nature communications

Article

https://doi.org/10.1038/s41467-026-68711-x

Neural and computational mechanisms underlying one-shot perceptual learning in humans

Received: 6 September 2024

Accepted: 7 January 2026

Published online: 04 February 2026

Check for updates

Ayaka Hachisuka[1,10], Jonathan D. Shor[1,10], Xujin Chris Liu[2,3,10], Daniel Friedman[4], Patricia Dugan[4], Ignacio Saez[5], Fedor E. Panov[6], Yao Wang[2,7], Werner Doyle[3], Orrin Devinsky[4], Eric K. Oermann[3,8,9,11] ✉ & Biyu J. He[1,4,7,9,11] ✉

The ability to quickly learn and generalize is one of the brain's most impressive feats and recreating it remains a major challenge for modern artificial intelligence research. One of the most mysterious one-shot learning abilities displayed by humans is one-shot perceptual learning, whereby a single viewing experience drastically alters visual perception in a long-lasting manner. Where in the brain one-shot perceptual learning occurs and what mechanisms support it remain enigmatic. Combining psychophysics, 7 T fMRI, and intracranial recordings, we identify the high-level visual cortex as the most likely neural substrate wherein neural plasticity supports one-shot perceptual learning. We further develop a deep neural network model incorporating top-down feedback into a vision transformer, which recapitulates and predicts human behavior. The prior knowledge learnt by this model is highly similar to the neural code in the human high-level visual cortex. These results reveal the neurocomputational mechanisms underlying one-shot perceptual learning in humans.

The human perceptual system is incredibly malleable even in adulthood. Visual perceptual abilities, from low-level contrast and color sensitivity to high-level expertise in recognizing clinical features in radiological images, can improve dramatically with repeated training[1,2]—"practice makes perfect". While perceptual learning is often studied in the context of slow, laborious training, it can also occur with a single experience in a drastic, long-lasting manner (an "aha!" moment), a phenomenon termed "one-shot perceptual learning"[3–5]. This phenomenon is famously illustrated by the Dalmatian Dog picture[6] and studied in the laboratory using the "Mooney image" paradigm, wherein degraded images are difficult to recognize initially, but effortlessly recognized once the subject views the corresponding original, clear images, and the learning effect lasts many months[3,5,7]. Thus, the visual system possesses very fast learning mechanisms without sacrificing stability or suffering from catastrophic interference. To date, the neural mechanisms underlying this rapid perceptual learning ability remain elusive.

Although artificial intelligence (AI) has shown tremendous progress in basic object recognition over the past decade, one- or few-shot learning remains an unmet need and has emerged as an active area of

[1]Department of Neuroscience, NYU Langone Health, New York, NY, USA. [2]Electrical and Computer Engineering, NYU Tandon School of Engineering, New York, NY, USA. [3]Department of Neurosurgery, NYU Langone Health, New York, NY, USA. [4]Department of Neurology, NYU Langone Health, New York, NY, USA. [5]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Department of Biomedical Engineering, NYU Tandon School of Engineering, New York, NY, USA. [8]Center for Data Science, New York University, New York, NY, USA. [9]Department of Radiology, NYU Langone Health, New York, NY, USA. [10]These authors contributed equally: Ayaka Hachisuka, Jonathan D. Shor, Xujin Chris Liu. [11]These authors jointly supervised this work: Eric K. Oermann, Biyu J. He.
✉e-mail: eric.oermann@nyulangone.org; biyu.he@nyulangone.org

research in recent years. These research efforts have focused on tasks belonging to concept learning, such as the classification or detection of a novel object based on a few or no training examples[8–12]. Approaches broadly involve learning representations that can be used to distinguish novel cases[12], learning parameters that can be easily adapted to novel tasks, or learning models of the generating process behind potential novel cases[13]. However, there are several reasons to consider one-shot perceptual learning and one-shot concept learning as fundamentally different phenomena. First, one-shot perceptual learning relies on existing concepts, without the need to form or handle new concepts. Second, existing evidence suggests that they likely rely on different brain structures: the hippocampus and associated medial-temporal lobe structures for concept learning[14,15], and hippocampus-independent cortical mechanisms for one-shot perceptual learning[4]. Third, very young children can learn and generalize new concepts quickly[16–18], while one-shot perceptual learning has a protracted developmental time course, reaching adult-level in adolescence[19,20].

What neural and computational mechanisms support one-shot perceptual learning in humans? Conventional wisdom holds that one-shot, fast learning requires the hippocampus, but a recent study[4] ruled out this possibility for one-shot perceptual learning: memory-impaired patients with bilateral hippocampal lesions were intact at one-shot perceptual learning. This study also demonstrated a clear dissociation between one-shot perceptual learning and episodic memory—both are fast, one-shot learning, but only episodic memory (about whether a picture was previously encountered) is impaired after damage to the hippocampus and associated medial temporal lobe structures.

However, this still leaves a vast hypothesis space for where the learning-related plasticity subserving one-shot perceptual learning might occur in the brain. Previous neuroimaging studies have shown widespread cortical activity changes before vs. after one-shot perceptual learning, from early visual cortex and high-level visual regions to frontoparietal (FPN) and default-mode (DMN) networks[21–25]. In all of these regions, after one-shot learning, neural activity patterns triggered by Mooney images contain more information about the image content and become more similar to the activity patterns triggered by the matching original images (which induced learning). However, not all of these brain regions are necessarily involved in the learning process, and it would be uneconomical for the brain to store multiple copies of prior knowledge (i.e., the knowledge learnt by viewing the corresponding original clear image). A more efficient solution would be to store the learnt prior knowledge in a particular site or a few interconnected sites, and, once reactivated by a matching visual input (degraded image post learning), it could exert widespread influences on neural processing. In this paper, we aim to investigate where priors are stored, their representational format, and potential computational mechanisms.

Because learning-induced plasticity from synaptic changes is not directly measurable by neuroimaging techniques, the site of prior storage (i.e., where learning/plasticity occurs) has remained unresolved. Previous neuroimaging work hypothesized that either FPN or DMN might encode the prior knowledge learnt in one-shot perceptual learning and send this prior information to visual regions[3,26]. However, this hypothesis was based on observations comparing neural activity driven by the same degraded image input before and after viewing the corresponding original clear image, and the activity differences might reflect a region's involvement in perceptual processing, which can be influenced by priors stored elsewhere.

In other lines of work, previous studies using slow, laborious training paradigms to induce perceptual learning have emphasized plasticity within the visual system[1,2]. And a recent study showed that monkey inferotemporal (IT) cortical neurons are equipped with a multiplexed neural code for object perception and long-term memory, such that familiarity (a form of episodic memory) can be read out from the same neuronal population as perception[27]. However, these studies did not specifically address neural plasticity involved in one-shot perceptual learning, which is distinct from episodic memory[4] and likely differs from slow, laborious perceptual learning[28]. In sum, the exact brain mechanisms supporting one-shot perceptual learning, including the site of learning-related plasticity, remain unknown.

To pinpoint the site of cortical plasticity and the involved computational mechanisms underlying one-shot perceptual learning in humans, we used several convergent approaches in this study: First, using psychophysics, we manipulated the prior-inducing image and assessed its effect on learning. This revealed what kind of information is stored in the prior knowledge encoded in the brain, which was then compared with neural coding properties assayed by 7 T fMRI to identify which brain regions have neural coding properties compatible with the information content of prior knowledge. Second, using intracranial recordings in neurosurgical patients, we assessed the timing latencies of neural activity changes in different brain regions; brain regions with the earliest prior-driven activity changes are more likely to be the site of prior knowledge storage. Third, we built a deep neural network (DNN) capable of one-shot perceptual learning, which both captured the overall magnitude of learning effects and predicted image-specific learning outcomes in humans. We then asked which brain region's neural code is similar to the prior information learnt by the DNN. The convergent results from these three lines of inquiry point to the high-level visual cortex (HLVC) as the site of learning-induced plasticity. Our work further reveals potential computational mechanisms involved in one-shot perceptual learning by developing a DNN model capable of capturing human behavior in this task.

Results

Invariance properties of perceptual priors

We first replicated previously observed behavioral effects[22,24,25,29] using a well-established one-shot perceptual learning paradigm (Fig. 1a). Subjects were instructed to verbally identify the content depicted in the Mooney or grayscale image. In "original" trials, the Mooney images and their matching original grayscale images were presented at the same size, retinal location, and orientation (Fig. 1b, top). In "catch" trials, the grayscale image did not match the corresponding Mooney image, which controlled for repetition effects (Fig. 1b, middle). Similar to previous studies, we found robust learning effects in "original" but not "catch" trials. A two-way repeated-measures ANOVA on image recognition rate revealed significant main effects ([Pre vs. Post]: $F_{1,29} = 115.5$, $p < 0.001$; [Original vs. Catch]: $F_{1,29} = 12.9$, $p = 0.001$) and, critically, a significant interaction effect ($F_{1,29} = 39.6$, $p = 7 \times 10^{-7}$; for full statistics, see Supplementary Table 1) (Fig. 1d).

To investigate the information content of prior knowledge acquired during one-shot perceptual learning, we manipulated the matching grayscale image (Fig. 1b, bottom) in multiple ways across two experiments (Fig. 1c). We reasoned that if a particular manipulation did not impair learning as compared to the "original" trials (Fig. 1e, H1), it would suggest that the stored perceptual priors are invariant to this manipulation (i.e., did not encode the specific information altered by this manipulation). By contrast, if a particular manipulation abolished the learning effect (Fig. 1e, H3), it would suggest that the perceptual priors are stored in a specific format that the manipulation disrupted. Finally, if a particular manipulation significantly reduced learning but did not abolish it (Fig. 1e, H2), it would suggest that the perceptual priors are partially invariant to that manipulation. Then, the invariance properties of the perceptual priors will indirectly point to where in the brain they are stored, given known neural coding properties in different brain regions, which we will further validate via an fMRI experiment. We note that our experimental logic is similar to previous psychophysics studies on slow, gradual visual perceptual learning, investigating whether the learning effect is specific to the trained condition or transfers to other conditions as a way to shed light on the potential brain loci of learning and plasticity[1,2,30].

https://doi.org/10.1038/s41467-026-68711-x

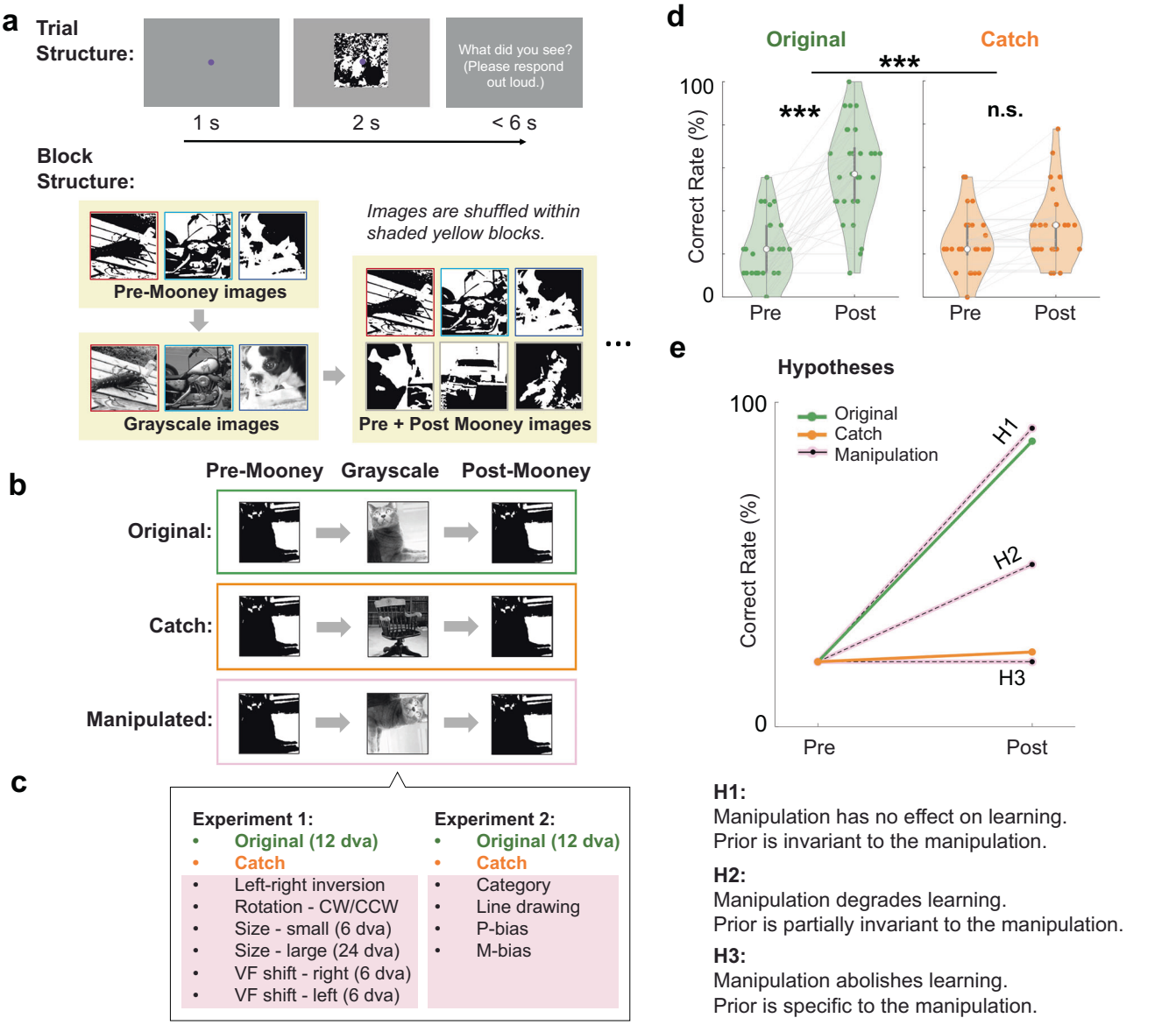

Fig. 1 | Paradigm and hypotheses for the psychophysics experiment. **a** Paradigm. Top: Trial-level timing; images were presented for 2 s, followed by a verbal response. Bottom: Block structure; pre- and post Mooney images were shuffled to prevent low-level priming effects. Border colors reflect paired Mooney-grayscale images and were not shown to subjects. **b** For each subject, a given Mooney image and its paired grayscale image are presented in one of three conditions (original, catch, manipulated). **c** Grayscale image manipulation conditions in each experiment. **d** Image identification accuracy for pre- and post-Mooney images in original and catch trials. Data from Experiment 1 ($n = 30$ subjects), reproduced from Fig. 2a. Asterisks denote statistically significant interaction effects in a two-way repeated measures ANOVA, ***: $p < 0.001$. The central white dot of each violin plot represents the median, the gray vertical bar represents the inter-quartile range (25th to 75th percentiles), the violin plot bounds represent the minima and maxima, and the plot curvature represents the density estimate of the data distribution. Source data are provided as a Source Data file in Fig. 2. **e** Hypotheses. Grayscale image manipulation may have no effect on learning (H1), degrade learning without abolishing it (H2), or abolish learning (H3). All images adapted from the Caltech 101[70] and Pascal VOC[71] databases.

Importantly, to test for one-shot perceptual learning, each Mooney image (presented in both pre and post stages) and its associated grayscale image were presented to a given subject only once under a particular grayscale image condition (original, catch, or a specific manipulation condition; see Fig. 1b). Different images and conditions were presented to different subjects in a counterbalanced design and the results were pooled across unique images and subjects (for details, see "Methods", "Behavioral Experiment 1").

