing level (x axis). Changes are expressed on a power-of-2 logarithmic scale. The strength of the effect depends upon the processing level that is clamped during imagery (curves illustrate three different clamping levels; see legend).

(C) Top row shows population tuning to spatial frequency ($\log_2$[cyc /stim], x axis) for vision (blue) and imagery (orange) for units at each level in the network when the top level of the network is clamped (dashed box). Bottom two rows show receptive fields (RFs) for individual units at each level of the network (circle radius is one SD of Gaussian RF; circle color scales with radius) for vision (middle) and imagery (bottom).

(D) For each level of clamping, units below the clamped level exhibit lower spatial frequency preference (top), larger RFs (bottom), and lower eccentricity (right) during imagery relative to vision.

See Figure S1 for additional information.

encoding of s during imagery at level l is thus likely to be limited by the encoding of s at level $l+d$ during vision.

For practical applications, high-level representations in deep generative networks are often optimized to give near-lossless stimulus reconstruction. However, in high-level visual areas, brain activity can be quite invariant to many aspects of stimulus variation—an inevitable consequence of forming abstract representations that are useful for cognition [35–37]. Thus, a predicted effect of $\Omega_{l,l+d}$ was that low-level visual areas should encode variation in mental images with less precision than seen images.

To illustrate this effect, we performed *in silico* vision and imagery experiments on a specific generative network that embodied a deep latent Gaussian model [24, 38] of natural scenes. We implemented a network with linear connections between processing levels in order to obtain exact solutions for the mean activity patterns during imagery and vision. Neural units with complex-like responses were modeled by combining pairs of network units under a sum-of-squares operation (see Method S1.1.7 for details). Although processing in the human visual system undoubtedly involves more complex nonlinearities, this simple model was sufficient to capture the basic tuning properties we analyzed in our subsequent human neuroimaging experiment.

We first trained the generative network to optimize the log-likelihood of natural scenes in a large image database. A unique and essential aspect of our approach is that the training objective also constrained the network to exhibit brain-like responses to visual stimulation. Units at higher levels of the network were encouraged to exhibit lower spatial frequency preference [33, 34], a more steep receptive field size-eccentricity relation [30, 32, 39, 40], and greater foveal coverage [32, 41] *during vision* than units at lower levels. We emphasize that these constraints were placed on activity generated by the network during vision only. No explicit constraints were placed upon the activity generated by the network during imagery, and imagery activity patterns played no role in training the model.

Once training was complete, we generated a large corpus of visual and imagery activity patterns in response to a new set of natural scenes. From these activity patterns, we first estimated signal amplitude, noise, and their ratio (SNR) during vision and imagery.

Signal amplitude during imagery was attenuated relative to signal amplitude during vision at levels below the clamped level (Figure 1B, top left), and the amount of attenuation depended upon distance from the clamped level. Attenuation of signal amplitude was a direct and obvious consequence of clamping the lowest level to 0 during imagery.

Interestingly, noise was reduced during imagery at all levels (Figure 1B, bottom left). Reduction of noise is caused by clamping two levels during imagery (i.e., the lowest level plus one higher level) instead of just one (i.e., the lowest level alone). This additional clamping reduces the number of random variables in the network and therefore reduces noise.

SNR during imagery was attenuated relative to SNR during vision at levels below the clamped level (Figure 1B, right). As with signal amplitude, the amount of attenuation depended upon distance from the clamped level. Attenuation of SNR occurred because the fixed reduction in noise during imagery at all levels was small relative to the attenuation of signal amplitude in lower levels.

We then analyzed visual and imagery responses to estimate distinct spatial frequency tuning functions and receptive fields for each unit in the generative network. As anticipated, tuning during imagery was very different from tuning during vision. Spatial frequency preference was reduced, relative to vision, for units below the clamped level (Figures 1C and 1D, top). Receptive field sizes were larger, and receptive field centers were shifted toward the fovea (Figures 1C and 1D, bottom). Thus, tuning to imagined features at lower levels more closely resembled tuning to seen features at the clamped level. Like SNR, the size of these changes increased with distance from the clamped level, so that the effect of $\Omega_{l,l+d}$ was strongest in the level furthest below the clamped one. Consequently, by clamping higher in the hierarchy, the effects became stronger and more widespread; by clamping lower, the effects become weaker and restricted only to the lowest processing level (Figures 1B and 1D).

Imagery-Encoding Models Link Brain Activity to Imagined Stimuli

To test these predictions, we developed a method for inferring how mental images are encoded in human brain activity patterns. In visual neuroscience, the encoding of stimuli in brain activity patterns is revealed by visual-encoding models [18, 22, 30, 42–44]. Formally, the visual-encoding model is specified by a forward transformation from a stimulus to an expected visual activity pattern (e.g., $\mu_l^{\text{vis}} = T_l^0[s]$). The transformation determines the visual features (e.g., spatial frequency) that are extracted from the stimulus and encoded into the expected activity pattern. It also describes spatial sensitivity to features in the form of a visual receptive field.

Formulating an imagery-encoding model that would reveal tuning and spatial sensitivity to imagined features has been a formidable conceptual challenge because imagery activity is not driven by a measurable stimulus. However, an important implication of Equation 1 is that, if visual activity is reinstated *somewhere* in the network, then the imagery activity pattern μ_l^{img} can be expressed as an encoding of the stimulus s, even though s is not seen during imagery. Thus, if a set of known stimuli are imagined, it should be possible to use those stimuli to construct imagery-encoding models. Once constructed, these imagery-encoding models could be compared to visual-encoding models to test for differences in the encoding of seen and imagined stimuli. This comparison should reveal the distortion effect of $\Omega_{l,l+d}$.

To construct imagery-encoding models, we used 7-Tesla fMRI to measure whole-brain blood-oxygen-level-dependent (BOLD) activity as human subjects viewed and then in separate scans imagined 64 pictures they had memorized prior to the experiment (Figure 2A). Pictures depicted a range of natural objects, artifacts, humans, and animals (Figure 2B). On each trial, a small 6-letter cue at the center of the display indicated the picture to be displayed (imagined). The color of the cue indicated the position of the seen (imagined) picture. All pictures were displayed (imagined) at one of 8 possible positions (Figure 2C). The 8 different positions were delineated with colored brackets that were all visible at all times throughout both visual and imagery scans. Thus, subjects saw (imagined) 64 pictures at 8 positions each for a total of 512 distinct seen (imagined) stimuli across the viewing (imagery) scans. Viewing and imagery scans alternated during each experimental session. Subjects were instructed to maintain fixation on the small central cue or dummy cue at all times.

We estimated distinct voxelwise visual and imagery encoding models [22] from activity measured during viewing and imagery sessions, respectively (Figures 3 and S2). The visual-encoding model for each voxel specified tuning to seen spatial frequency and a receptive field in visual space. The imagery-encoding model for each voxel specified tuning to imagined spatial frequency and a receptive field in imagined space.

Imagery-Encoding Models Accurately Predict Brain Activity during Mental Imagery and Identify Mental Images

To validate visual-encoding models, we used them to predict brain activity evoked by stimuli not included in the training set. Visual-encoding models made accurate predictions of brain activity throughout the visual cortex up through lateral occipital cortex (LO) in the ventral stream and intra-parietal sulcus (IPS) in the dorsal stream (Figures 4A, 4B, and S6). We confirmed that visual receptive fields exhibited expected retinotopic

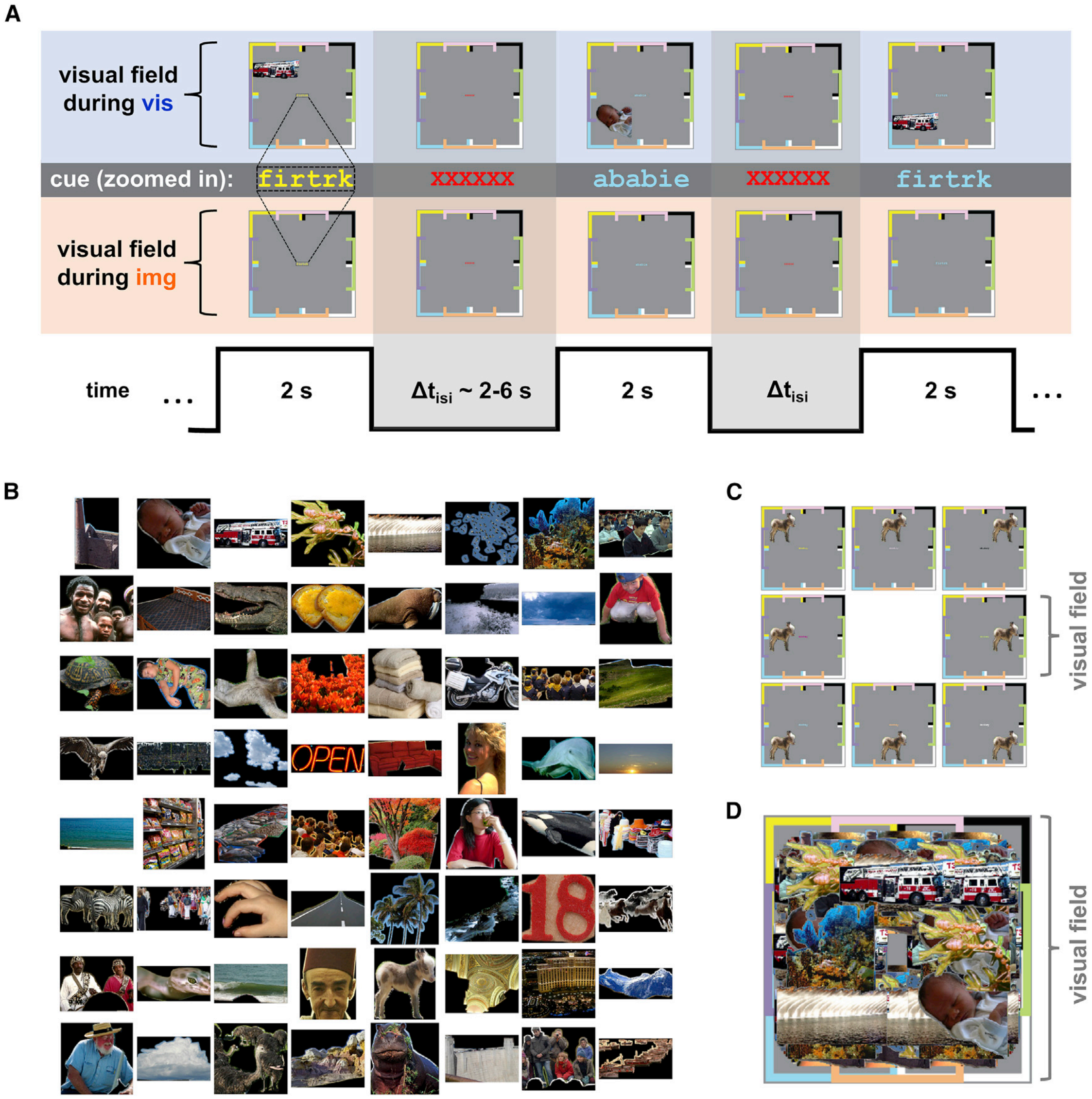

Figure 2. Stimulus Presentation and Timing

(A) Top: the stimulus displayed on the viewing screen during vision runs. Second from top: enlargements of the cues visible during both vision and imagery runs are shown. Third from top: the display during imagery runs is shown. Bottom: timing of stimulus on- and offset and interstimulus interval (Δt_{isi}) is shown.

