NEUROSCIENCE

Cortical response to naturalistic stimuli is largely predictable with deep neural networks

Meenakshi Khosla[1], Gia H. Ngo[1], Keith Jamison[2], Amy Kuceyeski[2,3], Mert R. Sabuncu[1,2,4]*

Naturalistic stimuli, such as movies, activate a substantial portion of the human brain, invoking a response shared across individuals. Encoding models that predict neural responses to arbitrary stimuli can be very useful for studying brain function. However, existing models focus on limited aspects of naturalistic stimuli, ignoring the dynamic interactions of modalities in this inherently context-rich paradigm. Using movie-watching data from the Human Connectome Project, we build group-level models of neural activity that incorporate several inductive biases about neural information processing, including hierarchical processing, temporal assimilation, and auditory-visual interactions. We demonstrate how incorporating these biases leads to remarkable prediction performance across large areas of the cortex, beyond the sensory-specific cortices into multisensory sites and frontal cortex. Furthermore, we illustrate that encoding models learn high-level concepts that generalize to task-bound paradigms. Together, our findings underscore the potential of encoding models as powerful tools for studying brain function in ecologically valid conditions.

INTRODUCTION

How are dynamic signals from multiple senses integrated in our minds to generate a coherent percept of the world? Understanding the neural basis of perception has been a long-standing goal of neuroscience. Previously, sensory perception in humans has been dominantly studied via controlled task-based paradigms that reduce computations underlying brain function into simpler, isolated components, preventing broad generalizations to previously unencountered environments or tasks (*1*). Alternatively, functional magnetic resonance imaging (fMRI) recordings from healthy subjects during free viewing of movies present a powerful opportunity to build ecologically sound and generalizable models of sensory systems, known as encoding models (*2*–*7*).

To date, however, existing works on encoding models study sensory systems individually and often ignore the temporal context of the sensory input. In reality, the different senses are not perceived in isolation; rather, they are closely entwined through a phenomenon now well known as multisensory integration (*8*, *9*). For example, specific visual scenes and auditory signals occur in conjunction, and this synergy in auditory-visual information can enhance perception in animals, improving object recognition and event detection as well as markedly reducing reaction times (*10*). Furthermore, our cognitive experiences unfold over time; much of the meaning we infer is from stimulation sequences rather than from instantaneous visual or auditory stimuli. This integration of information from multiple natural sensory signals over time is crucial to our cognitive experience. Yet, previous encoding methodologies have precluded the joint encoding of this rich information into a mental representation of the world.

Accurate group-level predictive models of whole-brain neural activity can be invaluable to the field of sensory neuroscience. These models learn to disregard the idiosyncratic signals and/or noise within each individual while capturing only the shared response relevant to the stimuli. Naturalistic viewing engages multiple brain systems and involves several cognitive processes simultaneously, including auditory and visual processing, memory encoding, and many other functions (*11*). Group-level analysis in this paradigm is enabled by the synchrony of neuronal fluctuations in large areas of the cortex across subjects (*12*). Thus far, intersubject correlation (ISC) analysis (*12*) has been a cornerstone tool for naturalistic paradigms because of its ability to characterize the shared response across individuals. Group-level encoding models adopt an alternative approach for capturing shared response, one grounded in out-of-sample prediction and generalization (*1*). This allows them to model neural activity beyond a constrained stimulus set. However, there is a clear gap between the two mediums of analysis. While ISC analysis suggests that large areas of the cortex exhibit fluctuations that are consistent across subjects, existing neural encoding models have largely focused on predicting activity within predefined functional areas of the brain such as visual and auditory cortices. It is unclear how they may be scaled to develop a single predictive model for whole-brain neural responses, given that naturalistic scenes produce widespread cortical activations. Here, we aim to fill this gap: Provided adequate characterization of stimuli, we hypothesize that the stable component of neural activity across a subject population, i.e., the stimulus-related activity, should be predictable. In the present study, we aim to quantify and improve the encoding of this widespread stimulus-driven cortical activity using rich stimulus descriptions.