First, to test whether the learnt prior knowledge contains orientation-specific or orientation-invariant information, we left-right inverted the grayscale images or rotated them by 90° (Fig. 2b, c, top). Previous work has shown that orientation-invariant object representations emerge within the primate inferior temporal (IT) cortex[31,32], where posterior IT is orientation-specific and anterior IT is orientation-invariant[32], with a similar trend in the human HLVC[33–35]. We found that both rotation and inversion significantly degraded the learning effect without abolishing it (Fig. 2b, c). A two-way repeated-measures ANOVA comparing each manipulation condition to the "original" trials showed a significant interaction effect ([pre vs. post] × [original vs. manipulated]; inversion: $F_{1,29} = 7.4$, $p = 0.011$; rotation: $F_{1,29} = 11.2$, $p = 0.002$). Similarly, an ANOVA comparing each condition to the "catch" trials also showed a significant interaction effect (inversion: $F_{1,29} = 51.1$, $p = 7 \times 10^{-8}$; rotation: $F_{1,29} = 20.7$, $p = 9 \times 10^{-5}$). Thus, perceptual priors are partially invariant to orientation manipulation.

https://doi.org/10.1038/s41467-026-68711-x

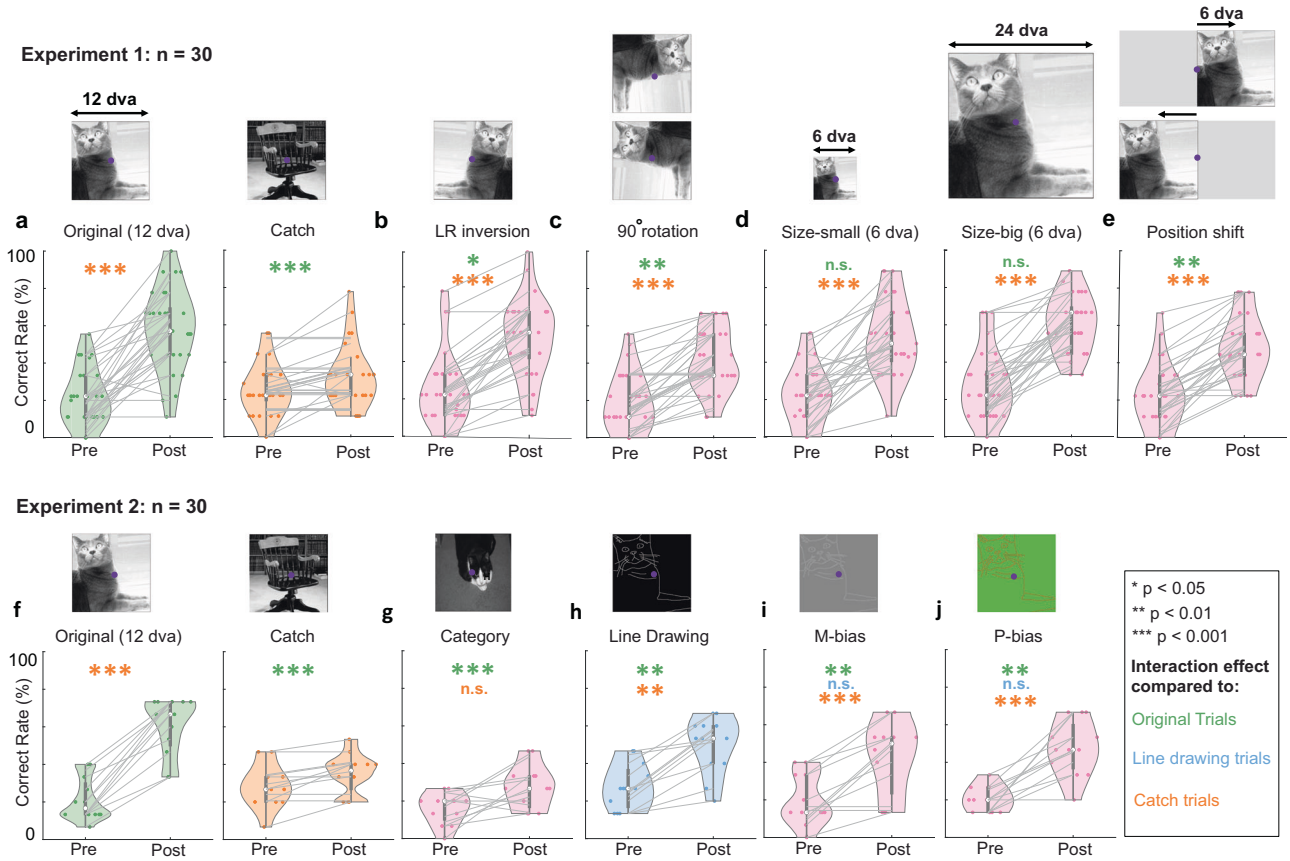

Fig. 2 | Mapping invariance properties of perceptual priors. Top row: Experiment 1 ($n = 30$ subjects). Bottom row: Experiment 2 ($n = 12$ subjects). **a** and **f** Learning effect in response to the original grayscale images and catch images in Experiments 1 and 2, respectively. Data underlying (**a**) are identical to those plotted in Fig. 1d. **b**–**e** Learning effects in response to left-right inverted grayscale images (**b**), 90° rotated grayscale images (**c**), size-manipulated grayscale images (**d**), and left/right visual-field shifted grayscale images (**e**). **g**–**j** Learning effects in response to a different grayscale image from the same category (**g**), high-contrast line drawings (**h**), magnocellular pathway-biasing low-contrast images (**i**), and parvocellular pathway-biasing red-green iso-luminant images (**j**). In (**i**), image contrast is artificially increased for visualization purposes. Asterisks denote statistically significant interaction effects in a two-way repeated measures ANOVA compared to original (green), high-contrast line drawing (blue), or catch (orange) trials. The central white dot of each violin plot represents the median, the gray vertical bar represents the interquartile range (25th to 75th percentiles), the violin plot bounds represent the minima and maxima, and the plot curvature represents the density estimate of the data distribution. Source data are provided as a Source Data file. All images adapted from the Caltech 101[70] and Pascal VOC[71] databases.

Next, we tested for size and positional invariance of the perceptual priors. Given the increasing receptive field size along the visual hierarchy, a given position or size change of the image input may completely alter neuronal encoding in a low-level region while having a modest influence on a higher-level region. Based on previous reports of RF sizes in the ventral visual stream[36,37], we chose the following size and position manipulations. The original images were presented at central fixation with 12 degrees of visual angle (dva). For the size manipulation, we decreased the image size to 6 dva or increased it to 24 dva (Fig. 2d, top). For the position manipulation, we shifted the image 6 dva to the left or 6 dva to the right (Fig. 2e, top). A control analysis investigated these manipulations' impacts on neural coding based on published population receptive field (pRF) data from the human ventral visual stream[37,38]. A voxel's pRF measures the center and size of its receptive field based on measured fMRI BOLD signal, reflecting an average property across neurons sampled within that voxel. This analysis showed that in anterior HLVC, our chosen size and orientation manipulations have relatively small impacts on neural coding (70–100% pRFs retain diagnostic feature), while position shifts had relatively large impacts (20–40% pRFs) (see Supplementary Fig. 1 and Supplementary Result). In early visual regions, all manipulations have larger impacts on neural encoding (Supplementary Fig. 1).

Strikingly, we found that presenting the grayscale image at double or half the original size had no impact on the learning effect, as shown by non-significant interaction effects compared to the "original" trials (reduced size: $F_{1,29} = 1.8$, $p = 0.189$, $BF_{10} = 0.7$; increased size: $F_{1,29} = 0.9$; $p = 0.354$, $BF_{10} = 0.4$) and significant interaction effects compared to the "catch" trials (reduced size: $F_{1,29} = 31.5$, $p = 5e\text{-}6$; increased size: $F_{1,29} = 45.4$; $p = 2 \times 10^{-7}$). Given that size manipulation significantly alters neural coding in early visual cortex (Supplementary Fig. 1c, V1-hV4), these results suggest that the perceptual priors are likely not encoded in the early visual cortex.

We found that position shifts significantly degraded the learning effect, yet without completely abolishing it (Fig. 2e), as evidenced by a significant interaction effect as compared to the original trials ($F_{1,20} = 8.4$, $p = 0.007$) as well as a significant interaction effect compared to the catch trials ($F_{1,20} = 31.2$, $p = 5 \times 10^{-6}$). In a control analysis, we excluded trials where subjects shifted their gaze more than 3 dva away from central fixation. The results were unchanged with both interaction effects remaining significant ($p = 0.031$ and $p = 0.015$; Supplementary Fig. 2).

In sum, orientation manipulations and position shifts significantly impaired learning, yet without abolishing it (following H2, Fig. 1e), while size manipulations had no impact on learning (following H1). These results are inconsistent with early visual cortex being a principal

site for storing perceptual priors and instead point to HLVC as a likely candidate region. In particular, since orientation invariance emerges within HLVC[32], if both posterior and anterior HLVC regions are involved in storing the perceptual priors, it would explain the observed pattern of partial invariance to orientation manipulations.

Perceptual priors are encoded in a perceptual, not conceptual, space

The above experiment used orientation, size, and position manipulations to probe invariance properties of the perceptual priors. These can be compared to known neural coding properties along the ventral visual stream, where invariance to these manipulations gradually increases across successive stages of neural processing. In a second experiment, we broadened our investigation along two additional lines. First, we probed whether the prior is stored in the perceptual space or at an abstract, conceptual level. To this end, we replaced the grayscale image with another image exemplar from the same object category (Fig. 2g). This manipulation completely abolished learning (following H3, Fig. 1e), as evidenced by a significant interaction effect compared to the original trials ($F_{1,11} = 34.4$, $p = 1 \times 10^{-4}$, $BF_{10} = 3311.7$), and a non-significant interaction effect compared to the catch trials ($F_{1,11} = 4.29$, $p = 0.063$, $BF_{10} = 2.2$). This suggests that the perceptual priors are stored in the perceptual space rather than at the conceptual knowledge level, compatible with our hypothesis that it is stored in HLVC, since IT neurons encode category information in a primarily perceptual space with explicit representation of many perceptual features[39,40].

We further probed whether the magnocellular or parvocellular visual pathway could each support the acquisition of perceptual priors. These two pathways originate from different populations of retinal ganglion cells and have stronger contributions to the dorsal and ventral visual pathways, respectively, but this separation is not absolute[41,42]. Following previous studies[43–45], we created line drawings that are either low contrast or red-green iso-luminant, based on the original grayscale images. The low contrast images bias visual processing toward the magnocellular pathway (M-bias); the red-green iso-luminant images bias visual processing toward the parvocellular pathway (P-bias). As a control, we created high-contrast line drawings to substitute for the original grayscale images. The high-contrast line drawings induced a significant learning effect that was lower than the original grayscale images (compared to catch: $F_{1,11} = 17.0$, $p = 0.002$; compared to original: $F_{1,11} = 9.9$, $p = 0.009$), presumably due to the loss of texture and other detailed information. Interestingly, compared to the high-contrast line drawings, neither the M-bias nor the P-bias images caused a significant reduction in the learning effect (M-bias: $F_{1,11} = 0.1$, $p = 0.79$, $BF_{10} = 0.4$; P-bias: $F_{1,11} = 0.01$, $p = 0.91$, $BF_{10} = 0.4$), and both sets of images induced robust learning effects (interaction effect compared to catch, M-bias: $F_{1,11} = 24.6$, $p = 4 \times 10^{-4}$; P-bias: $F_{1,11} = 35.5$, $p = 1 \times 10^{-6}$). These results suggest that either the magno- or the parvo-cellular pathway alone can support one-shot perceptual learning. Although the magnocellular pathway has a stronger contribution to the dorsal visual stream, it has collaterals reaching the IT cortex[41]. Therefore, these findings are compatible with our overall hypothesis that the perceptual priors are stored in HLVC.

Finally, a control analysis excluding any trials in which the prior-inducing image was not correctly identified yielded similar results in all conditions of both experiments (Supplementary Fig. 3).

Neural code in the HLVC matches the invariance properties of perceptual priors. To confirm that HLVC indeed has neural coding properties compatible with the information content of the perceptual priors uncovered in our behavioral experiments, we conducted a 7 T fMRI experiment using a subset of the grayscale images (Fig. 3a) employed in the behavioral experiments. On each trial, subjects ($N = 10$) viewed a grayscale image presented in the original condition or one of the manipulation conditions employed in Experiment 1 for 500 ms, followed by a 1.5–3 s inter-trial interval.

For each subject and region of interest (ROI), we computed a neural representational dissimilarity matrix (RDM) comprising cross-validated (c.v.) Euclidean distances between every pair of image-condition combination computed from voxel-wise fMRI activity patterns. Given 10 unique images and 7 image conditions, this generated a 70 × 70 matrix (Fig. 3b and Supplementary Fig. 5a). ROIs covered early visual cortex (EVC, including V1–V4), HLVC (including LO1, LO2, and FC), as well as FPN and DMN previously shown to be involved in this task[25,26] (Fig. 3c; for ROI details, see Supplementary Fig. 4 and "Methods").

We first tested which ROIs exhibited significant neural invariance to image manipulations. To this end, we averaged within-image, between-condition neural distances (green squares in the RDM shown in Supplementary Fig. 5a; values shown as green bars in Fig. S5b), which were compared against between-image neural distances (sampled from the yellow region of the RDM in Fig. S5a; values shown as yellow ribbon in Fig. S5b; for details, see SI Methods). A significant difference in this test would suggest that the neural representation has significant invariance to image manipulation, since different conditions of the same image are represented more similarly than different images. Significant neural invariance was found in HLVC regions (LO1, LO2, FC) and V4 (all $p < 0.01$, permutation test, FDR-corrected; black asterisks in Fig. S5b; for full statistics see Supplementary Table 2). A whole-brain searchlight analysis yielded convergent results, with a single significant cluster located within the FC ROI ($p < 0.03$, cluster-based permutation test; center-of-mass MNI coordinates: [−38, −54, −14], 66 voxels).

To directly probe neural representations that have similar invariance properties as those identified in our psychophysical experiment for the perceptual priors, we conducted a model-based representational similarity analysis (RSA). We created a model RDM based on the psychophysical results showing that size manipulation had no impact on learning, while orientation and position-shift manipulations significantly degraded the learning effect (Fig. 3b, bottom). Thus, the model RDM contains three levels of neural distance—low (between size manipulation and original), medium (between orientation/position manipulations and original), and high (between different exemplar images). Across all ROIs, model RDM only correlated significantly with neural RDM from HLVC (LO2: $p = 0.03$, FDR-corrected; Fig. 3c). A searchlight analysis across the whole-brain also identified a significant cluster within the HLVC (Fig. 3d, $p = 0.02$, cluster-based permutation test; MNI = [−44, −78, 0], 580 voxels). This result is consistent with previous work showing that invariant object representation emerges within the IT cortex[32].

Together, these results show that neural representations within the HLVC are uniquely endowed with similar invariance properties as those of the perceptual priors identified by our psychophysical experiment, supporting the notion that HLVC is the prime candidate region for storing the priors in one-shot perceptual learning.