(B) All 64 individual object pictures viewed and imagined during the experiment.

(C) An object picture displayed in each of the 8 positions bounded by the framing brackets.

(D) A superposition of all 64 object pictures showing the visual field coverage of the stimuli.

mapping (Figures S5A and S5B) and size-eccentricity relationships across distinct visual areas (Figure S5C).

To validate imagery-encoding models, we used them to predict brain activity during imagery of stimuli not included in the training set. Predictions made by imagery-encoding models exceeded a cross-validated threshold on accuracy for many voxels in all visual cortical areas considered here (Figures 4A–4C; Table S1; for subsequent analyses, we used only above-threshold voxels that, additionally, did not respond to the cue—see Voxel Selection in STAR Methods). Remarkably, cross-prediction analyses (Figure S7) revealed that imagery-encoding models were even able to accurately predict activity during vision.

Importantly, predictions of the imagery-encoding models in all regions of interest (ROIs) were accurate enough to identify the position of (Figure 4D) and the picture in (Figure 4E) the imagined stimuli. This result confirmed that subjects had indeed performed the instructed imagery task. Moreover, accurate identification of imagined pictures could not have been possible if variation in spatial attention, eye position, or visual cues were the determinants of imagery-encoding model prediction accuracy, allowing us to rule out these potential confounds. These results licensed us to inspect imagery- and visual-encoding models for evidence of the predicted differences in the encoding of seen and mental images.

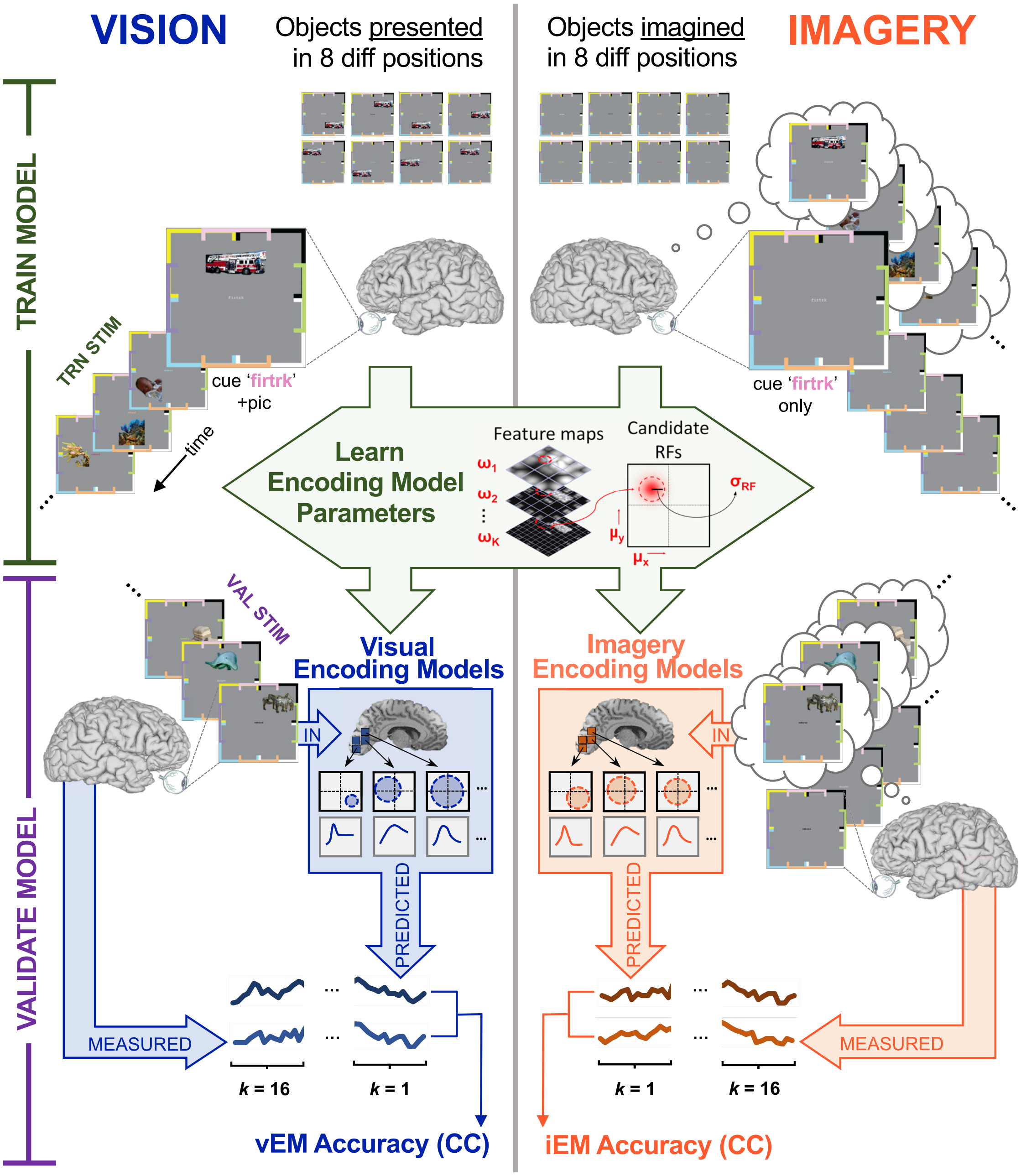

Figure 3. Data and Procedures for Estimating Visual- and Imagery-Encoding Models

Whole-brain fMRI (7T) measured blood-oxygen-level-dependent (BOLD) activity as subjects viewed (top left) or imagined (top right) 64 unique pictures at 8 distinct positions. Unlike vision runs, pictures were not displayed during imagery runs; rather, subjects imagined the picture associated with the cue in the position indicated by the cue color. Encoding models were trained (central diamond) separately using vision and imagery training data (top half), respectively, resulting in a distinct visual-encoding model (vEM) (blue box) and imagery-encoding model (iEM) (orange box) for each voxel. Each vEM/iEM (consisting of an estimated receptive field and spatial frequency tuning curve) was used to predict activity in response to held-out stimuli. Predictions were compared to measured vision/imagery brain activity (k-fold cross-validation, k = 16, bottom half) to produce a vEM/iEM prediction accuracy score (Pearson correlation) for each voxel. See Figure S2 for additional information.

Imagery-Encoding Models and Brain Activity Reveal a Hierarchical Progression of Prediction Accuracy, Signal, and Noise

Having validated the imagery-encoding models, we compared their prediction accuracy, tuning functions, and receptive fields to visual-encoding models across areas in the visual cortical hierarchy.

We expected that the prediction accuracy of the visual- and imagery-encoding models would be closest to parity in high-level visual areas, because our premise is that visual activity

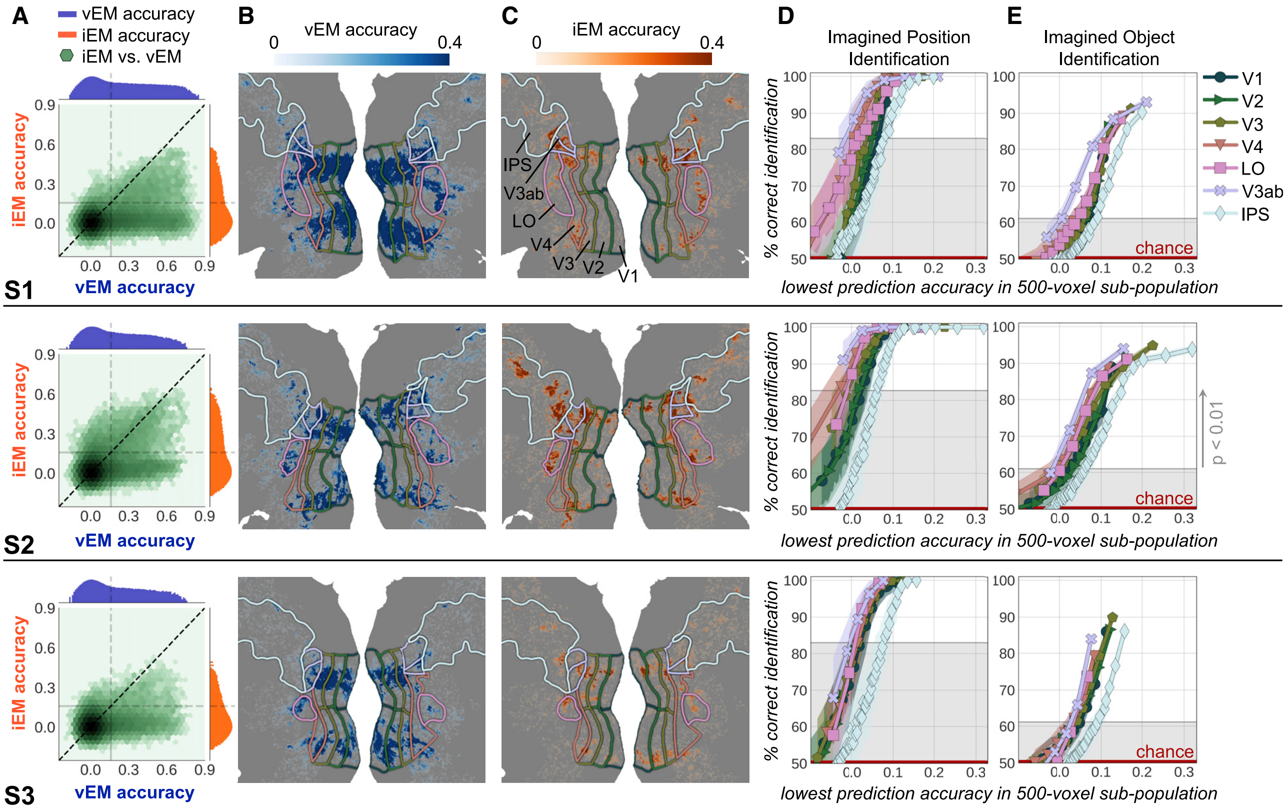

Figure 4. Imagery-Encoding Models Accurately Predict Brain Activity during Mental Imagery and Identify Mental Images

(A) Joint histogram (green) and marginal histograms of prediction accuracy for imagery- (orange) and visual (blue)-encoding models across all voxels for subjects 1–3. The iEM made accurate predictions of imagery activity (Pearson correlation ≥ 0.16; $p < 0.001$; dashed gray lines) in all subjects.