Brain responses in real-world conditions are highly complex and variable. Owing to their high expressive capacity, deep neural networks (DNNs) are well suited to model the complex high-dimensional nature of neural activity in response to the multitude of signals encountered during movie watching. Recently, DNNs optimized for image or sound recognition have emerged as powerful models of computations underlying sensory processing (*2*, *4*, *5*, *7*), surpassing traditional models of image or sound representation based on Gabor filters (*3*) and spectrotemporal filters (*13*), respectively, in higher-order processing regions. In this approach, the stimuli presented during brain activity recordings are fed as input to pretrained neural networks, and activations of individual layers are linearly transformed into predictions of neural responses in different regions of the

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. [2]Department of Radiology, Weill Cornell Medicine, New York, NY, USA. [3]Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [4]Nancy E. & Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA.
*Corresponding author. Email: msabuncu@cornell.edu

brain. This approach affords a useful interpretation of these feature spaces as outcomes of a task-constrained optimization, shedding light on how high-level behavioral goals, such as recognition, may constrain representations in neural systems (*2*). While useful, task-driven features may diverge from optimal neural representations, and tuning these features to better match the latter may be both feasible and beneficial (*14*). This approach can help bridge the quantitative gap in explaining neural responses under realistic conditions while improving our understanding of the nature of information processing in the brain. From a purely modeling standpoint, our methodological innovations are threefold. First, we propose an end-to-end deep learning–based encoding model that extracts semantic feature maps from audio and visual recognition networks and refines them jointly to predict the evoked brain response. To this effect, we demonstrate that using different modalities concurrently leads to improvements in brain encoding. Second, we note that cognitive perception during movie watching involves maintaining memory over time and demonstrates the suitability of recurrent neural networks (RNNs) to capture these temporal dynamics. Last, on the basis of existing evidence of hierarchical information processing in visual and auditory cortices (*5*, *7*), we adopt features at multiple levels of abstraction rather than low-level or high-level stimulus characteristics alone. We embed these inductive biases about hierarchy, long-term memory, and multimodal integration into our neural architecture and demonstrate that this comprehensive deep learning framework generalizes remarkably well to unseen data. Specifically, using fMRI recordings from a large cohort of subjects in the Human Connectome Project (HCP), we build group-level encoding models that reliably predict stimuli-induced neuronal fluctuations across large parts of the cortex. As a demonstration of application, we use these encoding models to predict neural activity in response to other task-based stimuli and report excellent transferability of these models to artificial stimuli from constrained cognitive paradigms. This further suggests that these encoding models are able to capture high-level mechanisms of sensory processing.

Approaching multisensory perception through the predictive lens of encoding models has several advantages. Because of their unconstrained nature, encoding models can enable data-driven exploration and catalyze new discoveries. Using six neural encoding models with different temporal scales and/or sensory inputs, trained only on ~36 min of naturalistic data per subject, we can replicate findings from a large number of prior studies on sensory processing. First, by prominently highlighting the transition from short to long temporal receptive windows as we move progressively from early to high-level auditory areas, we can distinguish the cortical temporal hierarchy. Next, by differentiating unisensory cortices from multisensory regions such as the superior temporal sulcus and angular gyrus, we can reproduce the multimodal architecture of the brain. Last, by synthesizing neural responses to arbitrary stimuli such as faces, scenes, or speech, we can demonstrate the functional specialization of known brain regions for processing of these distinct categories. Together, our results highlight the advantages and ubiquitous applications of DNN encoding models of naturalistic stimuli.

MATERIALS AND METHODS

Dataset

We study high-resolution 7T fMRI data of 158 individuals from the Human Connectome Project movie-watching protocol comprising four audiovisual movie scans (*15*, *16*). The movies represent a diverse collection, ranging from short snippets of Hollywood movies to independent Vimeo clips. All fMRI data were preprocessed following the HCP pipeline, which includes motion and distortion correction, high-pass filtering, head motion effect regression using Friston 24-parameter model, automatic removal of artifactual time series identified with independent component analysis (ICA), as well as nonlinear registration to the Montreal Neurological Institute (MNI) template space (*16*). Complete data acquisition and preprocessing details are described elsewhere (*15*, *16*). Last, whole-brain fMRI volumes of size 113 × 136 × 113 are used as the prediction target of all proposed encoding models. Rest periods and the first 20 s of every movie segment were discarded from all analyses, leaving ~12 min of audiovisual stimulation data per movie paired with the corresponding fMRI response. We estimated a hemodynamic delay of 4 s using region of interest (ROI)–based encoding models, as the response latency that yields the highest encoding performance (fig. S2; see the Supplementary Materials for details). Thus, all proposed models are trained to use the above stimuli to predict the fMRI response 4 s after the corresponding stimulus presentation. We train and validate our models on three audiovisual movies with a 9:1 split and evaluate our models on the first three clips of the held-out test movie. Because the last clip in the held-out movie is repeated within the training movies, we excluded it from our analysis.