Learning-induced neural activity changes onset first in the HLVC

The above results show that HLVC is a plausible region for implementing learning-induced plasticity and storing the priors. To further test this hypothesis, we probed the timing properties of neural activity changes induced by one-shot perceptual learning using intracranial EEG (iEEG) recordings in 19 patients undergoing neurosurgical treatment of epilepsy. We reasoned that perceptual priors are stored in latent synaptic connectivity (since one-shot perceptual learning's effect is long-lasting[4,5]) and, once reactivated by a matching sensory input (e.g., a Mooney image), can trigger widespread shifts in neural activity towards the prior knowledge such as those observed in non-invasive neuroimaging[24–26] Therefore, the brain region with the earliest shift in neural activity toward the relevant prior knowledge is the most likely region for storing the perceptual prior.

https://doi.org/10.1038/s41467-026-68711-x

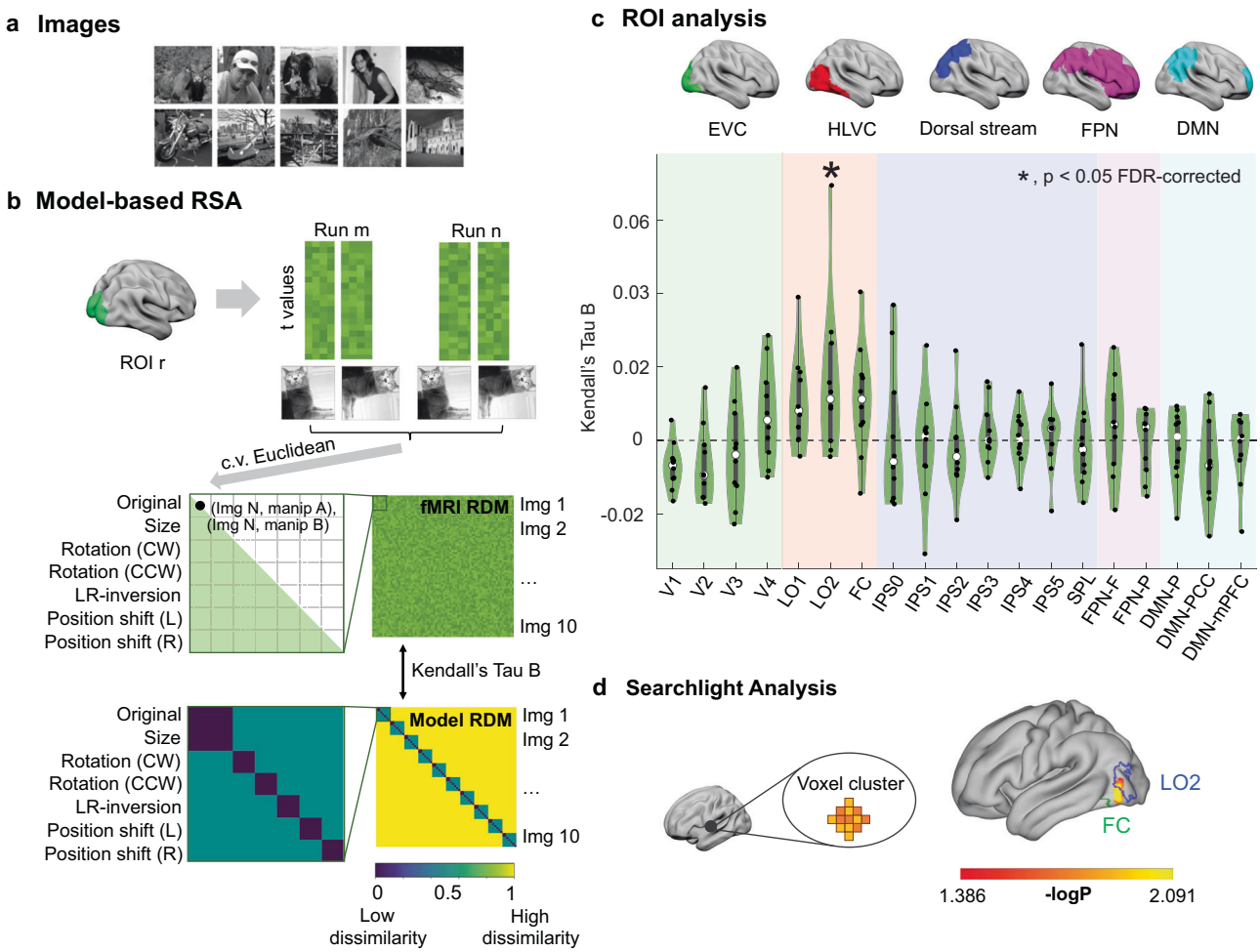

Fig. 3 | Model-based RSA results ($n = 10$ subjects). a Images used in the fMRI experiment, selected from those used in the psychophysics experiment and evenly distributed between inanimate and animate categories. Images were adapted from the Caltech 101[70] and Pascal VOC[71] databases. **b** Voxel-wise fMRI activation patterns are extracted from each ROI, and cross-validated (c.v.) Euclidean distances were calculated for all pairs of image-condition combinations to generate a 70 × 70 RDM. Then, the fMRI RDM is correlated (using Kendall's Tau-B) with the model RDM that corresponds to behavior results, wherein neural distances are assumed to be high (yellow), intermediate (teal), or low (navy). **c** Top: locations of ROIs, grouped by networks; for detailed ROI locations, see Supplementary Fig. 4. Bottom: Kendall's Tau-B values correlating fMRI and model RDMs, for each ROI in the ventral and dorsal streams, as well as FPN and DMN. The central white dot of each violin plot represents the median, the gray vertical bar represents the interquartile range (25th to 75th percentiles), the violin plot bounds represent the minima and maxima, and the plot curvature represents the density estimate of the data distribution. For detailed statistics, see Supplementary Table 2. **d** A searchlight analysis shows significant correlation between model RDM and fMRI RDM in a voxel cluster within HLVC ($p = 0.02$, one-sample t-test across subjects, FWE-corrected). Source data are provided as a Source Data file.

In total, 1886 electrodes were recorded in 19 patients (Fig. 4a; see Supplementary Tables 3 and 4 for demographic, clinical, and electrode information) while patients performed the classic Mooney image task involving "original" and "catch" conditions. Careful screening of patients and collected iEEG data was performed to minimize the potential contribution of pathological activity to the analyzed data (see SI Methods). In all patients, the iEEG electrodes had extensive coverage (Supplementary Fig. 6a) outside the seizure focus (Supplementary Table 4).

Patients exhibited similar learning effects as healthy subjects (Supplementary Fig. 6b, c; [pre vs. post] × [original vs. catch]: $F_{1,18} = 34.2$, $p = 1.5 \times 10^{-5}$). In addition to the five networks used in the fMRI analysis (Fig. 3b), we also assessed the limbic network, which included the cingulate, insular, and orbitofrontal cortices, given recent results showing the limbic network's involvement in conscious visual perception[46–48]. Between 32 and 479 electrodes were recorded in each network. In order to maximize the number of trials collected, image presentation ended when a response was given, with a maximal duration of 2 sec (see "Methods" and Supplementary Fig. 7). For each subject, only images that triggered the classic disambiguation effect—recognized in the post stage and not recognized in the pre stage—entered into the following analysis unless otherwise stated (see "Methods" for details).

We first assessed neural activation time courses, as indexed by high gamma (50–120 Hz) power (HGP)[49,50], for each perceptual stage (Pre, Grayscale, Post). Both early visual cortex (EVC) and HLVC activated early ($p < 0.05$, cluster-based permutation test), at approximately 50 ms after image onset for both pre- and post-Mooney images (Fig. 4c, blue and orange bars, $p < 0.001$). Importantly, post-Mooney images elicited significantly higher neural activity than pre-Mooney images (Fig. 4c, black bars) in the HLVC at approximately 430–623 ms ($p = 0.036$, onset time 95% confidence interval [CI]: 191–551 ms), followed by FPN at approximately 668–1143 ms ($p < 0.001$, onset time 95% CI: 609–943 ms), and later in EVC at approximately 844–1045 ms ($p = 0.01$, onset time 95% CI: 744–893 ms).

Dorsal stream, FPN, and DMN all had a relatively early and transient neural activation for pre-Mooney images (significant clusters found at 92–250 ms, $p = 0.032$; 57–234 ms, $p = 0.028$; and 49–186 ms,

https://doi.org/10.1038/s41467-026-68711-x

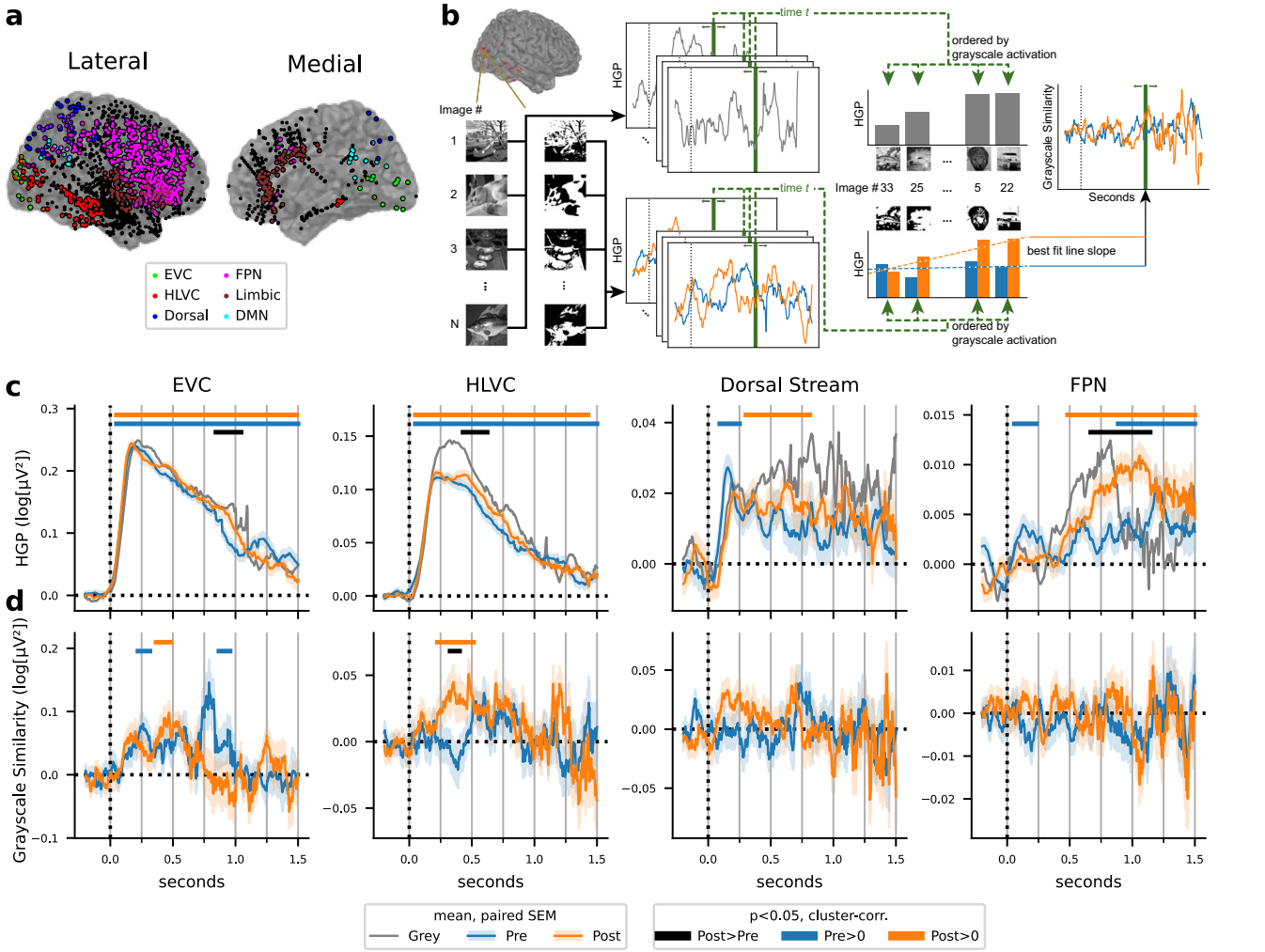

Fig. 4 | Timing properties of learning-induced activity changes. a Locations of recorded electrodes from all $n = 19$ patients, colored by ROI inclusion. Electrodes in more than one ROIs are correspondingly two-toned. Electrodes not assigned to any ROIs are black. Electrodes in the left hemisphere are shown mirrored across the midline. See Supplementary Table 3 for exact electrode counts by ROI. **b** Schematic for Image Preference analysis; for details, see "Methods". Images were adapted from the Caltech 101[70] and Pascal VOC[71] databases. **c** Results for mean ROI activation analysis, showing the HGP time course in each ROI for each condition. **d** Results for Image Preference analysis, showing time courses for the neural tuning similarity of pre- and post-images as compared to Grayscale images. Significance bars: $p < 0.05$, cluster-based permutation test (based on one-sided Wilcoxon signed-rank test), see Supplementary Table 5 for exact p-values. Shaded areas for the Pre and Post time courses denote SEM corresponding to the paired tests (Post > Pre)[78,79]. See Supplementary Fig. 7 for electrodes included across the time course of each ROI. Source data are provided as a Source Data file.

$p = 0.048$, respectively; Fig. 4c and Supplementary Fig. 8a, blue bars). This early and transient neural activation to pre-Mooney images (<250 ms) was likely triggered by bottom-up visual activation that subsided quickly when recognition was unsuccessful. In addition, FPN had higher neural activity to post- than pre-Mooney images at approximately 668–1143 ms, which may be related to recognition-triggered decision-related activity. We did not observe significant neural activation to pre- or post-Mooney images in limbic regions.

To pinpoint neural activity specifically related to prior-guided perceptual processing, we followed an earlier approach[26] to identify time points at which the pre- or post-Mooney image-elicited activity has a similar neural tuning profile as neural activity triggered by the grayscale image (i.e., if an electrode is tuned towards a certain grayscale image, it also exhibits high activity to the matching Mooney image; for analysis schematic, see Fig. 4b). A shift in neural activity toward the relevant prior knowledge would manifest as higher neural similarity between post and grayscale images than between pre and grayscale images. The brain region with the earliest such activity would be the most likely candidate for storing the prior knowledge.

In HLVC, we found that post-Mooney images elicited similar neural tuning profiles as grayscale images at approximately 225–516 ms (Fig. 4d, orange bar, $p < 0.0001$; onset time 95% CI: 152–420 ms), and this similarity is significantly higher than the pre-grayscale similarity (black bar, $p = 0.047$). Importantly, this effect in HLVC preceded that in EVC (at 365–483 ms; Fig. 4d, orange bar, $p = 0.018$; onset time 95% CI: 242–455 ms), and EVC did not exhibit a significant post vs. pre difference. This result, showing earlier and stronger prior-guided neural activity in HLVC, suggests that feedback from HLVC to EVC could have carried prior-related information. EVC also had two time clusters in which pre-Mooney images had similar neural tuning as grayscale images (Fig. 4d, blue bars, $p = 0.047$, $p = 0.036$), which can be explained by similar visual features between Mooney images and their matching grayscale images, such as co-localized contours. We did not observe similar neural tuning between pre/post-Mooney images and grayscale images in any other networks (Fig. 4d and Supplementary Fig. 8b).

To further test the idea that the similar neural tuning in HLVC between disambiguated post-Mooney images and their matching

grayscale images reflects the influence of one-shot perceptual learning, we performed a control analysis using pre-Mooney images that were spontaneously and correctly recognized before seeing the matching grayscale images. Recognizing the Mooney image prior to viewing the greyscale original version occurs in a minority of trials (Supplementary Fig. 6c) and indicates an alternative source of prior knowledge derived from lifelong experiences, distinct from the one-shot priors acquired by viewing the original greyscale images. We found that spontaneously recognized pre-Mooney images did elicit similar neural tuning profiles as greyscale images, but with a distinct temporal profile to that of the disambiguated post-Mooney images described above (Supplementary Fig. 9). Recognized pre-Mooney images exhibited similarity for two short-lived periods at approximately 262–345 ms ($p = 0.032$) and 641–740 ms ($p = 0.024$)—possibly related to a feedforward and a feedback wave[51]; by contrast, disambiguated post-Mooney images exhibited similarity in one cluster at approximately 225–516 ms. The broader temporal cluster with a later peak for disambiguated post-Mooney images (at 355 ms as compared to 301 ms) suggests that additional processing is required in HLVC to bring the recently learned prior knowledge to bear as compared to basic object recognition guided by lifelong knowledge.

Together, these results show that a shift in neural activity towards prior knowledge onsets first in HLVC (at approximately 225 ms), preceding that in EVC. Strikingly, we did not find a similar shift in neural activity in the dorsal visual stream, FPN, DMN, or limbic network. This result, obtained from extensive iEEG sampling across cortical networks, provides strong evidence that the perceptual priors are stored and reactivated in HLVC.

A top-down transformer captures human behavior during the one-shot perceptual learning task

To shed light on potential computational mechanisms underlying one-shot perceptual learning in humans, we sought to develop an image-computable DNN model that can recapitulate human behavior on this task. Instead of modeling specific brain regions, we optimized the model to match human performance, thus avoiding circularity when using the model to localize prior representations in the brain. To preview, we constructed a DNN model which, given a sequence of images, stores accumulated information in a prior module and uses it to modulate visual information processing. We show that our DNN model achieves one-shot perceptual learning capability similar to that of humans, has similar error patterns as human subjects and can be used to predict human learning outcomes for a specific image, thereby proving its efficacy to approximate perceptual priors learnt by human subjects. We further show that the prior information learnt by the model has the highest correspondence to neural representation in the human HLVC.