(B) Prediction accuracy (color bar) of the vEM mapped on the flattened cortical surface.

(C) iEM prediction accuracy.

(D) Imagined position identification. Percentage (500 trials) of correct pairwise identifications (y axis; colored shading indicates ± SE; gray shading indicates statistical significance threshold of $p < 0.01$; permutation test) for subpopulations of 500 voxels (markers on curves) from the indicated visual area. Identification is correct when the brain activity pattern predicted by iEMs for the true imagined position is more correlated to the measured brain activity pattern than the pattern predicted by the iEMs for a randomly selected position. Subpopulations are ordered along the x axis according to the lowest iEM prediction accuracy among all voxels in that subpopulation.

(E) Imagined object identification. Format is as in (D). Identification is correct when the brain activity pattern predicted by iEMs for the true imagined object is more correlated to the measured brain activity pattern than the pattern predicted by the iEMs for a randomly selected object.

See Figures S6 and S7 for additional data.

patterns are faithfully reinstated in one or more of these areas during imagery. In agreement with our expectations, we found that the prediction accuracy of imagery-encoding models monotonically approached parity with the prediction accuracy of visual-encoding models with ascent from V1 toward IPS (Figures 5A and 5B for subject 1 and Figure S3 for all subjects). Visual areas in IPS are at the highest level of visual processing for which our visual-encoding models generated accurate predictions (Figure S6).

The progression of relative encoding model prediction accuracy matched the progression of SNR (Figure 5C). As predicted (Figure 1B), SNR reduced with descent toward V1. The attenuation of SNR during imagery in the brain tracked an attenuation in signal amplitude (Figure 5D), although noise was uniformly reduced during imagery at all processing levels (Figure 5E), again in agreement with our predictions (see Figure 1B).

Differences in Tuning to Seen and Imagined Spatial Frequency

As predicted (Figure 1D), preferred imagined spatial frequency decreased relative to preferred visual spatial frequency with descent toward V1 (Figure 6). Consistent with our analysis of generative feedback, this decrease had the effect of bringing the preferred imagined spatial frequency in all visual areas close to the preferred seen spatial frequency in higher areas (i.e., V3A/B and IPS; see Figure 6D). Unlike encoding model prediction accuracy, loss of SNR in early visual areas cannot account for these effects (Figure S4B).

Differences in Receptive Field Size and Location for Seen and Imagined Space

We predicted that imagery receptive fields should be increasingly dilated and displaced toward the fovea relative to visual receptive fields with descent toward V1 (Figure 1D, bottom two

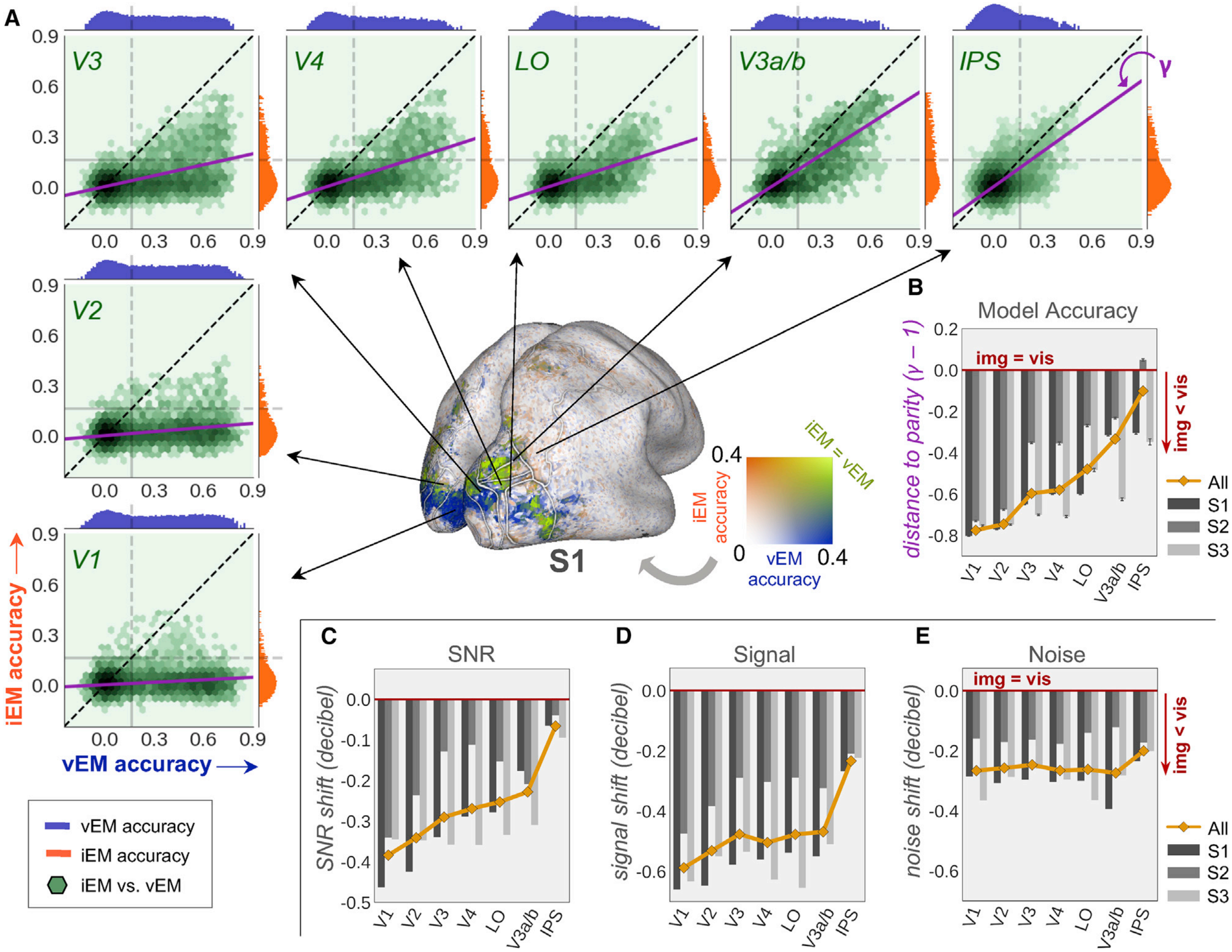

Figure 5. Hierarchical Progression of Imagery-Encoding Model Prediction Accuracy, Signal, and Noise

(A) Relative vEM and iEM accuracies (2D color map) for each voxel projected onto the inflated brain surface (center) for subject 1 (see Figure S3 for data from all subjects). Surrounding joint and marginal prediction accuracy histograms (format as in Figure 4A) show progression of relative iEM/vEM prediction accuracy (slope of best linear fit, γ; purple line) across indicated visual areas.

(B) Distance of relative iEM/vEM prediction accuracy from parity $(\gamma - 1)$ for individual subjects (bars) and all subjects combined (yellow curve).

(C) Median signal-to-noise ratio (SNR) for imagery activity relative to visual activity. See Figure S6 for cortical maps of SNR.

(D and E) (D) Relative signal, S (i.e., activation amplitude) and (E) relative noise, N. These signal and noise results qualitatively match *in silico* results in Figure 1B. See also Figure S7.

panels). After confirming that visual receptive fields conformed to known anatomical and spatial organization and size-eccentricity relationships (Figure S5), we compared the size and location of the visual receptive fields to those of imagery receptive fields. In V1, imagery receptive fields were larger (Figures 7A and 7B) and more foveal (Figures 7A, 7C, and 7D), relative to vision for each subject. Consistent with predictions, evidence for differences in receptive field eccentricity and size weakened with ascent toward high-level visual areas.

DISCUSSION

The observed differences between visual and imagery activation amplitude, noise level, SNR, encoding model prediction accuracy, spatial frequency preference, receptive field location, and size all offer novel support for the hypothesis of distinct brain activity codes for seen and mental images. Using a simple model of visual and imagery activity in a generative network, we have shown that these differences can be explained as the result of generative feedback from a clamped, high-level representation. We now discuss some implications and alternative explanations of our results.

Clamping

We argue that the similarity between imagery and vision is limited by the hierarchical level at which activity patterns are clamped. If clamping occurs in the brain, it need not be limited to a single brain area and probably has dynamics that our work sheds no light on. Nonetheless, our results localize brain activity during imagery that is consistent with our notion of clamping.

At and above V4, differences in spatial frequency tuning and receptive field attributes during mental imagery relative to vision

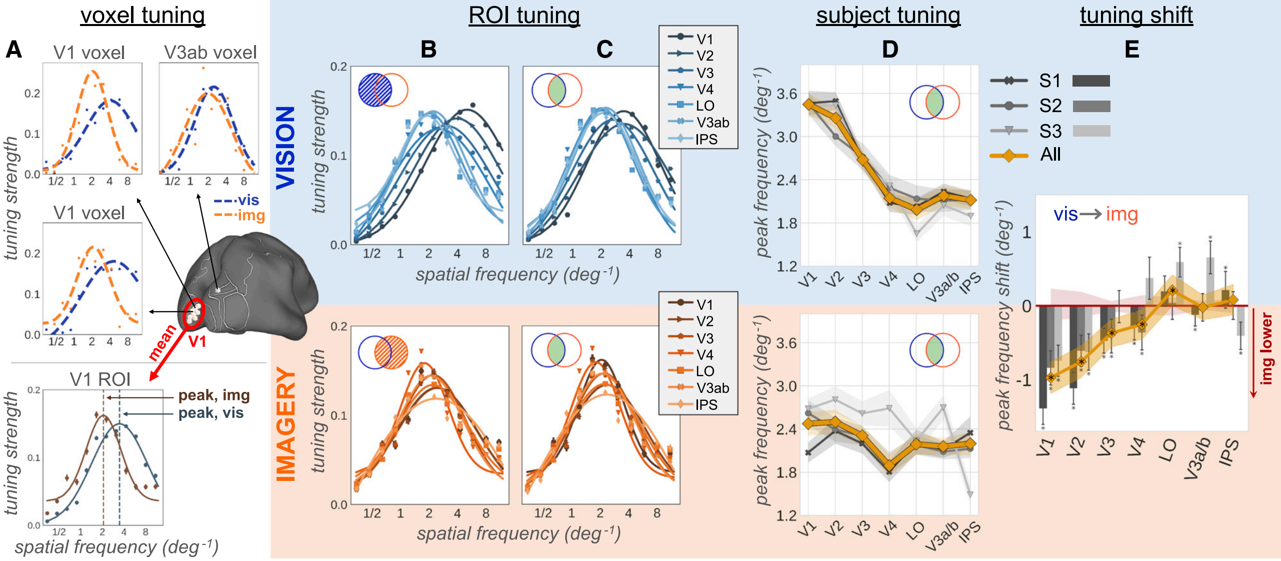

Figure 6. Differences in Tuning to Imagined and Seen Spatial Frequency

(A) Top: visual (blue) and imagery (orange) spatial frequency tuning curves for several individual voxels. Bottom: population tuning curves for voxels in V1 with an accurate iEM or vEM are shown.