Methodology

We train six encoding models using different facets of the complex, dynamic movie stimulus. These include (i) audio 1-s and (ii) audio 20-s models, which are trained on single audio spectrograms extracted over 1-s epochs and contiguous sequences of 20 spectrograms spanning 20 s respectively; (iii) visual 1-s and (iv) visual 20-s models, trained with last frames of 1-s epochs and sequences of 20 evenly spaced frames within 20-s clips respectively; (v) audiovisual 1-s and (vi) audiovisual 20-s models, which use audio and visual input as described above, jointly. All models are trained to minimize the mean squared error between the predicted and measured whole-brain response. Figure 1 depicts the overall methodology for training different encoding models.

Stimuli

Audio. We extract mel-spectrograms over 64 frequency bands between 125 and 7500 Hz from sound waveforms to represent auditory stimuli in ~1-s epochs, following (*17*). The audio spectrogram is treated as a single grayscale 96 × 64 image, denoted by x_t^a, for the short-duration model. For the longer-duration model, the input is simply a contiguous sequence of 20 of these grayscale images, represented as $s_t^a = \{x_i^a\}_{i=t-19}^{t}$. This representation of auditory input is also supported by strong evidence that suggests the cochlea may be providing a spectrogram-like input to the brain for information processing (*18*).

Visual. All videos were collected at 24 fps. We extract the last frame of every second of the video as a 720 × 1280 × 3 RGB (red green and blue channels) input, denoted by x_t^v, for the 1-s models. We emphasize that the input here is a single RGB frame, and we are using the 1-s terminology only to be consistent with the nomenclature for audio models. We further arrange the last frame of every second in a 20-s clip into a sequence of 20 images, denoted by $s_t^v = \{x_i^v\}_{i=t-19}^{t}$, to represent the continuous stream of visual stimuli. These are presented to the longer-duration visual 20-s and audiovisual 20-s models.