We converted the Mooney image learning task to a computational benchmark to recreate our experimental setup in silico. Using this benchmark, we developed a top-down transformer architecture engineered to solely rely on top-down signaling for one-shot learning[52] (Fig. 5a, see "Methods" and Supplementary Fig. 10 for details). There are two main components in our model. The first component is a vision backbone (using the transformer architecture), which is pre-trained using self-supervised learning. The second component, key to recapitulating the one-shot learning behavior, is a prior storage module that is responsible for storing prior knowledge about the images seen. Using these two components, we designed two pathways for computing the visual representations suitable for the one-shot perceptual learning task: the bottom-up pathway and the top-down pathway. When an image is first presented to the bottom-up pathway, the vision backbone produces visual representations that are unmodified by previous experiences. The output of the bottom-up pathway is not directly involved in the decision-making process but is used as a query to retrieve relevant representations from the prior storage module. The relevant context from the prior storage module is then used as top-down conditioning to modulate the model in the top-down pathway. Here, the same vision backbone computes image features of the currently shown image again, but this time with the conditioning provided by the prior storage module. Finally, the output of the modulated computation is used to obtain a classification label and update the prior module to incorporate the current information.

We first evaluated the model's performance and perceptual learning effect on 1000 image sequences generated from randomly chosen grayscale images from the ImageNet 1k dataset and their Mooney image counterparts (automatically generated, see "Methods"), following the same task structure as the human psychophysics study (see block structure in Fig. 1a, without any manipulated grayscale images). We define one-shot perceptual learning effect here as the increase in accuracy in the post-Mooney phase compared to the pre-Mooney phase. Our top-down transformer model displayed an average perceptual learning effect of 16.62% (post–pre; Fig. 5b; pre vs. post, Mann–Whitney U test: $p < 0.005$, $N = 1000$). This increase is much higher than the mere repetition-induced learning effect of 3.11% (Fig. 5b; post vs. repetition, Mann–Whitney U test: $p < 0.005$, $N = 1000$), indicating that the model exhibits genuine one-shot perceptual learning.

To further evaluate the model's ability for one-shot perceptual learning against humans, we conducted an online behavioral study ($N = 12$) using a larger set of Mooney images ($n = 219$) (see SI Methods). The 90 images on which human subjects showed the greatest degree of perceptual learning were chosen for the in-person human psychophysics experiments described earlier. We exposed the model to the identical task and image sequences as the human subjects, and plotted model performance against human performance for the top 90 images. Overall, evaluated on the identical task, the model exhibited a similar perceptual learning effect as human subjects (Fig. 5c), with the absolute post-phase human accuracy at 72% compared to the model's at 66%.

We compared our top-down transformer model with existing well-known neurobiologically motivated DNNs (henceforth "baseline models"), including BLT[53] and CORnet[54]. Our model significantly outperformed these baseline models on the one-shot perceptual learning task, as shown by model performance in the evaluation phase (Supplementary Fig. 11a). In addition, when exposed to the image sequences used in the human online psychophysics experiment, CORnet and BLT had sharply degraded performance and failed to maintain the learning effect (Supplementary Fig. 11b). This drop in performance for baseline models was likely due to their inability to maintain the long-term storage of visual priors (the psychophysics task had much longer image sequences than those used in the model training/evaluation phase).

To confirm that the top-down conditioning from the prior storage module is key to our model's success, we corrupted this conditioning signal by using a weighted average of the conditioning tokens and norm-matched Gaussian noise. As the weight of the noise increases from 0.05 to 0.8, the model's performance improvement from the pre to post stage drops sharply (Supplementary Fig. 13), confirming that the conditioning by the prior storage module is key to the model's success at one-shot perceptual learning.

To examine whether our DNN model exhibits any behavioral alignment to humans beyond mimicking the overall accuracy, we analyzed the error patterns of human subjects and our model. The twelve human subjects were each presented with a unique image sequence (consisting of the same set of Mooney and grayscale images). We thus tested our model with the same 12 image sequences presented to human subjects. This resulted in 12 error sequences for humans and the model, respectively. We then asked whether there is similarity between these error patterns, as measured by AUROC (Fig. 5d). We found that the model showed a high but imperfect self-agreement at

https://doi.org/10.1038/s41467-026-68711-x

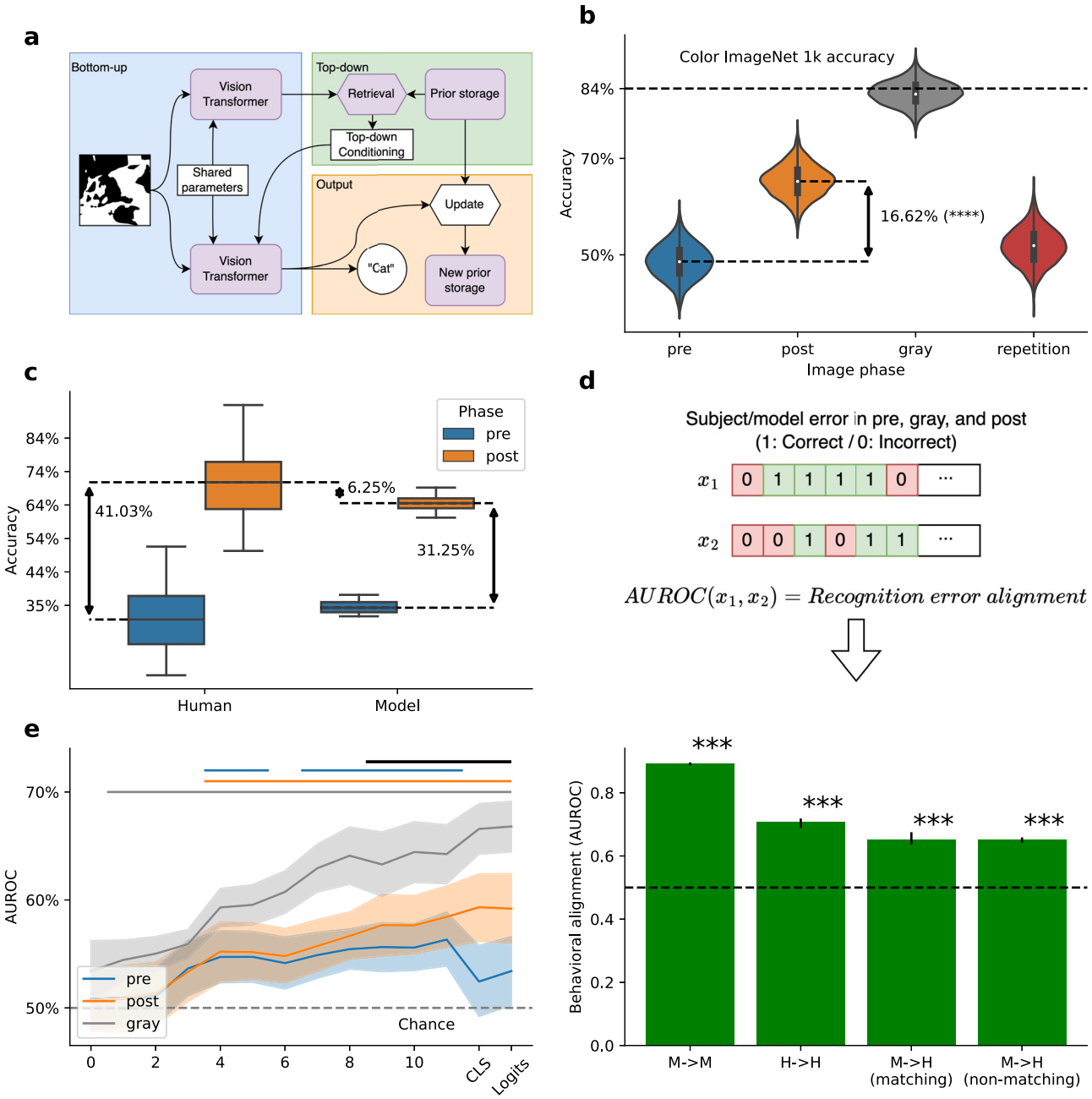

Fig. 5 | Model displays perceptual learning effect and predicts human learning outcomes. a DNN model schematic. The model compares bottom-up features with the "state" module representing prior knowledge, produces top-down conditioning features, and then produces a final output, which is then used to update the model state. For details, see "Methods" and Supplementary Fig. 10. **b** Model accuracy on 1000 test sequences (sequence length: 630, including 210 unique Mooney images) constructed from grayscale ImageNet 1k images and their Mooney image counterparts. The repetition effect was evaluated on the same Mooney images presented twice in a sequence without matching grayscale images (sequence length: 420). The plot shows the distribution of aggregate phase performance across different synthetic image presentation orders ($n = 1000$ image presentation orders). **** indicates $p < 0.00001$, Mann–Whitney test. The central white dot of each violin plot represents the median, the gray vertical bar represents the interquartile range (25th to 75th percentiles), the violin plot bounds represent the minima and maxima, and the plot curvature represents the density estimate of the data distribution. Detailed model performance for "pre", "post", and "gray" conditions is plotted in Supplementary Fig. 11c, broken down by the position of a trial within the long image sequence. **c** Model learning performance as compared to human subjects, where the model was presented with identical image sequences as human subjects. Whiskers show min and max of accuracy, box sides show 25 and 75 percentile, center line shows median accuracy ($n = 219$ images). **d** Image recognition error pattern alignment between human subjects (H- > H), between model fed with difference sequences of the same images (M- > M), and between model and human subjects with matching image presentation (M- > H matching) and non-matching image presentation order (M-H non-matching), measured by AUROC. Bar height shows median value across measurements and error bars indicate 95% CI of AUROC. Dashed line indicates chance level AUROC. *** indicates statistical significance above chance ($p < 0.0005$, one-sided t-test). **e** Human learning outcome prediction. On the x-axis, numbers refer to model layer and CLS, Logits refer to model's representation following its last layer. Black horizontal bar at the top indicates significant difference in prediction when using model features from pre- vs. post-phase ($p < 0.05$, two-sided t test with FWE correction, $n = 12$ subjects). Blue, orange, and gray horizontal bars show significant prediction as compared to the chance level ($p < 0.05$, two-sided t test with FWE correction, $n = 12$ subjects). Individual image learning outcomes are significantly predicted using the model's grayscale image representation starting from the second layer onward. Center line shows mean AUROC across subjects, with shaded areas show 95% CI across subjects ($n = 12$ subjects).

AUROC = 0.89 ($p < 0.0005$ above chance, Mann–Whitney U test, $n = 66$, pairwise between the 12 model instantiations). Human subjects also showed a significant agreement between each other at AUROC = 0.71 ($p < 0.0005$ above chance, Mann–Whitney U test, $n = 66$). Importantly, the model showed significant error pattern similarity to humans at AUROC = 0.65 ($p < 0.0005$ above chance, Mann–Whitney U test, $n = 12$) with the matching image presentation order. When shown non-matching image sequences, the model shows an AUROC = 0.65 ($p < 0.0005$ above chance, Mann–Whitney U test, $n = 132$) that is not significantly different from when the matching sequences are shown to humans and the model ($p = 0.769$, Mann–Whitney U test).

These results suggest that our DNN exhibits behavioral alignment to human subjects and largely recognizes the same images as humans do; moreover, the strength of this behavioral alignment is largely invariant to the order of the presented images. The fact that the specific image sequence has little effect on model-human error alignment (Fig. 5d, rightmost two bars) suggests that the model uses the specific matching grayscale image—instead of all previously seen images or only recently seen images—to disambiguate a given Mooney image, similar to the one-shot perceptual learning phenomenon in humans. Specifically, a mechanism that uses all previously seen images would be similar to life-long priors[55] instead of priors obtained from one-shot learning, and a mechanism that uses only the most recently viewed images would be similar to working memory in humans, which is known to be distinct from one-shot perceptual learning[29].

To evaluate whether the internal representations of the model contain information relevant to how human subjects recognize the Mooney images, we used model internal features (from the vision transformer and its outputs after top-down conditioning; for details see "Methods") to predict human subjects' learning outcomes for individual images (learned vs. not learned). Accurate prediction of human learning outcomes would suggest that the model extracts features that are relevant to humans' learning success. Using the model's representation features for grayscale images, prediction accuracy for human subjects' Mooney image learning outcomes was significant from the 2nd layer onward (Fig. 5e, gray; the 1st layer has index 0), and increased monotonically from early to late layers with a peak AUROC of 66%. The visual features extracted from pre- or post-Mooney images are also significantly predictive of human learning outcome in certain layers, but not as predictive, with post features reaching 59% and pre features reaching 56%. In addition, from layer 8 onwards, model features from the post-phase predict human learning outcome significantly better than model features from the pre-phase (Fig. 5e, black bar), suggesting that the incorporation of pertinent prior information improves the prediction of human learning outcome. Overall, these results show that the features extracted by the model from the prior-inducing grayscale image are highly predictive of humans' learning success rate.

Finally, we tested whether the model shows similar invariance properties as human subjects (Fig. 2a–e). To this end, we fed shuffled blocks of image to the model (see Fig. 1a), with random manipulation applied to the grayscale image, and recorded the model's recognition performance. The model's performance in the grayscale-manipulated condition was then compared to the original condition or the catch condition, similar to the human experiment. The results are shown in Supplementary Fig. 12. Similar to human subjects, the model exhibits invariance to orientation, size, and position manipulations of the grayscale image, as evidenced by a significant interaction effect when comparing each manipulation condition to the catch condition (all $p < 0.001$). Because our model was never designed to capture the invariance properties directly, the emergence of invariance in the model's one-shot perceptual learning ability is nontrivial and adds to the evidence that our model captures the human one-shot perceptual learning phenomenon behaviorally.

The model suggests that prior-related information is concentrated in HLVC

Armed with a DNN model that recapitulates one-shot perceptual learning ability of humans, has human-aligned error patterns, and predicts human learning success at an image-to-image level, we next used the model to shed light on the computational mechanisms implemented in the human brain. Because the model contains an explicit representation of the prior information, we asked which brain region contains neural code similar to the prior information learnt by the model. To this end, we compared the prior information accumulated in our DNN model, which guides the model's learning behavior, with human brain activity recorded during the Mooney image task performance measured by 7 T fMRI ($n = 19$; data from ref. 24), and assessed the ability of prior information encoded in the model to predict voxel-level neural activity in each brain region.

Given the same sequence of images presented to the human subjects, we predicted each subject's neural activity using the model's internal features representing accumulated visual information (state component; see Supplementary Fig. 10 for details), and compared this to a set of baseline predictions. These baselines were obtained from counterfactual catch trials—image sequences that mimic the task format but offer no stimuli for encoded priors, similar to "catch" trials in the psychophysics experiment (Fig. 6a, left; see "Methods" for details). Since the model's state component encodes information related to the visual prior information, the improvement in brain prediction score (see "Methods" for details) as compared to the catch image sequence (shown in orange in Fig. 6a, pooled across all images) indicates the utilization of information related to perceptual priors. We found that the fusiform cortex (FC), a region that is part of HLVC, contained the highest proportion of voxels containing prior-related information (29.7%), followed by DMN (13.9%) and FPN (11.2%) (Fig. 6a, right). Outside of FC, we observed a steadily increasing trend from early visual regions (<5%) to higher-level regions like DMN.