(B) Top: population tuning curves during vision for all voxels in the indicated area that have an accurate vEM. Bottom: population tuning curves during imagery for voxels that have an accurate iEM are shown. Populations in top (blue circle in Venn diagram) and bottom (orange circle) plots are overlapping, but not identical.

(C) Population tuning curves for all voxels in the indicated area that have an accurate vEM *and* iEM (i.e., a completely overlapping. population of voxels). All subsequent panels use this population.

(D) Peak spatial frequency of tuning curves in (C).

(E) Difference between peak spatial frequency during imagery and vision for individual subjects (bars) and all subjects combined (yellow curve). The red shaded area indicates significance level $p < 0.01$ (permutation test) for combined subject data (yellow curve). Asterisks indicate significant difference from null value (red line; $p < 0.01$; permutation test); shading on curves indicates ± SE. Compare to *in silico* results in Figure 1D (top).

See Figure S4 for control analyses. See Table S1 for number of voxels per ROI per subject for this analysis.

were much weaker than in areas lower in the hierarchy. Voxels in IPS exhibited the strongest signal and highest SNR during imagery (Figures 5C and 5D) and showed little evidence for any differences in the encoding of seen and mental images (Figures 6 and 7). A more fine-grained analysis (Figure S3) showed a trend of decreasing prediction accuracy for both the imagery-encoding models (iEM) and the visual-encoding models (vEM) with progression from IPS0–5. Remarkably, in IPS0/1, the iEM predicted visual activity as accurately as the vEM. In IPS2–5, neither model accurately predicted either imagery or visual activity.

We found mixed evidence for engagement of prefrontal cortex during the specific mental imagery task used here. SNR was high (relative to vision) for one subject and low for another (Figure S6). The iEM and the vEM did not accurately predict imagery or visual activity in prefrontal cortex.

In all subjects, ventral stream areas immediately anterior to V4/LO exhibited relatively high SNR during the visual task, but not during imagery (Figure S6). Further anterior, there was, as with prefrontal cortex, mixed evidence for engagement during imagery.

These findings isolate IPS0/1 as a region in high-level cortex where our analyses of brain activity fail to distinguish between imagery and vision. Thus, our results point to IPS0/1 as the likely source of "clamped" activity in this particular imagery experiment. Previous findings support engagement of parietal areas during imagery [2, 45]. Additional studies with more subjects and, ideally, varying imagery tasks will be needed to confirm the generality of this finding.

Our approach assumes availability of a predictive, image-computable, visual-encoding model. Future work will require the development of models that, to our knowledge, currently do not exist (e.g., for prefrontal cortex) or require more samples of imagery activity than we acquired in this experiment (e.g., for anterior ventral stream areas) and may possibly require acquisition at higher spatial resolution (i.e., voxel volume $< (1.6\text{mm})^3$).

Alternate Explanations

Here, we consider alternate explanations of our results.

Failures of Imagination

When asked to imagine a specific remembered picture of, say, a baby, subjects might have thought of the general category "baby" without imagining details of the remembered picture. Might the reported reduction in imagined spatial frequency preference be explained as a consequence of such failures of imagination?

In our model, imagination failure can be equated with the level of clamping. Clamping high under-specifies mental images, even if the clamped activity pattern faithfully reinstates an expected visual activity pattern. Better imaginers, or ones imagining simpler stimuli, might clamp lower than the subjects in our study. This could induce imagery activity patterns in low-level visual areas that have greater overlap with corresponding visual activity patterns than we observed. Although our theory accommodates variable clamping levels (see Figures 1B and 1D), the overlap between imagery and visual brain activity in

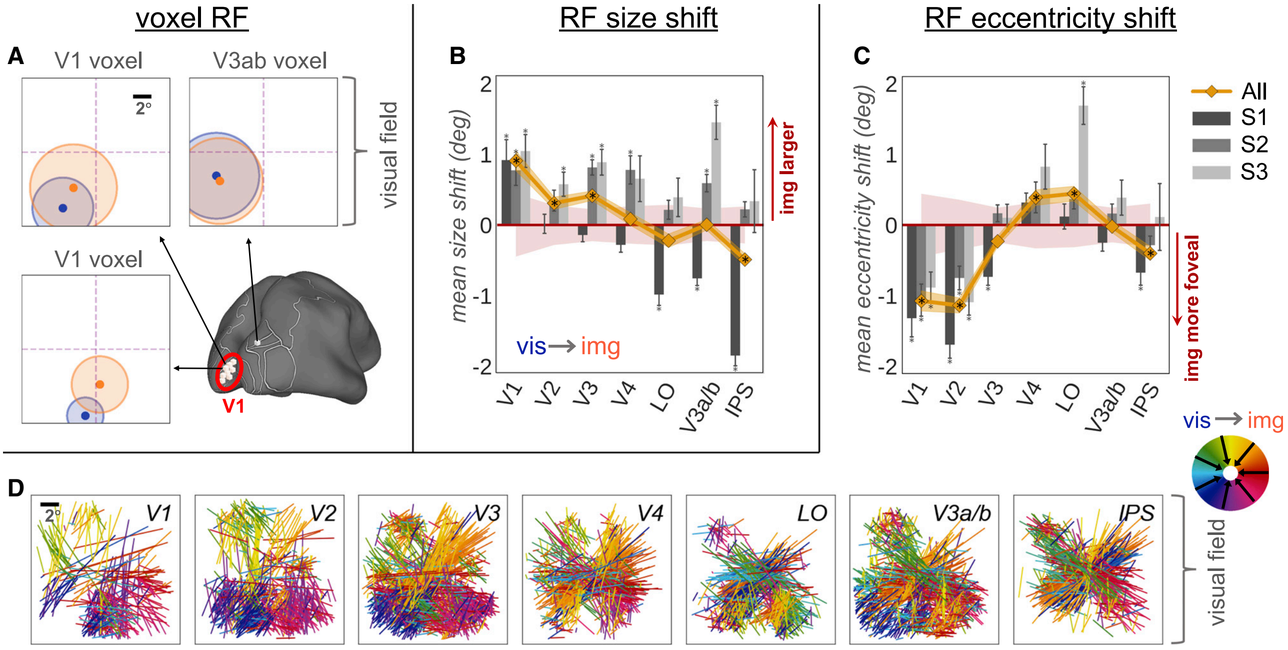

Figure 7. Differences in Receptive Field Location and Size between Vision and Imagery

(A) Example visual and imagery RFs for single voxels.

(B) Average signed change in RF size from vision to imagery. Positive (negative) values indicate dilation (shrinkage). Compare to Figure 1D (bottom left).

(C) Average signed magnitude of shift in RF location from vision to imagery. Negative values indicate a shift toward fovea. Compare to Figure 1D (bottom right). The red shaded area in (B) and (C) indicates significance level $p < 0.01$ (permutation test) for combined subject data (yellow curve). Asterisks indicate significant difference from null value (red line; $p < 0.01$; permutation test); shading on curves indicates ± SE.

(D) Orientation and magnitude (line segments) and direction (color wheel at far left) of RF location shifts (same voxels as in B and C) from vision to imagery. See Figure S5 for additional data. See Table S1 for number of voxels per ROI per subject for this analysis. See also Figure S4.

low-level visual areas is consistently weaker than in high-level visual areas [2, 17, 46, 47].

Adversarial Mental Imagery

Perhaps imagery cues trigger representations stored as associative memories in V1. Feedforward activity could then act like an adversarial image [48] that incompletely encodes details of the remembered picture but nonetheless results in the "right" abstract representations at higher levels of the network. In this case, generative feedback would not be needed for imagery.

This adversarial explanation has two flaws. First, our imagery task did not depend entirely on associative memory. Subjects did not memorize all 512 stimuli; rather, they memorized 64 pictures and a rule linking the color of cues to the position of pictures. We are unaware of evidence for such rule-based processing in human V1. Second, the adversarial explanation would require weak signals triggered in V1 to become amplified as they ascend the visual hierarchy. It is unlikely that a network that amplified signals in this way would be stable.

Weak Vision

The "weak vision" model assumes that vision and imagery are equivalent up to a bound on signal amplitude: $\mu^{\text{img}} = \varepsilon\mu^{\text{vis}} + \eta$, where $\varepsilon < 1$ and η is additive noise. This model is implied when visual-encoding models are used to predict, decode, or reconstruct mental imagery [10–13]. Our work shows that imagery may also induce tuning and receptive field changes at the level of single voxels. Thus, the lack of spatial specificity in reconstructions of mental images, relative to seen images, may be a consequence not only of noisy signal but also of changes in spatial sensitivity within cortical localities no larger than the volume of the voxels in this study ($(1.6\text{mm})^3$). This implies a limit on the specificity of mental image reconstructions that is independent of the abundance or quality of data. Our analysis further implies that this limit is imposed by the nature of visual representation.

Vision in Reverse

The "vision in reverse" (VIR) model of voluntary mental imagery holds that imagery is a form of sensory-memory recall [28]. Under this model, activity in high-level visual or mnemonic brain areas serves as the effective source of input to visual cortex. This intuitive model of mental imagery is generally consistent with our formal model. However, the VIR model does not make predictions about tuning to imagined features or receptive fields in imagined space. It does not specify whether imagery and vision are supported by a common or by separate neural populations [47], although our results support the former scenario. It does not localize the source of the "reversed" activity, although our results provide rather specific localization information (see above). Finally, the VIR model does not specify the computational significance of imagery, although our work argues that imagery may be interpreted, like vision, as a form of inference in a generative network.