Fusiform cortex information strength assessed by the model predicts the learning effect in humans

Lastly, we evaluated whether successful perceptual learning in humans is related to the strength of learning-related information as measured by the model in each ROI. To measure this information strength, for each image, we quantified the proportion of decrease in brain activity prediction error for a typical image sequence compared to the catch image sequence (which induced no learning). We then pooled these results across images at the ROI level. Following earlier work[24], we defined successful learning in human subjects as 4 or more (out of 6) presentations reported as recognized in the post phase of a Mooney image. We found that the learning-related information strength in the dorsal visual stream is negatively related to the subject's successful perceptual learning (Fig. 6b), with an increase in dorsal stream information strength from the 50th percentile to 100th percentile reducing the average subject perceptual learning success rate from 81% to 61%. This suggests that prior-related information in the dorsal visual stream is inversely related to learning success, a surprising result that hints at a potential competition between the dorsal and ventral visual stream.

We also evaluated whether the reliability of successful perceptual learning in humans is related to the learning-related information strength in an ROI. Taking only the successfully learnt images as defined above, we measured reliability as the proportion of post-phase images that are reported as recognized (varying from 4/6 to 6/6), with 100% being always recognized in the post-phase (hence, most reliable). We found that the fusiform cortex (FC)'s information strength was positively associated with the reliability of the perceptual learning effect (Fig. 6c), with an increase in FC information strength from the 50th percentile to 100th percentile increasing the average subject perceptual learning reliability from 84% to 95%. No other ROI's information strength was associated with success rate or learning reliability,

https://doi.org/10.1038/s41467-026-68711-x

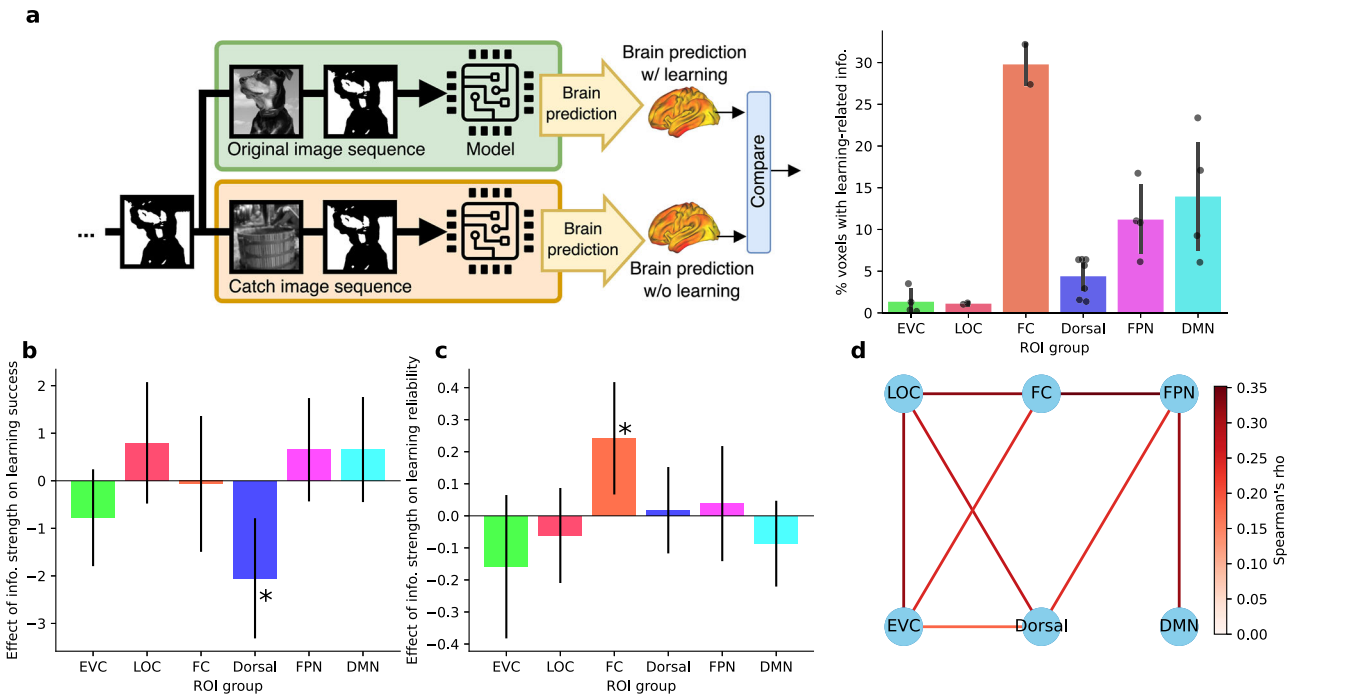

Fig. 6 | Brain prediction contrast reveals FC's strong involvement in learning. **a** Average percentage of voxels in each ROI that show significant improvement over baseline in prediction score at the group level (Pearson's *r*, TFCE 10 k permutation, $p < 0.05$, $n = 19$ subjects for each individual point). Baseline brain prediction is obtained by feeding alternative sequences of images into the model, where no learning happens. FC shows the highest percentage of significant voxels. Each point shows the individual ROI within group. The center line shows the average of that ROI group, error bars show the 95% CI across individual ROIs. **b** Higher dorsal stream median information strength is associated with lower learning success. The center line shows the estimated parameter value, error bars show 95% CI of parameter estimate. *: significant parameter (logit link Binomial family GEE parameter *t*-test, $p < 0.05$, FDR-corrected). **c** Higher FC median information strength is associated with higher learning reliability. The center line shows the estimated parameter value, error bars show 95% CI of the parameter estimate. *: significant parameter (log link gamma family GEE, parameter *t*-test, $p < 0.05$, FDR-corrected). **d** Information strength connectivity pattern associated with successful perceptual learning effect (Spearman's rho parameter estimate using GEE, *t*-test $p < 0.05$, FDR-corrected).

suggesting that the perceptual learning effect is specifically related to information present in FC.

We also evaluated whether the perceptual learning effect is associated with interactions across ROIs by investigating pairwise connectivity between ROIs, where connections are defined by the correlation of learning-induced information strength across different images. We found that when an image is successfully learned, there is significant information connectivity across the entire brain network, with the fusiform cortex being a central node. Specifically, FC is connected with both EVC and FPN, with FPN further connected to DMN. LOC occupies a more peripheral location in the network graph, being connected only to other visual regions (Fig. 6d, $p < 0.05$, FDR-corrected). An alternative but weaker path exists from EVC to FPN through the dorsal stream. By contrast, when the image is not learned, we only observed significant connectivity between the dorsal stream and FPN, and no other connections were significant (not shown).

Together, these DNN-informed results demonstrate a central role of the fusiform cortex in representing prior visual information and predicting human subjects' learning success for a specific image.

Discussion

"Aha" moments, flashes of insight, and other phenomena of one-shot perceptual learning are mysterious and impressive feats of the human brain. Despite decades of research, the site of plasticity and learning underpinning one-shot perceptual learning—fast, long-lasting learning effects in the perceptual domain—remained unknown, in large part due to the learnt prior knowledge being encoded in latent synaptic connectivity (so as to be robust and long-lasting) and difficult to measure using neuroimaging approaches that only capture active neural dynamics.

Here, using convergent approaches from psychophysics, neuroimaging, intracranial recordings, and deep learning, we pinpointed the human HLVC as the seat of neural plasticity subserving one-shot perceptual learning and revealed the potential involved computational mechanisms. The information content of perceptual priors, assayed by psychophysics, uniquely matched the neural coding properties of HLVC, measured by fMRI. Using iEEG, we found that HLVC was the brain region showing the earliest-onset neural signature of prior-guided stimulus processing, suggesting that the latent priors may be encoded and reactivated locally within HLVC. Finally, a vision transformer-based DNN incorporating top-down feedback that shapes visual processing with accumulated prior information was able to recapitulate the one-shot perceptual learning phenomenon in humans and predict the image-to-image human recognition outcome, and the accumulated prior information in the model had the highest correspondence to neural representations in the human HLVC. These multiple strands of converging evidence point to a crucial role of HLVC in one-shot perceptual learning.

Our fMRI experiment was carried out under passive viewing of grayscale images to investigate which brain region has neural coding properties compatible with the invariance properties of perceptual priors uncovered by the behavioral experiment. The logic here is that viewing the grayscale image leaves a "trace" in the activated neural populations, and if the corresponding Mooney image is presented a while later, disambiguation of the post-Mooney image happens, which thereafter becomes a long-lasting memory through consolidation processes. The exact cellular mechanisms involved remain unclear and

await future study (we conjecture that the activity trace might be similar to activity-silent working memory[56]). Nonetheless, the invariance properties of neural activation during passive viewing of grayscale images should be equivalent to the invariance properties of priors stored in the one-shot perceptual learning task, because the latter is inherited from viewing of grayscale images during the one-shot perceptual learning task. We note that the same logic was adopted in a long line of research on visual perceptual learning (VPL)—slow, gradual perceptual learning in the visual domain[2].

Conventional wisdom holds that one-shot learning requires the hippocampus, but a recent lesion study[4] ruled out this possibility for one-shot perceptual learning and instead placed it under the phenomenon of priming[57]. Our observation that one-shot perceptual learning is invariant to size manipulation is reminiscent of earlier studies showing size-invariance in both priming[58,59] and VPL involving object recognition[30]. Up until now, the relationship between priming, VPL, and one-shot perceptual learning at a mechanistic level has been unclear, with studies on priming focusing on changes in neural activity magnitudes before and after exposure[60], and studies on VPL focusing on delineating plasticity at different levels of the visual hierarchy[2,28,61]. Our results are compatible with the view that priming and perceptual learning lie on a continuum[28,62], with one-shot perceptual learning being a special case of priming that has especially long-lasting effects, and a special case of perceptual learning with an especially fast acquisition phase. Interestingly, while three-year old children have similar magnitudes of priming effects as college students, one-shot perceptual learning ability does not reach adult level until adolescence[19,20]. This raises the intriguing possibility that one-shot perceptual learning relies on a perceptual system already fine-tuned by experience.

Previous neuroimaging studies found widespread changes in stimulus-driven neural activity, including a shift in neural activity toward prior knowledge, following one-shot perceptual learning[3,22,24–26], but could not pinpoint where learning takes place in the brain. Using intracranial recordings sampling widespread cortical networks, we observed that neural activity changes driven by prior knowledge—manifesting as a shift in the neural activity toward the relevant prior knowledge—emerged first in HLVC, prior to similar changes in EVC (Fig. 4d), suggesting that top-down feedback from HLVC to EVC could have carried prior-related information. An early primate study using a similar task reported fast changes in IT neuronal firing rates after learning, but did not reveal the time course of these neural activity changes or assess other cortical regions. Interestingly, we did not see a similar neural activity shift towards the relevant prior in higher order brain regions, including FPN and DMN, where such shifts were previously observed in fMRI[24,26]. This is likely due to differences in the recording modalities—high-gamma power is well known to reflect local population neuronal firing rates, whereas fMRI signal can also reflect synaptic inputs and field potential changes uncorrelated to firing rates[63]. Importantly, combining evidence from psychophysics, iEEG, and modeling, the present study underscores the key role of HLVC in one-shot perceptual learning, and updates a previous proposal based on fMRI, suggesting that the prior knowledge learnt from one-shot perceptual learning is encoded in FPN and DMN.

As part of this investigation, we derived a transformer architecture to model the one-shot perceptual learning phenomenon based on top-down mechanisms that convey learnt prior information. We showed that learnt prior information in the model is similar to that contained in the human HLVC, and that the existence of this type of information in the human HLVC predicts more reliable learning in humans. Our network analysis offers a hypothetical mechanism as to how priors shape the perceptual processing. We hypothesize that higher order regions, such as FPN might serve as a controlling center for the usage of prior information, which is stored and activated in HLVC and then communicated to other visual areas, such as EVC, through top-down feedback. The dorsal visual stream, on the other hand, is associated with a lower learning success rate when its information strength is high, suggesting a potential competitive role with the ventral visual stream, consistent with our overall conclusion that the HLVC is critical to one-shot perceptual learning.

Interestingly, while we were developing our model, several similar architectures were described in the machine learning literature that bear a strong resemblance to our model conceptually but were motivated by purely computational considerations with regard to extending the sequence length of transformer-based models (RMT, TransformerFAM, and infini-attention[64]). We see this as a broadly encouraging development in line with other work[65–67] suggesting a convergence between computational neuroscience research and deep learning.

This work is not without its limitations. First, although our DNN model can predict image-to-image human recognition outcomes, its behavioral alignment with human subjects is still below the alignment between two different human subjects (Fig. 6d), potentially due to the absence of additional mechanisms (such as the separation between dorsal and ventral pathways) that we do not account for. A more accurate understanding of the one-shot perceptual learning phenomenon can inform the development of better computational models that can explain individual human brain activity patterns and learning effects. In addition, in our modeling efforts, we focused on the storage and retrieval of content-specific priors purely based on activation changes, rather than model weight changes. For improved modeling of the long-term retention of learned perceptual priors, model weight updates might be necessary.

Second, the circuit- and cellular-level mechanisms supporting learning-related plasticity in HLVC remain to be uncovered. HLVC is known to support slow, gradual visual perceptual learning (VPL) that occurs at the object level[1,30]. A recent study also revealed that IT neurons encode familiarity—a form of long-term episodic memory[27]. An open question for future investigation is whether the neural code in IT cortex for slow VPL, one-shot perceptual learning, and long-term episodic memory rely on the same or overlapping group of neurons and, if so, whether the neural subspaces representing these distinct types of memories are orthogonal or correlated. In addition, one-shot perceptual learning effects persist for months to years, and, therefore, consolidation of the learnt prior knowledge is likely required, and its detailed mechanisms remain to be investigated.

Finally, although we employed a wide range of grayscale image manipulations to delineate the information content encoded in the priors, additional manipulations are possible and can be investigated in future studies. A related question is whether one-shot perceptual learning of low-level visual features, which—although rare—exists in special case scenarios[62], might rely on other visual regions such as EVC.

Human perceptual learning is a critical type of learning in humans, allowing us to modify how we perceive the world without radically shifting the underlying concepts used to perceive it. One-shot perceptual learning is the crown jewel of this general ability. Our work, localizing the underlying learning process to the HLVC and capturing the learning phenomenon in a DNN with top-down feedback, sheds light on this impressive human feat both biologically and computationally. We anticipate that our work will inspire further research into these novel mechanisms of one-shot learning and support the development of AI models with human-like perceptual mechanisms and computational properties. Furthermore, since altered one-shot perceptual learning reflecting an over-reliance of perception on prior knowledge is observed in multiple neuropsychiatric illnesses involving hallucinations[68,69], our findings help to pave the knowledge foundation to better understand the pathophysiological processes contributing to these perceptual disorders.

Table 1 | Counterbalanced group design for the main experiment (Experiment 1)

Images 1–9	Images 10–18	Images 19–27	Images 28–36	Images 37–45	Images 46–54	Images 55–63	Images 64–72	Images 73–81	Images 82–90	
M1	M2	M3	M4	M5	M6	M7	M8	Original	Catch	Group 1
Catch	M1	M2	M3	M4	M5	M6	M7	M8	Original	Group 2
Original	Catch	M1	M2	M3	M4	M5	M6	M7	M8	Group 3
M8	Original	Catch	M1	M2	M3	M4	M5	M6	M7	Group 4
M7	M8	Original	Catch	M1	M2	M3	M4	M5	M6	Group 5
M6	M7	M8	Original	Catch	M1	M2	M3	M4	M5	Group 6
M5	M6	M7	M8	Original	Catch	M1	M2	M3	M4	Group 7
M4	M5	M6	M7	M8	Original	Catch	M1	M2	M3	Group 8
M3	M4	M5	M6	M7	M8	Original	Catch	M1	M2	Group 9
M2	M3	M4	M5	M6	M7	M8	Original	Catch	M1	Group 10

M1–M8 denotes the image manipulation conditions: size-small (6 dva), size-large (24 dva), visual field shift (6 dva left/right shift), left-right inversion, 90° rotation (CW/CCW), as well as an additional condition not relevant to the present study. Subsequently, all images were randomly shuffled across all manipulation groups, to prevent trial blocks of the same manipulation. See above text for additional details.