Imagery and Attention

Previous fMRI studies have shown that changes in signal amplitude, receptive fields, and tuning are also induced by variations of attentional state [32, 49–54]. Consistently, these effects are small or insignificant in V1 and V2 but large and significant in higher visual areas. In contrast, in our study, the largest differences between imagery and vision were seen in V1 and V2. These differences weakened with ascent of the visual hierarchy. In addition, in our study, decreased eccentricity (of imagery receptive fields [RFs] relative to visual RFs) was accompanied by an increase in RF size, whereas in an at least one attention study [54], decreased eccentricity was accompanied by a decrease in RF size. Thus, it is clear that previously observed attention effects were not replicated in our study. Although some modeling work [55–56] suggests that inference in a generative network can explain many of the effects of attention, it is very likely attention calls into play mechanisms that do not seem to explain our imagery results. The connections between imagery and attention are an important topic for future work.

Mental Imagery versus Generative Networks in Artificial Intelligence

Mental imagery is subjectively vague, and we have shown that mental images are encoded with lower spatial frequency preference and larger receptive fields than seen images. However, samples drawn from generative networks developed for AI applications (e.g., [57]) can be quite detailed. What might explain this discrepancy?

Image synthesis in generative networks is performed by sampling from a distribution over images [57–59]. However, we explicitly do not equate such samples with mental imagery. We model the activity induced by clamping a relatively high processing level to the expected activity pattern that would have been evoked had a mental image been seen, while clamping activity at the sensor level to what is actually seen with the eyes during imagery (i.e., a blank screen). This is a principled choice because the retina (analogized by the sensor level in our model) receives no feedback and is therefore always clamped—even during imagery.

Generative networks developed in AI learn representations that are importantly different from representations in the brain. The brain maintains a relatively deep hierarchy of increasingly invariant representations [36]. The visual responses of our generative network reflect the most basic aspects of this invariance by exhibiting decreasing spatial frequency preference and increasingly large receptive fields with ascension of a network hierarchy. As we show in Figure S1, our generative network *must* exhibit these basic properties of the brain's visual responses in order to reveal the kind of tuning to imagined features that we observe in the brain.

In contrast, generative networks in AI often consist of only a single level of latent representation that is mapped to the sensor level through a nonlinear deterministic mapping. Although learning deeper hierarchies of representation in generative networks is an ongoing challenge in AI [29, 60], a hierarchy of brain-like latent representations is evidently not needed to generate detailed image samples. This suggests that optimizing the learning objective used to train many generative networks is not sufficient to learn brain-like representations. Our results further suggest that an objective that leads to deeper and more brain-like hierarchies of representation may come at the expense of less-detailed image samples [61,62] or that the details need to be "dreamed up by sampling" [63].

Mental Imagery and Inference

We have provided evidence for a link between mental imagery and generative networks. Considerable previous evidence links generative networks to vision [26, 27, 64–67]. In visual cortex, spontaneous dynamics [68], stimulus response nonlinearities [64, 69, 70], the encoding of prediction error [71, 72], and the structure of visual representations in low-level [73] and high-level [74] visual areas can be explained by modeling vision as forward inference in a generative network that embodies a probabilistic model of the visual world. Formally, this means that activity patterns during vision $(r_1, \ldots, r_L)$ are sampled from the posterior distribution $p(r_1, \ldots, r_L | s)$.

In our model, activity patterns during mental imagery can be interpreted as samples from a related but importantly different conditional distribution. In our simulations, activity patterns that encode mental images are sampled from $p(r_1, \ldots, r_{L-1} | s = 0, r_L = \mu_L^{\mathrm{vis}})$. Thus, clamping during imagery may be interpreted as probabilistic conditioning on a blank stimulus and a reinstated sensory expected activity pattern. Imagery generation may likewise be interpreted as sampling from this conditional distribution the features encoded at intervening levels $(r_1, \ldots, r_{L-1})$.

This interpretation of mental imagery as inference supports the intuition that we imagine to "see" the visual consequences of visual predictions and memories. Our work thus extends the relevance of the generative perspective on vision. Previous results relating vision to inference have supplied evidence that representations in biological visual systems are adapted to the structure of the visual environment [68, 69, 73]. Other studies linking the related concept of predictive coding to vision have supplied evidence that knowledge of the structure of the visual environment can be combined with contextual information to represent the visual structure of occluded scenes [75] and of illusory contours [76]. The current results provide additional compelling evidence that highly structured representations can emerge independently of retinal input [68, 77], allowing the visual system to reason coherently about the visual environment, even when there is nothing to see.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Subjects
- METHOD DETAILS
  - ○ Deep generative network
  - ○ Experimental design and stimuli
  - ○ MR Acquisition parameters

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Calculation of signal-to-noise ratio
  - ○ Data pre-processing
  - ○ Encoding model design and analyses of parameters
  - ○ Seen and mental image identification

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.04.014.

ACKNOWLEDGMENTS

We thank Dr. Truman Brown, Dr. Kendrick Kay, Dr. Logan T. Dowdle, and Dr. Bruno Averbeck for helpful comments and discussions on the manuscript. We gratefully acknowledge three anonymous reviewers for their helpful critiques. This work was supported by grant NIH R01 EY023384 to T.N.

AUTHOR CONTRIBUTIONS

C.A.O., J.L.B., and T.N. designed the experiment. C.A.O., J.L.B., and T.N. acquired the data. G.S.-Y. and T.N. developed the theory. J.L.B., G.S.-Y., and T.N. analyzed the data. All authors contributed to writing and editing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 20, 2019
Revised: February 3, 2020
Accepted: April 6, 2020
Published: April 30, 2020

REFERENCES

1. Wheeler, M.E., Petersen, S.E., and Buckner, R.L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. Proc. Natl. Acad. Sci. USA *97*, 11125–11129.
2. Winlove, C.I.P., Milton, F., Ranson, J., Fulford, J., MacKisack, M., Macpherson, F., and Zeman, A. (2018). The neural correlates of visual imagery: a co-ordinate-based meta-analysis. Cortex *105*, 4–25.
3. O'Craven, K.M., and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stiimulus-specific brain regions. J. Cogn. Neurosci. *12*, 1013–1023.
4. Stokes, M., Thompson, R., Cusack, R., and Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. J. Neurosci. *29*, 1565–1572.
5. Reddy, L., Tsuchiya, N., and Serre, T. (2010). Reading the mind's eye: decoding category information during mental imagery. Neuroimage *50*, 818–825.
6. Cichy, R.M., Heinzle, J., and Haynes, J.-D. (2012). Imagery and perception share cortical representations of content and location. Cereb. Cortex *22*, 372–380.
7. Lee, S.H., Kravitz, D.J., and Baker, C.I. (2012). Disentangling visual imagery and perception of real-world objects. Neuroimage *59*, 4064–4073.
8. Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., and de Lange, F.P. (2013). Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. *23*, 1427–1431.
9. Bosch, S.E., Jehee, J.F.M., Fernández, G., and Doeller, C.F. (2014). Reinstatement of associative memories in early visual cortex is signaled by the hippocampus. J. Neurosci. *34*, 7493–7500.
10. Naselaris, T., Olman, C.A., Stansbury, D.E., Ugurbil, K., and Gallant, J.L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage *105*, 215–228.
11. Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage *33*, 1104–1116.
12. Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. *8*, 15037.
13. Senden, M., Emmerling, T.C., van Hoof, R., Frost, M.A., and Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. Brain Struct. Funct. *224*, 1167–1183.
14. Dijkstra, N., Bosch, S.E., and van Gerven, M.A.J. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas. J. Neurosci. *37*, 1367–1373.
15. Dentico, D., Cheung, B.L., Chang, J.-Y., Guokas, J., Boly, M., Tononi, G., and Van Veen, B. (2014). Reversal of cortical information flow during visual imagery as compared to visual perception. Neuroimage *100*, 237–243.
16. Dijkstra, N., Mostert, P., DeLange, F.P., Bosch, S., and van Gerven, M.A. (2018). Differential temporal dynamics during visual imagery and perception. eLife *7*, e33904.
17. Pearson, J., Naselaris, T., Holmes, E.A., and Kosslyn, S.M. (2015). Mental imagery: functional mechanisms and clinical applications. Trends Cogn. Sci. *19*, 590–602.
18. Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.
19. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624.
20. Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446.
21. Güçlü, U., and van Gerven, M.A.J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. *35*, 10005–10014.
22. St-Yves, G., and Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. Neuroimage *180* (Pt A), 188–202.
23. Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience (MIT).
24. Bishop, C.M. (2006). Pattern Recognition and Machine Learning (Springer).
25. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT).
26. Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836.
27. Lee, T.S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. J. Opt. Soc. Am. A Opt. Image Sci. Vis. *20*, 1434–1448.
28. Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. Nat. Rev. Neurosci. *20*, 624–634.
29. Zhao, S., Song, J., and Ermon, S. (2017). Learning hierarchical features from deep generative models. In Proceedings of the 34th International Conference on Machine Learning (Journal of Machine Learning Research), pp. 4091–4099.
30. Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.
31. Grill-Spector, K., and Malach, R. (2004). The human visual cortex. Annu. Rev. Neurosci. *27*, 649–677.
32. Kay, K.N., Weiner, K.S., and Grill-Spector, K. (2015). Attention reduces spatial uncertainty in human ventral temporal cortex. Curr. Biol. *25*, 595–600.
33. Broderick, W.F., Benson, N.C., Simoncelli, E.P., and Winawer, J. (2018). Mapping spatial frequency preferences in the human visual cortex. In Annual Meeting, Vision Sciences Society (Vision Sciences Society).
34. Henriksson, L., Nurminen, L., Hyvärinen, A., and Vanni, S. (2008). Spatial frequency tuning in human retinotopic visual areas. J. Vis. *8*, 1–13.

35. Rust, N.C., and Dicarlo, J.J. (2010). Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. J. Neurosci. *30*, 12978–12995.

36. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

37. Leibo, J.Z., Liao, Q., Anselmi, F., and Poggio, T. (2015). The invariance hypothesis implies domain-specific regions in visual cortex. PLoS Comput. Biol. *11*, e1004390.

38. Frey, B.J., and Hinton, G.E. (1999). Variational learning in nonlinear gaussian belief networks. Neural Comput. *11*, 193–213.

39. Kay, K.N., Winawer, J., Mezer, A., and Wandell, B.A. (2013). Compressive spatial summation in human visual cortex. J. Neurophysiol. *110*, 481–494.

40. Grill-Spector, K., Weiner, K.S., Gomez, J., Stigliani, A., and Natu, V.S. (2018). The functional neuroanatomy of face perception: from brain measurements to deep neural networks. Interface Focus *8*, 20180013.

41. Amano, K., Wandell, B.A., and Dumoulin, S.O. (2009). Visual field maps, population receptive field sizes, and visual field coverage in the human MT+ complex. J. Neurophysiol. *102*, 2704–2718.

42. David, S.V., and Gallant, J.L. (2005). Predicting neuronal responses during natural vision. Network *16*, 239–260, 239–260.