Methods

Behavioral experiment 1

Subjects. Thirty-three participants were recruited from the greater New York City area. Ages 20–70 (median age 28, std = 13.7), 18 were female. Sex/gender was based on self-report and not considered in the study design, since sex or gender-based differences in perception were not a focus of this study. Most of the participants (31 out of 33) were right-handed, and their vision was normal or corrected-to-normal. All participants were compensated $15/h for their time, and provided with a written informed consent, and the experiment was approved by the Institutional Review Board of New York University School of Medicine (protocol #S15-01323).

Complete data from 3 participants were excluded due to poor performance in the main task. 1 block was removed from 2 subjects due to an experimental script error and a request to leave early, respectively. Exclusion criteria was established prior to the beginning of the study. Data from a total of 30 participants were used in the final analysis.

Experimental stimuli. The task was created using PsychoPy 2020.1.3 and presented on a 1920 × 1080 monitor, placed 63 cm away from the participant's eyes. Participants placed their heads on a chin rest to minimize head movements and ensure a consistent viewing angle. In the original trials, the Mooney and grayscale images had the same retinal location, size, and orientation (12 dva in size, presented at central fixation). All images were taken from public databases: from Caltech 101[70] (https://data.caltech.edu/records/mzrjq-6wc02) and Pascal VOC[71] grayscale image (https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/index.html) databases. All images used in this study were selected for having a single object in a naturalistic background.

Experimental procedure. Participants first completed two blocks of practice trials, first without a time limit for task familiarity, and later with time limits used in the main task. In each trial, a purple fixation dot was presented for 1 s, followed by an image presentation (Mooney or grayscale) for 2 s. Subjects responded with a Yes/No recognition button press, followed by a verbal response (with an upper limit of 6 s), which was recorded in real-time. Participants completed the task inside a dimly lit, soundproof room designed for EEG studies; a microphone was fed through the cable mount so that verbal responses could be heard from outside the room. Eye-tracking data were recorded using EyeLink 1000, in the binocular mode with a sampling rate of 1000 Hz.

The main task consisted of 3 blocks of 90 trials each. In total, 90 unique Mooney images were assessed. The following grayscale image conditions were tested: original, catch, size-small (6 dva), size-large (24 dva), visual field shift (6 dva left/right shift), left-right inversion, 90° rotation (CW/CCW). Each unique Mooney image was assigned to a single condition for each subject (because each image can only be tested once/participant), and this assignment, as well as the presentation order of conditions were counterbalanced across participants. The counter-balance structure is shown in the table below (Table 1). M1–M8 denotes the 7 image manipulation conditions listed above, plus an additional condition not relevant to the present study. Thirty participants were included in this experiment, and were evenly distributed across the 10 groups (i.e., each group contained 3 subjects). All images were randomly shuffled across all manipulation groups to prevent trial blocks of the same manipulation. Results were pooled and averaged across all subjects to test learning outcomes for each condition. This counter-balancing design was necessary because each participant could only view each unique image in one condition, given the one-shot perceptual learning task.

Experiment 1: counterbalanced groups

Following previous studies[22,24], images were presented in mini-blocks that consisted of 3 grayscale images followed by 6 Mooney images. The 6 Mooney images included 3 "Post" images that correspond to the 3 grayscale images shown just before, and 3 "Pre" images that correspond to grayscale images that would be shown in the subsequent block, and their order was randomly shuffled. Before the first block, 3 "Pre" images were shown before the start of this block structure. For the last block, only 3 grayscale images and 3 "Post" images were shown.

fMRI experiment

Subjects. Twelve participants were recruited from the greater New York City area. Ages 21–42 (median age 23, std = 6.2), 9 were female, and all were right-handed with correct or corrected-to-normal vision. Sex/gender was based on self-report and not considered in the study design, since sex or gender-based differences in perception were not a focus of this study. All participants were provided with a written informed consent, and the experiment was approved by the Institutional Review Board of New York University School of Medicine (protocol #S15-01323). Data from 2 participants were entirely excluded: one immediately opted out due to nausea in the scanner, and the other was excluded due to suboptimal scan quality. Lastly, 4 out of 16 blocks from 1 participant were excluded due to scanner error.

Experimental stimuli. The task was created using PsychoPy 2021.2.3 and stimuli were presented using an MRI-compatible LCD monitor (BOLDScreen, Cambridge Research Systems) with a 120 Hz refresh rate. The monitor was located 198 cm behind the center of the scanner bore, and participants viewed the screen using an eye mirror that was placed 5 cm away from the participant's eyes, attached to the head coil. To test for the emergence of invariant object recognition, a subset of 10 grayscale images from the psychophysics study was used. The images were balanced between 5 animate and 5 inanimate objects. The images were shown in the following conditions that matched the psychophysics paradigm: Original image (11 dva), LR inversions, Rotation (CW), Rotation (CCW), Size-small (5.5 dva), VF shift (5.5 dva right), VF shift (5.5 dva left), and line drawings. Line drawings were not further analyzed to constrain analysis to size, viewpoint, and position invariance from Experiment 1. Size-big (24 dva) manipulations were excluded due to monitor size and placement limitations inside the scanner room. Also, the size and VF shift parameters had to be presented at a slightly smaller scale as compared to the behavioral experiments (from 12 to 11 dva, and from 6 to 5.5 dva) to accommodate the scanner screen size.

Task design. First, participants were shown all possible 80 images (10 exemplars × 8 conditions) before entering the scanner (on a gray background, in a similar format as the main task), for familiarity and to prevent any discrepancies in neural responses during the first and the subsequent runs. Each session consisted of anatomical scans and 16 runs of fMRI BOLD runs that were 5 min each, for a total of ~90 min in the scanner. During the task inside the scanner, participants were asked to passively view the screen while maintaining visual fixation, and to respond with a button press when the fixation cross changed from white to red for a 200 ms duration. In each trial, the image was presented for 500 ms, followed by a 1.5–3.5 s jittered ITI. Each run included 80 trials, in which each unique image was presented once in shuffled order, with the constraint that two different manipulation conditions of the same image cannot be presented in adjacent trials (to avoid any repetition suppression/priming effects). The fixation cross color change happened 16 times per run, at a random time after the trial onset (during image presentation or the ITI). At the end of each run, subjects were given visual feedback about the proportion of successful button-presses in response to fixation cross color changes, to maintain task engagement.

MRI data acquisition. Experiments were run in a Siemens 7 T MRI scanner using a 32-channel NOVA head coil at the NYU Center for Biomedical Imaging. T1 weighted MPRAGE images were acquired with 1.0 mm isotropic voxels, FOV 256 mm, 192 sagittal slices, TR 3000 ms, TE 4.49 ms, flip angle 6°, fat suppression on, bandwidth 130 Hz/Px. Proton density images were acquired for intensity normalization, with the following parameters: FOV 256 mm, 192 sagittal slices, 1.0 mm isotropic voxels, TR 1760 ms, TE 2.61 ms, flip angle 6°, bandwidth 280 Hz/Px. BOLD fMRI images were acquired using a GRE-EPI sequence with the following parameters: FOV 192 mm, 66 oblique slices covering all of cortex, voxel size 1.6 × 1.6 mm, slice thickness 1.6 mm with distance factor 10%, TR 1500 ms, TE 25 ms, multiband factor 2, GRAPPA acceleration 2, phase encoding direction posterior to anterior, flip angle 50°, bandwidth 1894 Hz/Px.

fMRI analysis

Preprocessing. Data preprocessing follows our published procedures[24,46]. All fMRI analyses were preprocessed using FSL's FEAT tool. Motion artifacts were corrected using MCFLIRT, which aligned each volume to the volume acquired in the middle of the run, and estimated 3 dimensions of head rotation and translation across time, with 6 DOF. Slice-timing correction accounted for the long whole-brain acquisition time of 1500 ms, which interpolated the signals from each slice to the middle of each TR. Then, the brain was extracted using BET, and spatial smoothing (3 mm FWHM) was applied. Lastly, ICA cleaning was used to remove artifacts related to the motion, arteries, or CSF pulsation. The data was initially passed through AROMA ICA, an automatic artifact classification method, and 60–70 components that explain ~80% of variance in the BOLD signal were manually inspected to select components that corresponded to artifacts. Functional images were registered to the individual subject's MPRAGE (T1).

General linear model (GLM). A general linear model (GLM) was used to extract stimulus-evoked activation, using the FEAT tool in FSL. For each task run, the following regressors were created: one regressor for each of the 80 unique images, as well as the button press events, for a total of 81 regressors per run. For the button press regressor, a boxcar function was applied for the duration between the onset of fixation cross color change and the button press; in the event of missed trials, the boxcar lasted 200 ms—the duration of color change. Then, beta estimates for each regressor were obtained. *t*-values were computed by dividing the beta estimate by its standard-error estimate (output from FSL), and were used for the rest of the analysis to suppress the contribution of noisy voxels in the beta estimate[72]. All analyses were conducted within each subject, with t-values aligned to the subject (T1) space.

fMRI–Definition of ROIs

For the ventral and dorsal visual streams (V1-V4, LO1-LO2, IPS0-IPS5, and SPL), ROIs were defined using anatomical masks from a probabilistic atlas that used retinotopy to map ROIs[73]. All overlapping voxels were removed using a winner-take-all approach. Fusiform cortex (FC) was obtained from the "temporal occipital fusiform cortex" partition defined by the Harvard-Oxford atlas. Only voxels that had $a > 3\%$ probability of belonging in the ROI were included. FPN and DMN ROIs were derived from task-driven activity patterns (GLM and decoding results contrasting pre- and post-Mooney images, respectively) from an independent dataset reported in a previous paper[24]. Specifically, FPN was defined using a binarized statistical map from a whole-brain searchlight decoding analysis of unrecognized pre-Mooney vs. recognized post-Mooney images. DMN was defined using a GLM contrast of the learning effect (unrecognized pre-Mooney vs. recognized post-Mooney) that has been observed in multiple papers[3,24]. Previous control analyses have shown that the results obtained using these ROIs were similar to those obtained using FPN and DMN ROIs from a resting-state atlas[24,26]. For all ROIs, both hemispheres were combined for analysis. The analyses were performed in each subject's T1 space, with the ROI transformed to this space.

fMRI–Neural distance analysis

Invariant object representation was quantified by the measure of neural distances within the same unique image, across manipulations. *t*-values from the GLM, aligned to the subject's T1 space, were used in this analysis[72], and neural distance metrics were calculated using the rsatoolbox in Python[74]. To increase the signal-to-noise ratio, randomly chosen pairs of runs were binned into a single run, with t-values averaged across them, creating 8 total "runs" for cross-validation. To ensure robustness of the distance metric in RSA, we used cross-validated (c.v.) Euclidean distances as the unbiased distance estimator[74–76].

From each fMRI run, we construct one Q × P matrix for each ROI (or voxel cluster, in the searchlight analysis), where each of the Q rows is the activity pattern in response to a specific image input, and P corresponds to the number of voxels. Here, $Q = 80$, corresponding to 10 image exemplars × 8 conditions. Then, we used a leave-one-run-out cross-validation scheme to compute neural distances between *k*

images across different folds, as follows:

$$d_k = \frac{1}{M}\sum_{m=1}^{M} \delta_{k,m}\delta^{T}_{k,\widetilde{m}}/P,$$
$$\text{where } \delta_{k,m} = b_{i,m} - b_{j,m} \quad (1)$$
$$\text{and } \delta_{k,\widetilde{m}} = \frac{1}{(M-1)}\sum_{n=1,\ n\neq m}^{M} b_{i,n} - b_{j,n}$$

We define a distance metric (d_k), where m indexes the left-out run, and n indexes the rest of the runs. From the Q × P matrices, we have: for the mth run, $b_{i,m}$ is a 1 × P vector for the ith image input and $b_{j,m}$ is a 1 × P vector for the jth image input. The first term, $\delta_{k,m}$ is the difference in activity patterns between them ($b_{i,m}$, $b_{j,m}$) in the left-out m^{th} run. The second term, $\delta_{k,\widetilde{m}}$, is the averaged difference in activity patterns between the same two image inputs for the rest of the runs. The inner product between $\delta_{k,m}$ and $\delta_{k,\widetilde{m}}$ are averaged across all M folds, normalized by the number of voxels (P) in each ROI or voxel cluster.

Cross-validated Euclidean distances have been shown to be unbiased distance estimators, such that they are not conflated with noise[75,76]. This is because we can assume noise to be independent across partitions, so any measured noise between $\delta_{k,m}$ and $\delta_{k,n}$ should point in random directions in the high-dimensional space, and therefore be near-orthogonal. Taking the dot product of $\delta_{k,m}$ and $\delta_{k,n}$, noise should cancel out.

Because this analysis aimed to identify neural invariance properties that matched the invariance properties of the perceptual priors identified in Experiment 1, the line drawing condition was excluded from the analysis, leaving 7 manipulation conditions in total, including the original images (Fig. 3b). Using the 70 image inputs (10 image exemplars × 7 conditions), a 70 × 70 RDM is created using the c.v. Euclidean distances computed for every pair of image inputs (Fig. 3b). Then, to compute within-image, between-condition neural distances (green bars in Supplementary Fig. 5b), the values were averaged across all pairs of manipulation conditions for each image exemplar, then averaged across the 10 image exemplars (green squares in the matrix in Supplementary Fig. 5a). In total, for each subject, 210 values were averaged within each green bar of Fig. 3c, including 21 condition pairs × 10 image exemplars.

fMRI–Model-based RSA

A model RDM was created to capture the invariance properties of perceptual priors identified in Experiment 1. To this end, we created a 70 × 70 model RDM, in the same layout as the neural RDM (Fig. 3b). Because size manipulation did not significantly impact the learning effect, while orientation and position shift manipulations significantly degraded the learning effect, we assumed the distance between original and size-manipulated conditions to be low (distance = 0, in navy), those between original and rotation/inversion/position shift conditions to be intermediate (distance = 0.5, in teal), and distances between different image exemplars to be high (distance = 1, in yellow). The model RDM was correlated with neural RDMs from each ROI using Kendall's Tau-B. For statistics, a null distribution was created by shuffling image labels (across both exemplar image and condition labels). For each ROI, the real correlation value was compared to the null distribution to obtain an empirical p-value (right-tailed).

A whole-brain searchlight analysis was run in each subject's T1 space, using a 6-voxel radius sphere size. For each voxel cluster, the neural RDM was correlated with the model RDM using Kendall's Tau-B. The correlation values were transformed into Fisher's z, and normalized into standard MNI space. For group-level analysis, the z-value maps were spatially smoothed (12 mm FWHM) and submitted to a one-sample t-test across subjects. Significance was assessed by permutation (using FSL randomize) and the threshold-free cluster enhancement (TFCE) method[77], thresholded at a $p < 0.05$, FWE-corrected level.

iEEG experiment

Experimental setup. Participants performed the task while reclined in their hospital beds with the use of a laptop and USB keyboard. The laptop was placed on a hospital table, with the screen ~60 cm from the participant's eyes and in a position approximately level with their view. All images were presented with sides equaling 8.5° × 8.5° of visual angle. See additional details in SI Methods.

iEEG experimental paradigm. The task design was adapted from previous work with Mooney images[25] for the iEEG setting. Each trial started with a red fixation cross displayed for a pseudorandom amount of time between 1 s and 2 s, followed by image presentation and response collection (Supplementary Fig. 6b). Patients were instructed to fixate on the red cross whenever visible, and to respond "Yes" or "No" to the question "Can you name the object hidden in the image?" as soon as possible upon presentation of an image. "Yes" and "No" responses were collected from distinct hands on a USB keyboard. To maximize the number of trials collected within a limited amount of time, the trial concluded and the image was removed once the patient provided a response. In the absence of a response, after the 2 s image presentation, a response prompt was presented for up to 1 s to allow additional time for responding, after which the trial ended.