43. Wu, M.C.-K., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. *29*, 477–505.

44. Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron *63*, 902–915.

45. Ishai, A., Ungerleider, L.G., and Haxby, J.V. (2000). Distributed neural systems for the generation of visual images. Neuron *28*, 979–990.

46. Kosslyn, S.M., and Thompson, W.L. (2003). When is early visual cortex activated during visual mental imagery? Psychol. Bull. *129*, 723–746.

47. Dijkstra, N., Bosch, S.E., and van Gerven, M.A.J. (2019). Shared neural mechanisms of visual perception and imagery. Trends Cogn. Sci. *23*, 423–434.

48. Zhou, Z., and Firestone, C. (2019). Humans can decipher adversarial images. Nat. Commun. *10*, 1334.

49. Womelsdorf, T., Anton-Erxleben, K., Pieper, F., and Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. Nat. Neurosci. *9*, 1156–1160.

50. Çukur, T., Nishimoto, S., Huth, A.G., and Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. *16*, 763–770.

51. Klein, B.P., Harvey, B.M., and Dumoulin, S.O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. Neuron *84*, 227–237.

52. Vo, V.A., Sprague, T.C., and Serences, J.T. (2017). Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. J. Neurosci. *37*, 3386–3401.

53. Klein, B.P., Fracasso, A., van Dijk, J.A., Paffen, C.L.E., Te Pas, S.F., and Dumoulin, S.O. (2018). Cortical depth dependent population receptive field attraction by spatial attention in human V1. Neuroimage *176*, 301–312.

54. van Es, D.M., Theeuwes, J., and Knapen, T. (2018). Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. eLife *7*, e36928.

55. Chikkerur, S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a Bayesian inference theory of attention. Vision Res. *50*, 2233–2247.

56. St-Yves, G., and Naselaris, T. (2019). Cognition as inference: A unifying account of some neural effects associated with mental imagery and attention. 2019 Conference on Cognitive Computational Neuroscience. https://doi.org/10.32470/CCN.2019.1388-0. https://ccneuro.org/2019/proceedings/0000287.pdf.

57. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, eds. (MIT Press), pp. 2672–2680.

58. Rezende, D.J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. arXiv, arXiv:1401.4082v3. https://arxiv.org/abs/1401.4082.

59. Kingma, D.P., and Welling, M. (2013). Auto-encoding variational Bayes. arXiv, arXiv:1312.6114v10. https://arxiv.org/abs/1312.6114.

60. Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. (2019). BIVA: a very deep hierarchy of latent variables for generative modeling. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates), pp. 6548–6558.

61. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR. https://openreview.net/pdf?id=Sy2fzU9gl.

62. Alemi, A.A., Fischer, I., Dillon, J.V., and Murphy, K. (2016). Deep variational information bottleneck. ICLR. https://arxiv.org/pdf/1612.00410.pdf.

63. Zhao, J., Mathieu, M., Goroshin, R., and LeCun, Y. (2015). Stacked what-where auto-encoders. arXiv, arXiv:1506.02351v8. http://arxiv.org/abs/1506.02351.

64. Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

65. Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. *10*, 301–308.

66. Bar, M. (2009). The proactive brain: memory for predictions. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 1235–1243.

67. Spratling, M.W. (2016). Predictive coding as a model of cognition. Cogn. Process. *17*, 279–305.

68. Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science *331*, 83–87.

69. Karklin, Y., and Lewicki, M.S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. Nature *457*, 83–86.

70. Coen-Cagli, R., Kohn, A., and Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. Nat. Neurosci. *18*, 1648–1655.

71. Murray, S.O., Kersten, D., Olshausen, B.A., Schrater, P., and Woods, D.L. (2002). Shape perception reduces activity in human primary visual cortex. Proc. Natl. Acad. Sci. USA *99*, 15164–15169.

72. Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. J. Neurosci. *30*, 2960–2966.

73. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.

74. Stansbury, D.E., Naselaris, T., and Gallant, J.L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron *79*, 1025–1034.

75. Muckli, L., De Martino, F., Vizioli, L., Petro, L.S., Smith, F.W., Ugurbil, K., Goebel, R., and Yacoub, E. (2015). Contextual feedback to superficial layers of v1. Curr. Biol. *25*, 2690–2695.

76. de Haas, B., and Schwarzkopf, D.S. (2018). Spatially selective responses to Kanizsa and occlusion stimuli in human visual cortex. Sci. Rep. *8*, 611.

77. Vetter, P., Smith, F.W., and Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. Curr. Biol. *24*, 1256–1262.

78. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., and Smith, S.M. (2012). FSL. Neuroimage *62*, 782–790.

79. Kay, K.N., Rokem, A., Winawer, J., Dougherty, R.F., and Wandell, B.A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front. Neurosci. *7*, 247.

80. Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

81. Gao, J.S., Huth, A.G., Lescroart, M.D., and Gallant, J.L. (2015). Pycortex: an interactive surface visualizer for fMRI. Front. Neuroinform. *9*, 23.

82. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., and Torralba, A. (2010). SUN database: large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE), pp. 3485–3492.

83. Jezzard, P., and Balaban, R.S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. Magn. Reson. Med. *34*, 65–73.

84. Engel, S.A., Glover, G.H., and Wandell, B.A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex *7*, 181–192.

85. Hansen, K.A., Kay, K.N., and Gallant, J.L. (2007). Topographic organization in and near human visual area V4. J. Neurosci. *27*, 11896–11911.

86. Wang, L., Mruczek, R.E.B., Arcaro, M.J., and Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. Cereb. Cortex *25*, 3911–3931.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
FSL (5.0.9)	[78]	https://fsl.fmrib.ox.ac.uk/fsl/
AnalyzePRF (1.1)	[39]	https://kendrickkay.net/analyzePRF/
GLMDenoise (1.4)	[79]	https://kendrickkay.net/GLMdenoise/
Blender (2.78)	The Blender Foundation	https://www.blender.org/
Freesurfer (6.0.0)	[80]	https://surfer.nmr.mgh.harvard.edu/
PyCortex (1.1dev0)	[81]	https://github.com/gallantlab/pycortex
Custom Python 2.7 code	This article	https://github.com/styvesg/imagery-master

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources/code should be directed to the Lead Contact, Thomas Naselaris (tnaselar@musc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Python source code and notebook examples generated during this study are available on GitHub at https://github.com/styvesg/imagery-master. Neural datasets and stimuli have not been deposited in a public repository due to file size but are available from the Lead Contact on request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects

Three healthy adult subjects (2 females; age range: 26-40, mean age = 35.33) with normal or corrected-to normal vision participated in the study. Two subjects were authors. All subjects gave written, informed consent approved by the Institutional Review Boards at the University of Minnesota before participating in the study.

METHOD DETAILS

Deep generative network

The generative network used in the numerical illustrations in Figure 1 was a deep latent Gaussian model [24, 38]. The network contained $L = 5$ layers with 1736, 1352, 1096, 840 and 584 units, respectively. Collectively, the network specifies a distribution over the activities of units at each level $p(r_0, \ldots, r_L)$. We assume this joint distribution factors across levels

$$p(r_0, r_1, \ldots r_L) = p(r_L) \prod_{l=0}^{L-1} p(r_l | r_{l+1})$$

where the conditional distribution $p(r_l | r_{l+1})$ is Gaussian and has a mean that is determined by a linear transformation of the activity pattern in the level above it, $\mu_{l|l+1} = U_{l+1} r_{l+1}$, where U_{l+1} is a set of connection weights between level l and level $l+1$. The prior $p(r_L)$ is assumed to have zero mean and a diagonal covariance matrix. We assumed linear connections in order to take advantage of exact inference, and to expose feedback effects that did not depend upon a specific nonlinear interaction between levels. Further details on the structure of the model are provided in Methods S1.

The connection weights between levels were learned by optimizing an objective function that maximizes a lower bound on the log-likelihood of 50,000 small natural scenes, as well as the correspondence between units in the network and idealized neurons that exhibited spatial frequency tuning functions and receptive field sizes similar to those observed in the human visual cortex. Details on the constrained objective function and training algorithm are provided in Method S1.1.6. Details on the idealized neurons used to train the deep generative network (DGN) are provided in Method S1.1.5.

Once trained, we obtained expected visual activity patterns at each level by computing the expectation of the posterior distribution:

$$\mu_l^{\text{vis}} \equiv \mathbb{E}_{p(r_1,\ldots,r_L|r_0=s)}[r_l]$$

where the stimuli s were drawn from a validation set.

We obtained expected imagery activity patterns at each level by computing the expectation of a modified posterior that was further conditioned on a "clamped" high-level activity pattern:

$$\mu_l^{\text{img}} \equiv \mathbb{E}_{p\left(r_{l\ \{0,k\}}|r_k=\mu_k^{\text{vis}},r_0=0\right)}[r_l]$$

Formulas for calculating these expectations are provided in Methods S1.1.3 and S1.1.4.

This procedure gave us 10,000 imagery and 10,000 visual activity patterns. These activity patterns were then subjected to the same analysis pipeline as experimentally measured human brain activity patterns. Details of these analyses are provided below and in Method S1.1.8.

Experimental design and stimuli

The experimental scans were organized into separate 10-minute runs, each an uninterrupted succession of trials during which whole-brain BOLD activity was measured. Runs were of two types: vision runs and imagery runs. Each subject completed both vision and imagery runs. During vision runs, stimuli were presented on a viewing screen and viewed by the subjects. During imagery runs stimuli were not presented (only a cue was shown) on the viewing screen but were imagined by subjects (Figures 2A and 3). Stimuli were presented by a NEC NP4000 projector (1024×768 resolution, 60Hz framerate) onto a rear projection screen. During all runs, subjects fixated on a 6-letter cue (filling a 1.5° × 0.4° rectangle) at the center of a gray stimulus field (16° × 16°). Eight framing brackets with 8 distinct colors were displayed throughout each run. Each bracket bounded a different but overlapping portion of the stimulus field (8° × 8°) that framed a seen (vision runs) or imagined (imagery runs) object picture. Framing brackets were visible at all times during all runs and conditions. Cues were 6-letter descriptive abbreviations (e.g., 'firtrk' cued a picture of a fire truck, 'ababie' cued a picture of a baby) and always appeared at the same location and with the same dimensions (Figure 2).

During vision runs (Figure 2A, blue panel and Figure 3, left panel), the central cue and an object picture were displayed simultaneously. Subjects were instructed to fixate the cue and passively view the object pictures. The picture-cue pair was presented for a duration of 2 s and was followed by an inter-stimulus interval (ISI) during which no picture was shown and the 6-letter cue at center changed to a dummy cue ("XXXXXX"). The duration of the ISI varied randomly from 1 TR (2 s) to $j \times TR$ where j was sampled from a Poisson distribution ($\lambda = 0.4$; ≈ 2 to 6s).

During imagery runs (Figure 2A, orange panel and Figure 3, right panel), subjects viewed the same display as during the vision runs (i.e., the same cues, background, and framing brackets) except that the object picture itself was absent. Subjects were instructed to fixate on the cue and mentally project the cued object onto the portion of the visual field framed by the bracket whose color matched the color of the cue. For example, the cue 'firtrk' written in yellow cued the subject to imagine the fire truck picture in the upper left corner within the yellow framing brackets. Subjects were instructed to imagine the object for the duration of the 2 s trial and to stop imagining it when the object cue reverted to the dummy cue.

Objects (Figure 2B) were selected from the SUN labeled image collection [82] and were selected to span the 19 object categories specified in Naselaris et al. [44]. Each object was extracted from its background using the object mask provided by the SUN database. Masks were dilated by 10 pixels.

Eight unique object pictures were displayed during each run. Each object picture was displayed at each of the 8 framed locations (Figure 2C), for a total of 64 unique stimuli in a run, which were repeated twice per run for a total 128 stimulus presentations per run. There were 8 vision runs, for a total of 512 unique seen stimuli and 1024 stimulus presentations. There were also 8 imagery runs for a total of 512 unique imagined stimuli, and 1024 unique acts of mental imagery.

Prior to each session subjects familiarized themselves with the experimental stimuli using a self-paced version of the imagery experiment. Familiarization sessions halted when subjects felt confident that the 8 object pictures and associated 6-letter cues were committed to memory. These sessions varied in duration from 10-20 minutes.

MR Acquisition parameters

7T MRI data was acquired at the Center for Magnetic Resonance Research (CMRR) at the University of Minnesota. The experimental fMRI runs were collected using a 7T Siemens Magnetom scanner and a Nova Medical head coil (CP Transmit / 32 channel receive coil). Whole-brain functional data was acquired with a gradient-echo EPI sequence at a resolution of $(1.6\text{mm})^3$: TR 2000 ms, TE 22.8 ms, FOV 130×130, Partial Fourier $7/8$, 70 slices, GRAPPA R = 2, multiband acceleration factor 2, anterior-posterior phase encode, transverse slice orientation.

Prior to experimental runs we collected a 1-mm T1-weighted whole-brain anatomical volume at 7T for all subjects. We also collected whole-brain fieldmap phase and magnitude volumes for the correction of EPI spatial distortions [83].

QUANTIFICATION AND STATISTICAL ANALYSIS

Calculation of signal-to-noise ratio

Signal amplitude for units in the DGN model was defined as the variance of the expected value of the activity across a given dataset (i.e., "conditions"). The noise level was estimated from the average of the variance of the activity across that same dataset. Thus, the signal-to-noise ratio (SNR) per unit was

$$\mathrm{SNR}_{\mathrm{model}} = \frac{\mathrm{var}[\mu]_{\mathrm{conditions}}}{\mathrm{avg}[\mathrm{diag}[\Sigma]]_{\mathrm{conditions}}}, \qquad \text{(Equation 2)}$$

where μ and Σ are the expected value and the covariance matrix of the activity in a given condition. This parallels the way that SNR is estimated experimentally through bootstrapping the activation amplitudes over separate runs with the same condition. In this case, the SNR is defined as

$$\mathrm{SNR}_{\mathrm{exp}} = \frac{\mathrm{var}\left[\mathrm{avg}[\beta]_{\mathrm{trial}}\right]_{\mathrm{conditions}}}{\mathrm{avg}\left[\mathrm{var}[\beta]_{\mathrm{trial}}\right]_{\mathrm{conditions}}}, \qquad \text{(Equation 3)}$$

where β is the measured activation amplitude (beta value) and the subscript "trial" (bootstrap iteration) and "conditions" stands for the dimension with which the average (variance) is taken over. Figure 1B illustrates the characteristic tendencies of the signal, noise and SNR ratios derived from the solution of the trained DGN.

Data pre-processing

Functional image correction and alignment

Functional scans were corrected and aligned within subject only. For each run, time series motion correction was performed through rigid alignment of all volumes to the middle volume (FSL MCFLIRT). Acquired fieldmaps were then used for spatial B0 distortion correction (FSL FUGUE). Functional volumes were temporally resampled to correct for slice timing differences (FSL slicetimer). Spatial transformations up to this pre-processing stage were then concatenated and applied to un-corrected and un-registered volumes to minimize spatial resampling. An average of the time series from the run with the least amount of absolute movement was selected as the reference image for rigid alignment between runs (FSL FLIRT). Any residual misalignment was reduced via non-linear registration of all functional volumes to the same reference image (FSL FNIRT). Transforms for the last two registrations were concatenated and applied to the within-scan corrected images.

Time-series denoising and beta estimation

BOLD time-series modeling for each voxel in the corrected and registered functional volumes was performed using GLMdenoise [79] (Canonical HRF, visual cortex mask for PC-selecting voxels, noise-pool threshold defined as 99th percentile of R^2 values, minimum of 700 voxels with highest R^2 selected from visual cortex used to select number of principal components, 100 bootstrapping iterations). For each voxel this procedure output an estimate of activation amplitude (beta value) per unique seen stimulus and an independent estimate of activation amplitude per unique imagined stimulus. Activation estimates were bootstrapped to obtain confidence intervals.

Surface reconstructions

Structural T1 volumes were skull-stripped and used to obtain surface reconstructions (Freesurfer). Flatmaps used for displaying results and drawing retinotopic regions of interest (see section Region of interest identification) were prepared with pycortex [81]. Briefly, T1-weighted volumes were passed to Freesurfer's recon-all (version 6) for cortical reconstruction and segmentation, pial and white-matter surface rendering, and cortical inflation. We then made manual edits to the segmentations to ensure optimal surface quality. Digital cuts were made into the inflated surface using Blender (v2.78) and then processed by pycortex for flattening and rendering. Functional data to be displayed on surfaces were rigidly aligned to the above processed structural volumes using FSL FLIRT.

Region of interest identification

We conducted independent retinotopic mapping experiments to identify visual areas V1, V2, V3, V3a/b, V4, and LO. We utilized the mapping stimuli and population receptive field estimation (analyzePRF) technique from Kay et al. [39] to construct angle and eccentricity maps for subjects 1 and 3. Similar retinotopic maps were constructed using a standard traveling wave approach [84] for subject 2. These maps were overlaid onto flattened cortical surfaces and imported into Inkscape where phase reversals and eccentricity patterns were used to hand-draw continuous regions of interest (ROI) boundaries as described in Hansen et al. [85]. Ventral and dorsal regions were delineated for V1, V2, and V3. Surface-defined ROIs were then transformed back to functional 3D volumetric space using pycortex (get_roi_masks function with gm_sampler = 'thick'). A cortical ribbon mask (adopted from Freesurfer's earlier segmentation) was prepared for each ROI.

To identify cortex within the intraparietal sulcus (IPS) functional volumes were registered to MNI $(1\mathrm{mm})^3$ standard space (FSL FNIRT, 3mm warp resolution). ROI's were then defined using published probabilistic maps of ROIs in volumetric standard space [86]. These probabilistic maps were thresholded at 10%; any voxel belonging to multiple ROIs under this threshold was assigned to the ROI for which it had the highest probability of membership. The registration transform was then inverted to bring these ROIs from standard space back into individual subjects' native spaces.

Encoding model design and analyses of parameters

The voxel-wise fwRF encoding model

Imagery encoding models and visual encoding models were estimated using the feature-weighted receptive field (fwRF) approach [22] summarized in Figure S2. The fwRF is a voxel-wise encoding model, meaning that it can be used to predict activation in response to arbitrary stimuli for each voxel in a functional volume. The predictions are generated according to the following separable linear model:

$$\widehat{r}_t = \sum_{k=1}^{K} w_k \int_{-\frac{D}{2}}^{\frac{D}{2}} \int_{-\frac{D}{2}}^{\frac{D}{2}} g\left(x, y; \mu_x, \mu_y, \sigma_{\text{f.p.f.}}\right) \phi^k_{i(x)j(y)}(s_t) dxdy \qquad \text{(Equation 4)}$$

where $\widehat{r}_t$ is the predicted activity in response to image s_t and D is the total visual angle sustained by the image s_t. The discretization depends on the resolution of the feature map under consideration such that $i(x) = \lfloor (2x + D)/2\Delta \rfloor$ (likewise for $j(y)$) where $\Delta = D/n_k$ is the visual angle sustained by one pixel of a feature map with resolution $n_k \times n_k$. The size, $\sigma_{\text{f.p.f}}$, of the Gaussian pooling field $g(\cdot)$ is tied to the size of receptive field, σ_{RF}, through $\sigma^2_{\text{f.p.f}} + \sigma^2_{\text{p.f.}} = \sigma^2_{\text{RF}}$ where $\sigma_{\text{p.f.}}$ is the size of the Gabor wavelet envelope for the feature being pooled over (a graphical representation of these terms is shown in Figure S2). The fwRF model is an estimate, for each voxel, of a spatial receptive field's parameters and a separable visual feature tuning weight vector $\mathbf{w} = (w_1, \ldots, w_K)$. The receptive field location (μ_x, μ_y) and size σ_{RF} are estimated using a brute-force search over a grid of locations and sizes, while the parameters of the tuning function are learned via gradient descent. See Figure S5 for an example of the retinotopic organization of these estimates.

In this study visual features were constructed by Gabor wavelet transform of the visual stimuli at 4 uniformly distributed orientations times 12 spatial frequencies (log-spaced between $\omega = 0.35$ and $\omega = 11.0\ deg^{-1}$) for a total of 48 features. Gabor wavelets spanned 4 s.d. of the Gaussian envelope and were designed to have one cycle per s.d. Receptive fields were optimized over a uniform grid of 21×21 receptive locations times 12 receptive field sizes (log-spaced between 0.22 and 8.75 degrees of visual angle).

Training and cross-validation

The fwRF was applied to vision and imagery datasets independently. Data from the visual runs were used to estimate the visual encoding models, while data from the the imagery runs were used to estimate the imagery encoding models. In both cases, the input of the encoding models s_t during training is the exact visual stimulus that was presented during the visual experiment (including the cue). All of the following details therefore apply to both instances of fwRF training.

Feature weights and receptive field parameters were estimated using a k-fold cross-validation procedure. For each fold, validation subsets ($N_{\text{val}} = 32$) were sampled from the $N_{\text{total}} = 512$ samples of activation per voxel. Each validation subset contained activation in response to randomly sampled picture-locations that did not overlap across folds. For each fold, activation samples outside of the validation subset were treated as training data used to estimate a fwRF model for each voxel. This process was repeated $k = 16$ times resulting in 16 distinct fwRF models for each voxel.