Similar to earlier neuroimaging work[24,25], trials were grouped into blocks: three greyscale images were presented in a pseudorandom order, followed by their three corresponding Mooney images (Post images) and three novel Mooney images (Pre images), each presented twice in pseudorandom order, totaling 15 trials per block. A block was repeated to form a run of 30 trials. In total, each greyscale image was presented twice, and each Mooney image was presented four times before and four times after disambiguation. A minority of images were presented more than four times in the Pre/Post phase and more than twice for the Grayscale image due to experimenter error in a small number of recordings. At the end of each run, six verbal response trials were performed in which each Mooney image was presented for 2 s, and the patient was asked to verbally indicate the subject of the image if they could. The verbal responses were used to correct for any erroneous subjective recognition responses to screen out incorrect identification of Mooney images. Thus, any "Yes" keyboard responses to Mooney images during a run in which the corresponding verbal response trial was incorrect were treated as a "No" in further analyses.

iEEG–Data analyses

Electrodes were assigned to an ROI if their MNI coordinates lay within a voxel in that ROI. FPN and DMN ROIs were defined as above, from[24]. EVC, HLVC, and Dorsal ROIs were defined as above, except no voxels were excluded for low probability, and there is no prohibition that a voxel be in only one ROI. The Limbic ROI was defined as the union of the "Cingulate Gyrus, anterior division", the "Frontal Orbital Cortex", and the "Insular Cortex" regions of the Harvard Oxford Anatomical Atlas.

Only successfully disambiguated images were included in most analyses. Images were deemed successfully disambiguated when at least half the Pre trials were not recognized, at least half the Post trials were recognized, and the Grayscale image was viewed at least once (trials in which subjects were distracted away from the task were excluded). Trials presenting successfully disambiguated images are included for analysis if they are either an unrecognized Pre trial, a recognized Post trial, a recognized Grayscale trial, or the final Grayscale trial for the image. For the control analysis in Supplementary Fig. 9, recognized Pre trials were also used. Trials are clipped to end at image offset and then smoothed with a 100-ms moving average window centered on each timepoint. Significance testing was only performed on time points at least 50 ms after image onset to avoid mixing pre- and post-stimulus signals.

The electrode mean activation time course for each condition (Pre, Post, Grayscale) was calculated as the mean across trials for that condition and electrode. The mean time courses for each ROI (Fig. 4, top) were calculated as the average across electrodes within the ROI, for each condition. Shaded error bars were calculated as the paired SEM[78,79] across electrodes at each timepoint. Significance testing was performed using a paired *t*-test at each timepoint (one-tailed, to identify time points with higher HGP in the post than pre stage). We focused on the post > pre effect here to exclude potential task difficulty-related effects, since the pre phase is more difficult, whereas the post phase has heightened prior- and recognition-related processing. To correct for multiple comparisons across time points, a cluster-based permutation test was used (see below). Since trials terminated at image offset when a response was given, the number of trials with data varies across timepoints, especially towards the end of the trial. Once all trials of one condition for an electrode have terminated, the electrode is dropped from that and subsequent timepoints in the corresponding mean ROI time courses. See Supplementary Fig. 7b for each ROI's electrode survival time course.

The Image Preference Analysis quantifies how well each electrode's image selectivity during Pre and Post conditions aligns with its image selectivity during Grayscale image trials (Fig. 4, bottom). For each electrode, the mean HGP time course for each image was calculated, separately for the grayscale, pre, and post phases. At each timepoint, the relative values of each image's mean HGP in the mean Grayscale time courses defined an image sorting order. The Pre (/Post) images' mean HGP values at that timepoint were arranged in the same image sorting order and a best fit linear regression line was calculated. The best fit line's slope was extracted, and this procedure was repeated for every timepoint, producing a time course of slopes for the Pre (/Post) condition at each electrode. Mean traces, paired SEM, and significance were calculated similarly to that described above, except that Wilcoxon signed-ranked tests were used in place of paired *t*-tests. Timepoints at which there are not at least two images with trial data in both the grey condition and the Pre (/Post) condition were dropped. See Supplementary Fig. 7c for each ROI's electrode survival time course.

We used an adapted cluster-based permutation test[80] to accommodate the varying degrees of freedom (DOF) at different timepoints due to the varying trial durations. We defined a cluster's DOF to be the mean DOF for the timepoints it spans, rounded to the nearest integer. We calculated cluster summary statistics on our real data as the sum of the test statistic for each cluster, and also noted the DOF for each cluster. We then produced a permutation-derived null distribution for each DOF using the following procedure. These null distributions were defined so that all data points within one distribution were derived from clusters with the same DOF. For each permutation, we identified the largest cluster summary statistic and its DOF, and its summary statistic was assigned to the corresponding null distribution. A count of permutations that produce no significant clusters was maintained and a corresponding number of zero-valued data points was added to each null distribution proportionately to the number of data points it has once permutation generation has ended. Permutations were generated and this process was repeated until there are 1000 data points in all null distributions with DOFs matching the clusters from the real data, or all possible permutations had been produced, or 1,000,000 permutations had been produced. P-values were then assigned to each cluster by calculating the percentage of data points in the null distribution with the corresponding DOF that were greater than the cluster's summary statistic. This corresponded to a one-tailed cluster-corrected test.

Since massive univariate analysis on individual timepoints (followed by cluster-based permutation test to correct for multiple comparisons) does not statistically assess onset timing, we performed a bootstrap analysis to compare the relevant timings between ROIs in key conditions of interest[81]. For each condition of interest, we generated 2000 bootstrapped replications of the corresponding analysis described above, yielding 2000 sets of significant clusters. An estimate of the onset time distribution for each real cluster in a condition of interest was generated by extracting from each replication the earliest significant timepoint that was part of a cluster that overlapped with the real cluster. Only replications with significant clusters that overlapped with the real cluster contributed points to these distributions, and a minimum of 800 points contributed to the estimated onset time distribution for all conditions of interest[82]. The points at the 2.5th and 97.5th percentiles were reported as the 95% confidence interval (CI) onset.

DNN model architecture

Our model consists of two learnable components, a vision transformer and a cross-attention module, and a memory state module that is updated via a fixed rule. The vision transformer is initialized from the base sized DinoV2 vision transformer, which has 768 hidden units and 12 layers of self-attention mechanisms for processing visual features. The original vision transformer was pretrained using color images, whereas our task used grayscale images. We handled this difference by duplicating the grayscale intensity to the R, G, B channels so as to not bring additional information into the model. This initial vision transformer was then trained on the Mooney image recognition task using Mooney images and their matching grayscale images created from the ImageNet1k dataset (see "DNN Model Training and Evaluation" below). To keep training time reasonable, we use low-rank adaptation (LoRA)[83] to finetune the key and query weight matrices of the self-attention layers only. LoRA learns additional weights that are added to the existing weights, which ensures the visual features already learned is not lost. We use a matrix rank of $r = 64$, and an alpha value of 16. The cross-attention module is used to retrieve information from the state module and is fully trainable.

The model operates in two stages, first a feedforward run to produce retrieval queries, and then a conditioned run to generate an object classification given past knowledge using this query. When initially shown an image, the image is fed into the vision transformer without any previous information. This produces a priming feedforward output that encodes visual representations in the input image. The output of size (1 + 256 tokens, 768 hidden dimensions) is down-sampled to (1 + 9 tokens, 768 hidden dimensions) using adaptive down-sampling to reduce computation cost and generate a query.

This query is then used to compute the feedback conditioning for the conditioned feedforward run. The previous state (1 + 10 + 256 tokens, 768 hidden dimensions) is used as the key and value for the cross attention, while the query generated from the feedforward run is used as the query. This produces a feedback conditioning (1 + 9 tokens, 768 hidden dimensions) representation as the top-down signaling for the next stage of computation.

The second conditioned feedforward pass then occurs with the feedback conditioning providing conditional information from the prior state to the same vision transformer from the prior stage to generate an output. The conditioning tokens (10) are concatenated with the image patch tokens. The second run produces top-down informed outputs (1 + 10 + 256 tokens). The first token corresponds to the classification token (CLS token) from the model, a readout token that is forwarded with the visual tokens and used for classification at the end of the network. 10 tokens are from the conditioning tokens and the 256 tokens correspond to image patches produced from the current input image. Lastly, the state is updated with a moving average with a fixed constant of 0.5 during training and evaluation.

Notable differences from existing models

There are a few architectural details that can explain the difference in performance. Existing models such as BLT and CORnet rely on

convolutional neural networks (CNN)-based layers to extract visual features, whereas our model uses a vision transformer, which encourages more complex interaction between parts of the visual input.

A second crucial difference between our model and BLT and CORnet is how the recurrence is implemented. In our model, the state is a separate module that carries information about the past and gets updated with a simple moving average mechanism. This protects the priors from being overly affected by the current step of activations. On the other hand, BLT and CORnet do not keep a separate state module; all the prior information needs to be stored in the current activation, which sacrifices the expressivity of current visual representation and limits the storage capacity.

Another difference from existing models is that our model is explicitly selective about how prior knowledge affects the current visual processing. Our model achieves this selective retrieval by first computing bottom-up visual features without the conditioning, followed by the retrieval of relevant priors using cross attention. Existing models simply merge the past state with the current activation, where prior activations implicitly modulate current activations and lack fine-grained selectivity of past prior information.

DNN model training and evaluation

We train the network on a computational adaptation of the Mooney Task. To construct a sequence of images, we first obtain all unique color images from the ImageNet dataset at random, convert them to grayscale images, and apply random thresholding between pixel intensity 50 and 205 to each image to produce the binarized version of them ("Mooney image"). For each batch during training, we first obtain 3 pairs of grayscale and Mooney images. To encourage the reuse of information across time and not discard information right after usage, we repeat each constructed sequence 3 times. Finally, we shuffle the entire sequence before feeding it into the model. This same image sequence structure is used in the evaluation shown in Supplementary Fig. 11a. Below, we give an example of a typical sequence used for training the model. We first obtain 3 grayscale images, and notate them as G1, G2, G3. We then binarize them with a random threshold and obtain M1, M2, and M3. We construct a sequence wherein each M-G image pair occurs 3 times: [M1, G1, M1, G1, M1, G1, M2, G2, M2, G2, M2, G2, M3, G3, M3, G3, M3, G3]

Finally, we shuffle this sequence to encourage the model to be able to handle an arbitrary sequence order. E.g., a shuffled sequence produced from the procedure, as described, can be: [G2, G2, M3, G1, M2, G3, M1, G3, G1, M2, G2, M1, M2, M3, M1, G1, G3, M3].

For training, we accumulate gradients for 16 steps with a batch size of 32 sequences on 8× A100 GPUs with the PyTorch Lightning framework. All images in the sequence are weighted equally in the cost function. Supplementary Fig. 11a shows model performance during the evaluation phase, using held-out image sequences following the same shuffled structure. All image sequences used during the training and evaluation phase had a length of 18. Any Mooney image presented before (/after) the first corresponding Grayscale image is designated "Pre" (/"Post") in the results plot. Baseline models were trained and evaluated using the same approach.

We next evaluated our model on 1000 longer synthesized image sequences (sequence length: 630, using 210 unique Mooney images), again created from grayscale images that were randomly selected from ImageNet, and thresholded using a random threshold between 50 and 205 to generate the corresponding Mooney images. Repetition effect was assessed by presenting two repetitions of the same 210 Mooney image in the same sequence, without the corresponding grayscale images (sequence length: 420). Sequences followed the same block-shuffled structure as the human psychophysics experiment (Fig. 1a). The results from this evaluation are shown in Fig. 5b and Supplementary Fig. 11c.

In an exploratory analysis, we found that image patches that received a higher attention rank during the grayscale image input had greater changes in attention score from pre phase to post phase (Supplementary Fig. 11d). This suggests that the vision backbone's representation of a post-Mooney image shifts towards its representation of the corresponding grayscale image, similar to observations in human fMRI data[21,24].

Model behavior comparison with humans

From an online study, behavioral data were obtained from 12 subjects performing the Mooney Task (see SI Methods). The same 219 Mooney images and their matching grayscale images were presented to each subject in a unique order. We supplied each image sequence used in the human experiment to the model and obtained 12 sets of model behavioral data. When comparing model behavior to human behavior, we first take the top 90 Mooney images on which human subjects showed the strongest perceptual learning effect, measured by the average pre-to-post accuracy increase (the same 90 images were used in the in-person behavioral experiments). We then exclude any images that are not in our model's label space due to the model's inability to predict images outside of the label space. This resulted in 78 unique Mooney images (and their corresponding grayscale images) used in the model-human behavior comparison shown in Fig. 5c, d. Equivalent results for baseline models are shown in Supplementary Fig. 11b.

To obtain the results in Fig. 5d, we first transformed each subject's behavioral output or model output into a binary 1×234 vector denoting correct/incorrect identification of a specific image shown in the pre/gray/post phase. We then calculated pairwise AUROCs between human subjects and between the 12 sets of model outputs, yielding 66 error pattern similarity measures for each. We further obtained 144 (12×12) model-human error pattern similarity measures, which can be separated into those with matching image sequences (12) and those with non-matching image sequences (132).

Learning outcome prediction

Using the same behavioral data set mentioned above, for each subject ($n = 12$), we removed images that were already recognized in the pre phase. We then obtained internal model features from the vision transformer layers and their outputs including CLS token and logits (from the 2nd pass after top-down conditioning), and used an SVM classifier with a learnable linear kernel from the sklearn python package to predict whether the subject will recognize the image in the post phase. To avoid data leakage, we applied a 12-fold cross-validation and report the mean test performance over folds. AUROC was used to measure the performance.

Statistical test for Fig. 5e uses two-sided t-tests with FWE correction. FWE correction is done by permuting the trial labels for 5000 repetitions and collecting maximum t statistic over all layers as the null distribution. Finally, the p-value of the real t statistic is obtained by finding its percentile in this null distribution.

Neural (fMRI) data prediction from DNN

To predict the post phase brain beta values (outputs from the GLM at the individual image level, see ref. 24), we use a kernel ridge regression model with sparsity alpha = 1, using the latent visual portion of the recurrent model state as features (see Supplementary Fig. 10). To obtain a set of beta map predictions without overfitting, we use leave-one-image-out cross validation for each subject. All predictions were made after registering subject-level beta maps to the standard MNI152 space, and predictions from 4 different model seeds were averaged together.

The baseline predictions are obtained from constructing counterfactual model representations, using catch images. The counterfactual representation for an image I is obtained by replacing only the corresponding grayscale image for I (i.e., replacing the correct learning

material). The average counterfactual representation over all possible alternative learning material (32 alternative images) is used as the feature.

We first find the correlation between predicted betas and the true betas. We then apply z-transform to the Pearson's correlation of the predictions. Repeating this process for the normal beta predictions and the baseline beta predictions gives us two maps per subject. Subtracting the baseline score map from the normal score map, we obtain the map of score improvement.

To calculate the information strength of a region, we used the following procedure.

Let's denote the beta prediction for voxel j and image i as $\hat{\mathcal{Y}}_{ij}$, and the true beta there $\mathcal{Y}_{ij}$. We first calculate the relative error $\frac{\hat{\mathcal{Y}}_{ij}-\mathcal{Y}_{ij}}{\sum_{i'}\left|\mathcal{Y}_{ij}\right|/I}$ for both the normal prediction and the baseline predictions. This provides us one normalized error map per combination of subject and image. Then we compare the error of baseline and normal predictions. This is obtained with $\frac{e_{\text{baseline}}}{e_{\text{normal}}}-1$. This yields one map per subject and image, with 0 indicating no difference between both performances and a higher value indicating improvement in performance. In the following experiments, we only consider maps with strictly non-negative entries in this map. For the ROI level analyses in Fig. 6b–d, we find the median information strength within the ROI. The ROIs were as described earlier; here, we thresholded the ROI masks from the probabilistic atlases at 25% and assigned a voxel to the most likely ROI. We chose a stringent threshold intentionally to highlight the differences between ROIs and minimize overlap between ROIs.

To compare the prediction score with noise ceilings, after we obtain the prediction score, we average the z-transformed scores between 4 different model seeds. We obtain lower and upper noise ceiling maps for each subject with the Representation Similarity Analysis (RSA) method[84]. Specifically, to calculate the upper noise ceiling estimate for a voxel center, we first find its average RDM across subjects. The average similarity of this average RDM to all subject is the upper noise ceiling. To calculate the lower noise ceiling, we use leave-one-subject-out similarity. For each subject, we leave their RDM out, take the average RDM across the remaining subjects, and calculate the similarity between the held-out RDM and the average RDM. To compare the score against either lower or upper noise ceiling, we use a permutation-based 2-sample *t*-test in a whole-brain voxel-wise analysis, with threshold-free cluster enhancement (TFCE) correction with 10,000 permutations. A one-sided comparison with the max statistic was used to produce voxel-level significance values. The results were summarized by ROIs.

To compare the prediction score with the baseline score, we first average the z-transformed normal prediction scores between the 4 different model seeds. We use the same procedure to obtain scores for the baseline predictions. Then we use a permutation based 2-sample *t*-test with TFCE correction with 10,000 permutations. A one-sided comparison with the max statistic is used to produce voxel-level significance values.

To obtain the estimates for the analysis in Fig. 6b, we first obtain a binary label of "learned" versus "not learned" based on the subjective report of the subjects. We fit a binomial family generalized estimating equations (GEE) model grouped by subjects with the logit link function to predict whether the image is learned from information strength in each ROI. The model can be expressed in the R-style formula: learning ~ FC + EVC + LOC + DMN + Dorsal. After fitting the model, the parameters corresponding to each ROI's contribution to learning is tested using *t*-tests.

To obtain the estimates for the analysis in Fig. 6c, like in Fig. 6b, we first obtain the reliability score by excluding the images that are not learned, followed by finding the proportion of post-phase images that are reported as recognized (varying from 4/6 to 6/6 for each unique image). We fit a gamma family GEE model grouped by subjects with log link function to predict the reliability, which is between 0 and 1. The model can be expressed in the R-style formula: reliability ~ FC + EVC + LOC + DMN + Dorsal. After fitting the model, the parameters corresponding to each ROI's contribution to learning is tested using *t*-tests.

To obtain the estimates for the information connectivity between ROIs, we first calculate the Spearman's rho between the pairs of ROIs' information strength for each image. This produces a connectivity matrix between ROIs for each subject. To aggregate this connectivity across subjects, we take the lower triangle of this connectivity matrix and fit a GEE model grouped by subjects. This model can be written in the R-style formula as: connectivity ~ C(ROI pair). After fitting the model, the parameters corresponding to each unique pair of ROIs are tested using *t*-tests.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Online human psychophysics data and previously published fMRI data used for DNN development and testing, as well as model's checkpoints are uploaded to OSF and can be accessed at: https://osf.io/xphte. Source data to reproduce figures related to in-person human behavioral, fMRI, and iEEG analyses can be found at: https://github.com/BiyuHeLab/Mooney2024. Stimuli from the psychophysics paradigm, features used in the pRF analysis, and the full data set from the fMRI experiment can be found at: https://osf.io/jh2kt.

Source data for all figures are provided with this paper. Source data are provided with this paper.

Code availability

The code supporting this work is available in open Github repositories: Code related to human experimental work, including psychophysics, fMRI, and iEEG analyses: https://github.com/BiyuHeLab/Mooney2024. Code related to DNN modeling: https://github.com/nyuolab/MooneyComputationModeling2024.

References

1. Seitz, A. R. Perceptual learning. *Curr. Biol.* **27**, R631–R636 (2017).
2. Dosher, B. & Lu, Z.-L. *Perceptual Learning: How Experience Shapes Visual Perception* (The MIT Press, 2020).
3. Dolan, R. J. et al. How the brain learns to see objects and faces in an impoverished context. *Nature* **389**, 596–599 (1997).
4. Squire, L. R., Frascino, J. C., Rivera, C. S., Heyworth, N. C. & He, B. J. One-trial perceptual learning in the absence of conscious remembering and independent of the medial temporal lobe. *Proc. Natl. Acad. Sci. USA* https://doi.org/10.1073/pnas.2104072118 (2021).
5. Ludmer, R., Dudai, Y. & Rubin, N. Uncovering camouflage: amygdala activation predicts long-term memory of induced perceptual insight. *Neuron* **69**, 1002–1014 (2011).
6. Gregory, R. L. *The Intelligent Eye*. 1st edn. (McGraw-Hill, 1970)
7. Tovee, M. J., Rolls, E. T. & Ramachandran, V. S. Rapid visual learning in neurones of the primate temporal visual cortex. *Neuroreport* **7**, 2757–2760 (1996).
8. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
9. Sorscher, B., Ganguli, S. & Sompolinsky, H. Neural representational geometry underlies few-shot concept learning. *Proc. Natl. Acad. Sci. USA* **119**, e2200800119 (2022).
10. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. Matching networks for one shot learning. In *Adv. Neural Inf. Process. Syst.* (eds Lee, D. et al.) **29**, 3630–3638 (2016).

11. Sung, F, et al. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (eds Forsyth, D., Laptev, I., Oliva, A. & Ramanan, D.) 1199–1208 (2018).
12. Koch, G., Zemel, R. & Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*. (eds Hinton, G. et al.) Vol. 2. 1–30 (2015).
13. Wang, Y.-X., Girshick, R., Hebert, M. & Hariharan, B. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (eds Forsyth, D., Laptev, I., Oliva, A. & Ramanan, D.) 7278–7286 (2018).
14. Zeithamova, D. et al. Brain mechanisms of concept learning. *J. Neurosci.* **39**, 8259–8266 (2019).
15. Behrens, T. E. J. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
16. Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245 (2007).
17. Markman, E. M. M. *Categorization and Naming in Children: Problems of Induction*. (MIT Press, 1991).
18. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
19. Yoon, J., Winawer, J., Witthoft, N. & Markman, E. Striking deficiency in top-down perceptual reorganization of two-tone images in preschool children. In *IEEE 6th International Conference on Development and Learning*. (eds Demiris, Y., Mareschal, D., Scassellati, B. & Weng, J.) 181–186 (2007).
20. Milne, G. A. et al. Perceptual reorganization from prior knowledge emerges late in childhood. *iScience* **27**, 108787 (2024).
21. Hsieh, P. J., Vul, E. & Kanwisher, N. Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *J. Neurophysiol.* **103**, 1501–1507 (2010).
22. Gorlin, S. et al. Imaging prior information in the brain. *Proc. Natl. Acad. Sci. USA* **109**, 7935–7940 (2012).
23. van Loon, A. M. et al. NMDA receptor antagonist ketamine distorts object recognition by reducing feedback to early visual cortex. *Cereb. Cortex* **26**, 1986–1996 (2016).
24. Gonzalez-Garcia, C., Flounders, M. W., Chang, R., Baria, A. T. & He, B. J. Content-specific activity in frontoparietal and default-mode networks during prior-guided visual perception. *eLife* https://doi.org/10.7554/eLife.36068 (2018).
25. Flounders, M. W., Gonzalez-Garcia, C., Hardstone, R. & He, B. J. Neural dynamics of visual ambiguity resolution by perceptual prior. *eLife* https://doi.org/10.7554/eLife.41861 (2019).
26. Gonzalez-Garcia, C. & He, B. J. A gradient of sharpening effects by perceptual prior across the human cortical hierarchy. *J. Neurosci.* **41**, 167–178 (2021).
27. She, L., Benna, M. K., Shi, Y., Fusi, S. & Tsao, D. Y. Temporal multiplexing of perception and memory codes in IT cortex. *Nature* **629**, 861–868 (2024).
28. Ahissar, M. & Hochstein, S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* **8**, 457–464 (2004).
29. Chang, R., Baria, A. T., Flounders, M. W. & He, B. J. Unconsciously elicited perceptual prior. *Neurosci. Conscious* https://doi.org/10.1093/nc/niw008 (2016).
30. Furmanski, C. S. & Engel, S. A. Perceptual learning in object recognition: object specificity and size invariance. *Vis. Res.* **40**, 473–484 (2000).
31. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).
32. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
33. Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
34. Kourtzi, Z., Erb, M., Grodd, W. & Bülthoff, H. H. Representation of the perceived 3-D object shape in the human lateral occipital complex. *Cerebral Cortex* https://doi.org/10.1093/cercor/13.9.911 (2003).
35. Grill-Spector, K. et al. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* **24**, 187–203 (1999).
36. Freeman, J. & Simoncelli, E. P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).
37. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol.* **25**, 595–600 (2015).
38. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. A. Compressive spatial summation in human visual cortex. *J. Neurophysiol.* **110**, 481–494 (2013).
39. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
40. Bougou, V. et al. Neuronal tuning and population representations of shape and category in human visual cortex. *Nat. Commun.* **15**, 4608 (2024).
41. Merigan, W. H. & Maunsell, J. H. How parallel are the primate visual pathways?. *Annu. Rev. Neurosci.* **16**, 369–402 (1993).
42. Goodale, M. A. & Milner, A. D. Separate visual pathways for perception and action. *Trends Neurosci.* **15**, 20–25 (1992).
43. Kveraga, K., Boshyan, J. & Bar, M. Magnocellular projections as the trigger of top-down facilitation in recognition. *J. Neurosci.* **27**, 13232–13240 (2007).
44. Im, H. Y., Cushing, C. A., Ward, N. & Kveraga, K. Differential neurodynamics and connectivity in the dorsal and ventral visual pathways during perception of emotional crowds and individuals: a MEG study. *Cogn., Affect., Behav. Neurosci.* **21**, 776–792 (2021).
45. Cushing, C. A., Im, H. Y., Adams Jr, R. B., Ward, N. & Kveraga, K. Magnocellular and parvocellular pathway contributions to facial threat cue processing. *Soc. Cogn. Affect. Neurosci.* **14**, 151–162 (2019).
46. Levinson, M., Podvalny, E., Baete, S. H. & He, B. J. Cortical and subcortical signatures of conscious object recognition. *Nat. Commun.* **12**, 2930 (2021).
47. Wu, Y. H., Podvalny, E., Levinson, M. & He, B. J. Network mechanisms of ongoing brain activity's influence on conscious visual perception. *Nat. Commun.* **15**, 5720 (2024).
48. Huang, Z. et al. Anterior insula regulates brain network transitions that gate conscious access. *Cell Rep.* **35**, 109081 (2021).
49. Lachaux, J. P., Axmacher, N., Mormann, F., Halgren, E. & Crone, N. E. High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Prog. Neurobiol.* **98**, 279–301 (2012).
50. Crone, N. E., Korzeniewska, A. & Franaszczuk, P. J. Cortical gamma responses: Searching high and low. *Int. J. Psychophysiol.* **79**, 9–15 (2011).
51. Fyall, A. M., El-Shamayleh, Y., Choi, H., Shea-Brown, E. & Pasupathy, A. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *eLife* **6**, e25784 (2017).
52. Mittal, S. et al. Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. In *International Conference on Machine Learning*. (eds Daumé III, H. & Singh, A.) 6972–6986 (2020).
53. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).

54. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Adv. Neural Inf. Process. Syst.* (eds Wallach, H. M. et al.) **32** (2019).
55. Hardstone, R. et al. Long-term priors influence visual perception through recruitment of long-range feedback. *Nat. Commun.* **12**, 6288 (2021).
56. Stokes, M. G. Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
57. Tulving, E. & Schacter, D. L. Priming and human memory systems. *Science* **247**, 301–306 (1990).
58. Biederman, I. & Cooper, E. E. Size invariance in visual object priming. *J. Exp. Psychol.: Hum. Percept. Perform.* **18**, 121 (1992).
59. Cave, C. B. & Squire, L. R. Intact and long-lasting repetition priming in amnesia. *J. Exp. Psychol. Learn Mem. Cogn.* **18**, 509–520 (1992).
60. Gotts, S. J., Chow, C. C. & Martin, A. Repetition priming and repetition suppression: a case for enhanced efficiency through neural synchronization. *Cogn. Neurosci.* **3**, 227–237 (2012).
61. Watanabe, T. et al. Greater plasticity in lower-level than higher-level visual motion processing in a passive perceptual learning task. *Nat. Neurosci.* **5**, 1003–1009 (2002).
62. Rubin, N., Nakayama, K. & Shapley, R. Abrupt learning and retinal size specificity in illusory-contour perception. *Curr. Biol.* **7**, 461–467 (1997).
63. Logothetis, N. K. What we can do and what we cannot do with fMRI. *Nature* **453**, 869–878 (2008).
64. Munkhdalai, T., Faruqui, M. & Gopal, S. Leave no context behind: efficient infinite context transformers with infini-attention. Preprint at https://arxiv.org/abs/2404.07143 (2024).
65. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
66. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
67. Ramsauer, H. et al. Hopfield networks is all you need. Preprint at https://arxiv.org/abs/2008.02217 (2020).
68. Teufel, C. et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc. Natl. Acad. Sci. USA* **112**, 13401–13406 (2015).
69. Zarkali, A. et al. Increased weighting on prior knowledge in Lewy body-associated visual hallucinations. *Brain Commun.* **1**, fcz007 (2019).
70. Li, F.-F., Andreeto, M., Ranzato, M. & Perona, P. Caltech 101 (1.0) [Data set]. *CaltechDATA* https://doi.org/10.22002/D1.20086 (2022).
71. Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).
72. Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* **53**, 103–118 (2010).
73. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).
74. Schutt, H. H., Kipnis, A. D., Diedrichsen, J. & Kriegeskorte, N. Statistical inference on representational geometries. *eLife* https://doi.org/10.7554/eLife.82566 (2023).
75. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**, 188–200 (2016).
76. Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., Kriegeskorte, N. Comparing representational geometries using unbiased-distance-matrix similarity. Preprint at *ArXiv* https://doi.org/10.51628/001c.27664 (2021).
77. Smith, S. M. & Nichols, T. E. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44**, 83–98 (2009).
78. Baria, A. T., Maniscalco, B. & He, B. J. Initial-state-dependent, robust, transient neural dynamics encode conscious visual perception. *PLoS Comput. Biol.* **13**, e1005806 (2017).
79. Morey, R. D. Confidence intervals from normalized data: a correction to Cousineau (2005). *TQMP* **4**, 61–64 (2008).
80. Candia-Rivera, D. & Valenza, G. Cluster permutation analysis for EEG series based on non-parametric Wilcoxon–Mann–Whitney statistical tests. *SoftwareX* https://doi.org/10.1016/j.softx.2022.101170 (2022).
81. Mohsenzadeh, Y., Qin, S., Cichy, R. M. & Pantazis, D. Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife* https://doi.org/10.7554/eLife.36329 (2018).
82. Pernet, C. R., Latinus, M., Nichols, T. E. & Rousselet, G. A. Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: a simulation study. *J. Neurosci. Methods* **250**, 85–93 (2015).
83. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. *ICLR* (eds Finn, et. al.) **1**, 3 (2022).
84. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).

Acknowledgements

This work was supported by a W.M. Keck Foundation medical research grant (to E.K.O. and B.J.H.), an NSF grant (BCS-1926780, to B.J.H. and O.D.), as well as NYU Grossman School of Medicine. We would like to thank Kendrick Kay for providing the pRF data analyzed herein; Thomas Baumgarten, Richard Hardstone, and Daniel Haşegan for support with iEEG preprocessing and analysis; and Larry Squire for discussions on the psychophysics experiment. We would also like to thank Michael Costantino and the NYU Langone HPC team for their support of our AI research efforts.

Author contributions

A.H., J.D.S., X.C.L., E.K.O. and B.J.H. designed research. A.H. conducted psychophysics and fMRI experiments and analyzed data. J.D.S. conducted iEEG experiment and analyzed data. X.C.L. implemented computational model and conducted analyses on the model. D.F., P.D., F.E.P, W.D., O.D., and E.K.O. provided clinical care to iEEG patients and facilitated iEEG data collection. I.S. coordinated and facilitated iEEG data collection at Sinai site. Y.W. provided inputs on the computational modeling work. A.H., J.D.S., X.C.L., E.K.O., and B.J.H. co-wrote the manuscript.

Competing interests

E.K.O. reports equity in Artisight Inc., Delvi Inc., and Eikon Therapeutics. E.K.O. has consulting arrangements with Google Inc., and Sofinnova Partners. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-68711-x.

Correspondence and requests for materials should be addressed to Eric K. Oermann or Biyu J. He.

Peer review information *Nature Communications* thanks the anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.