Estimation of the feature weights for each model was performed using stochastic gradient descent with a learning rate of 5×10^{-3} and with a batch size of 96 for a maximum of 100 epochs. 40% of the training data was held-out as an early stopping set. Parameter updating halted early if the held-out loss began to increase. Estimation of the receptive field location and size was performed by brute-force search over the minimum hold-out loss reached by all possible candidates on a grid (see Figure S2).

To estimate model prediction accuracy, predictions from the 16 distinct models for each activation in the corresponding validation subsets were concatenated. A Pearson correlation coefficient between the model predictions and measured activations was calculated for each voxel. Prediction accuracy are shown in Figures 4 and 5. Error estimates on prediction accuracy values were obtained by sampling 100 times with replacement the $k = 16$ models for each voxel and recalculating the correlation coefficient for each sample.

We also developed a variant of the fwRF model that used ridge regression to optimize feature weights (instead of gradient descent). This variant offers a significant reduction in run time and was therefore used to create the cortex-wide maps of model prediction accuracy in Figures 5 and S6. The ridge parameter was selected from a range of 12 log-spaced value between 10^{-3} and 10^{3} based on validation accuracy on a holdout set size of 40% of the training data. The selected ridge parameter was such that the large majority fell within the extrema of the range above. We confirmed that results obtained with this variant were consistent with results obtained using gradient descent.

Voxel selection

Maps in Figures 4B and 4C project onto a flattened surface the encoding model prediction accuracy of all gray matter voxels with $\geq 10\%$ chance of belonging to any of the visual areas defined in the probabilistic atlas of Ref [86].

A significance threshold on prediction accuracy (dashed gray lines in Figures 4A, 5A, and S3) was defined such that values more extreme have $p < 10^{-3}$ for a null distribution over prediction accuracy that assumed no relationship between the model predictions and measured activities. This null distribution was built through 500 iterations of shuffling the model's predicted activity over conditions for each voxel and then measuring the correlation coefficient between this shuffled predicted activity and the corresponding measured activity for that voxel. Unless otherwise specified, analyses of receptive field attributes and spatial frequency tuning were applied only to voxels with a visual or imagery encoding model above this accuracy threshold (Pearson correlation coefficient ≥ 0.16).

To ensure that our results did not reflect any response to the slight changes in hue and shape of the cue with condition, the following procedure was used to identify and discard any voxel that showed sensitivity to the 6-letter cues during either the imagery or the visual runs.

The visual stimuli used to estimate each fwRF model included both the object picture and its associated cue. To test for sensitivity to the cue, we calculated cross-validated prediction accuracy using input stimuli that contained either the cues or the object pictures only. Voxels for which the cue-only stimuli resulted in above-threshold prediction accuracy for either the imagery or the visual encoding model were discarded from any analysis of receptive field attributes or feature tuning. Only voxels for which the picture-only stimuli resulted in above-threshold prediction accuracy were retained for receptive field and feature tuning analyses.

Table S1 enumerates the number of voxels per ROI and subject that satisfied these conditions. In Figure 6B of the main text, we retain voxels for which either the visual (top) or the imagery (bottom) encoding models had above-threshold predication accuracy. In Figures 6C–6E we retain only voxels for which both the visual and imagery encoding models had above-threshold prediction accuracy.

Figure S4A shows receptive fields of the discarded cue-responsive voxels. As expected, receptive fields tend to be small and are concentrated at the center of the visual field where the cue was displayed.

Spatial frequency tuning

Spatial frequency tuning was determined using a feature weight dropout procedure [22]. To determine the tuning strength to spatial frequency for each voxel, we first calculated the Pearson correlation coefficient (ρ) obtained by using feature weights associated with Gabor wavelets of a single spatial frequency alone (i.e., setting all weights to 0 except those belonging to feature maps generated using Gabor wavelets of a specific frequency). The value of ρ^2 was calculated for each spatial frequency and can be interpreted as a percentage of variance explained [42] by that frequency. In order to compare voxels to each other we normalized these frequency tuning curves to make them independent of the total variance explained. Thus, the tuning function was defined as the square of ρ for each specific frequency divided by the sum of the square of these ρ's over all frequencies (Figure 6A, top 3 plots, show examples for individual voxels). Thus, two voxels that show the same tuning profile, but different maximum explained variance would have the same tuning curve. A tuning value of 0 for a given spatial frequency means that the associated feature maps explained none of the variance in activation across stimuli; a tuning value of 1 means it uniquely explained all of the variance.

Averaging the tuning distributions of all voxels within a ROI produces a tuning distribution at the level of ROI. Averages (dots in Figures 6A–6C) and error estimates were obtained by sampling with replacement 100 times the 16 validation subsets and associated encoding models for each voxel and then averaging across all iterations and voxels in a single ROI.

Consistent with previous studies [33, 34] ROI-level tuning curves were found to empirically obey a log-Gaussian relationship. We thus performed nonlinear regressions to fit curves of this form to each tuning curve (curves Figures 6A–6C). This fit was used to estimate the peak frequency values of the tuning curves (Figure 6D) and its shift (Figure 6E). The error estimate on the difference in peak frequency between imagery and vision tuning curves takes into account the uncertainty in the fitting procedure as well as the uncertainty in the ROI tuning points.

To obtain significance estimates for Figure 6E we tested the hypothesis of a non-zero difference between imagery and visual peak spatial frequencies against the null hypothesis of no difference between peak frequencies. To construct the null distribution of peak frequency shift, we randomly shuffled the ‘‘vision’’ or ‘‘imagery’’ designation of voxel-wise tuning and calculated the mean frequency shift 1000 times. The region outside the red shaded area in Figure 6E indicates values with $p < 0.01$ for combined subject data. For each subject individually, the points with $p < 0.01$ are displayed with an asterisk.

Figure S4B shows the spatial frequency tuning functions for all subjects and ROIs during vision and imagery (shown together for subject 1 in Figure 6B). Figure S5A, bottom panel, demonstrates the cortical layout of peak spatial frequency during vision and imagery.

Receptive field size and location

As described above, for each voxel we fit $k = 16$ independent visual encoding models, and $k = 16$ independent imagery encoding models, each corresponding to a different training/validation split of the data. Thus, for each voxel we obtain 16 different estimates of receptive field size and location. Results on differences in the location (Figures 7C and 7D) and size (Figure 7B) between imagery and visual receptive fields were obtained by repeatedly sampling these estimates.

The colored lines in Figure 7D show the average (over all k samples per voxel) shift in receptive field location of individual voxels from vision to imagery. Their color represent the inward polar angle of the shift. To construct the plots in Figures 7B and 7C we sampled 1000 pairs of imagery and visual receptive field parameters at random from the $k = 16$ encoding models available for each voxel in order to capture the variance of these samples. The bars in Figures 7B and 7C show the average shift over all sampled pairs and voxels in each ROI per subject, and their error bars indicate one standard deviation of the distribution of shifts. The yellow curve in Figures 7B and 7C shows the same measures for the combined subject data, with the shading representing one standard deviation. Illustrations of receptive fields in Figure 7A show average receptive field locations and sizes over samples for individual voxels.

To obtain the significance estimates displayed for all receptive field size and location results in Figures 7B and 7C we tested the hypothesis of a non-zero mean difference between imagery and visual receptive field parameters against the null hypothesis of no mean difference. To construct the null distribution, we performed the same sampling process as above with the addition of randomly assigning the ‘‘imagery’’ or ‘‘vision’’ designation to each sampled value for each voxel in the indicated ROI. We then calculated the mean difference between the receptive field parameters between each group. This process was repeated 1000 times, resulting in a histogram of differences in mean receptive field parameters for each ROI. The region outside the red shaded area in Figures 7B and 7C indicates values at significance level $p < 0.01$ for combined subject data. For each subject individually, the observed values with $p < 0.01$ are highlighted by an asterisk on the mean observed value.

Seen and mental image identification

An important measure of the validity of an encoding model is how well it can discriminate target stimuli that correspond to the measured activity from other "lure" stimuli [10]. The model-based decoding analysis is known as "image identification" [18]. Here we used pairwise "hits" as our metric for identification accuracy. A "hit" occurs when the measured voxel activity pattern associated with a seen or imagined target stimulus is more correlated with the predicted activity pattern associated with the target stimulus than the predicted activity pattern associated with a lure stimulus. Identification accuracy for a given target stimulus is the percentage of hits accumulated across all lure images.

We performed two distinct types of identification. Position identification was used to determine if the encoding models successfully captured the way that object position was encoded in population activity. Similarly, object identification was used to determine if the encoding models successfully captured the way that specific objects, independent of position, were encoded in population activity. For both types of identification, the cue was not included as part of either the target or lure stimuli. Thus, model predictions did not include any information about the cue (and we again emphasize that cue-responsive voxels were not included in this analysis, see Voxel selection).

Both identification of locations and of objects was performed in two parts. First, half of the activation samples (256) for each voxel were randomly selected (but balanced such that all locations (objects) are represented in each set) to estimate a cross-validated prediction accuracy score (Pearson correlation coefficient) for each voxel. We then ranked these voxels and selected a population of 500 voxels to calculate identification accuracy on the remaining half of the activation samples. To perform location (object) identification, the measured and predicted activity patterns that corresponded to a single location (object) were concatenated across all 64 objects (8 locations) and across the selected voxels. This produced 8 (64) series of concatenated predictions and measured values for location (object) identification. We then evaluated the correlation matrix between all prediction series and measured series. The percentage of identification "hits" is then simply the fraction of entries for which the diagonal elements have a greater value than the other entries in the row corresponding to the "lure" predictions.

To produce the curves shown in Figures 4D and 4E, the previous procedure is repeated 100 times for overlapping brackets of decreasing voxel validation accuracy (i.e., the top bracket contains the 500 voxels with the most accurate encoding model predictions, the second bracket contains the voxels with the 251^{st}–751^{th} most accurate, etc.). Values on the x-axes of plots in Figures 4D and 4E give the largest (over the 100 repeats) of the smallest validation accuracy within each bracket. The standard error captures the variation of the identification "hits" percentage within each bracket.

To estimate the level of identification due to chance, the real identities of the locations (objects) underwent 5 shuffling per each of the 100 repeats discussed above and a common histogram was built from these 500 values for each bracket. Chance level for hits, the center of the null distribution, was always at 50%. The region outside the gray shading near the bottom of plots in Figures 4D and 4E correspond to identification score with significance level $p < 0.01$.