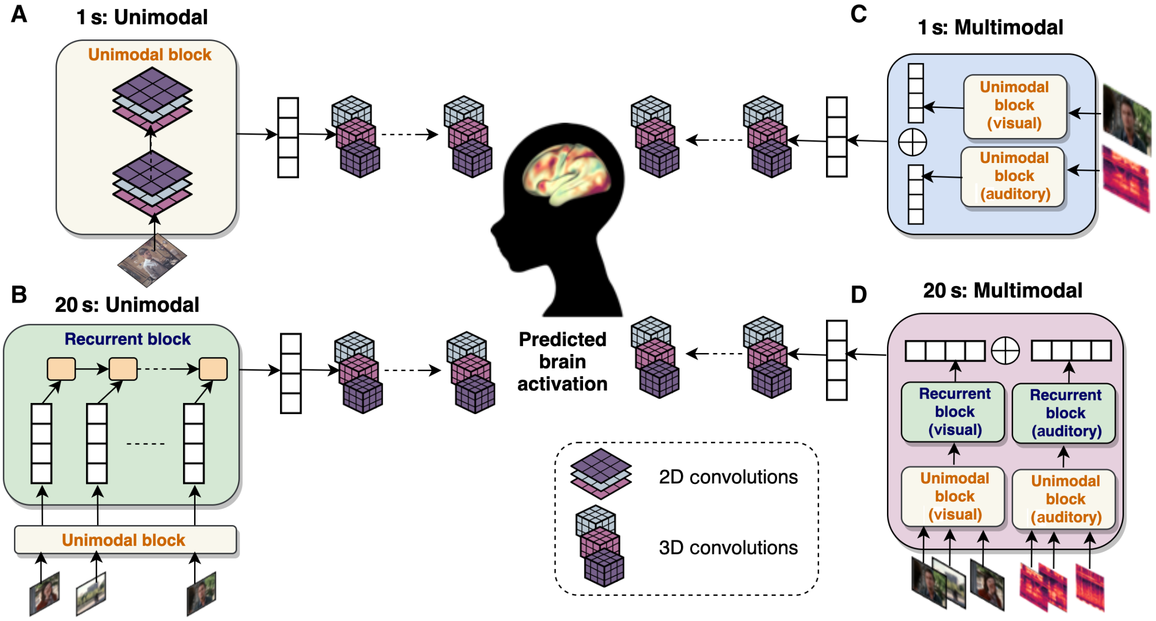

Fig. 1. Schematic of the proposed models. (**A**) The short-duration (1 s) auditory and visual models take a single image or spectrogram as input, extract multiscale hierarchical features, and feed them into a convolutional neural network (CNN)–based response model to predict the whole-brain response. (**B**) The long-duration (20-s) unimodal models take a sequence of images or spectrograms as input, feed their hierarchical features into a recurrent pathway, and extract the last hidden state representation for the response model. (**C**) The short-duration multimodal model combines unimodal features and passes them into the response model. (**D**) The long-duration multimodal model combines auditory and visual representations from the recurrent pathways for whole-brain prediction. Architectural details, including the feature extractor and convolutional response model, are provided in the Supplementary Materials.

The inputs to the audio 1-s, visual 1-s, audio 20-s, visual 20-s, audiovisual 1-s, and audiovisual 20-s models are thus given as x_t^a, x_t^v, s_t^a, s_t^v, $\{x_t^a, x_t^v\}$, and $\{s_t^a, s_t^v\}$, respectively.

Audio 1-s and visual 1-s models

Neural encoding models comprise two components: a feature extractor, which pulls out relevant features, **s**, from raw images or audio waveforms, and a response model, which maps these stimuli features onto brain responses. In contrast to existing works that use a linear response model (*4*, *7*), we propose a convolutional neural network (CNN)–based response model where stimulus features are mapped onto neural data using nonlinear transformations. Previous studies have reported a cortical processing hierarchy where low-level features from early layers of a CNN-based feature extractor best predict responses in early sensory areas, while semantically rich deeper layers best predict higher sensory regions (*5*, *7*). To account for this effect, we use a hierarchical feature extractor based on feature pyramid networks (*19*) that combines features from early, intermediate, and later layers simultaneously. The detailed architectures of both components, including the feature extractor and convolutional response model, are described in fig. S3. We use state-of-the-art pretrained ResNet-50 (*20*) and VGG-ish (*17*) architectures in the pyramid network to extract multiscale features from images and audio spectrograms, respectively. The base architectures were selected because pretrained weights of these networks optimized for behaviorally relevant tasks (recognition) on large datasets, namely, ImageNet (*21*) and YouTube-8M (*22*), were publicly available. ResNet-50 was trained on image classification with 1000 classes, while the VGG-ish network was pretrained on audio event recognition with ~30K categories. Furthermore, because of computational and memory budget, the ResNet-50 was frozen during training across all models. On the other hand, we were able to fine-tune the VGG-ish network in both the audio and audiovisual encoding models. We note that in contrast to images, there is a clear asymmetry in the axes of a spectrogram, where the distinct meanings of time and frequency might warrant one-dimensional (1D) convolutions over time instead of 2D convolutions over both frequency and temporal axes. However, we found the benefits of a pretrained network to be substantial in training convergence time and hence did not explore more appropriate architectures.

Audio 20-s and visual 20-s models

Audio 20-s and visual 20-s models use the same feature extractor and CNN response model as their 1-s counterparts. However, here, the feature extraction step is applied on each image in a sequence of 20 frames, followed by a long short-term memory (LSTM) module to model the temporal propagation of these features. The output dimensions of the LSTM unit are set to 1024 and 512 for the visual and auditory models, respectively, to ensure an equitable comparison with the corresponding 1-s models. The last hidden state output of this LSTM unit is fed into the CNN response model with the same architecture as the 1-s models.

Audiovisual 1-s and audiovisual 20-s models

Meaningful comparison across different models requires the control of as many design choices as possible. To ensure fair comparisons, the audiovisual 1-s model uses the same feature extractors as the visual 1-s and audio 1-s models. The only difference, here, is that the corresponding 1024-D and 512-D feature representations are concatenated before presentation to the CNN response model, and the concatenated features are passed into a bottleneck layer to reduce the final feature dimensionality to the maximum among audio and visual feature dimensions, i.e., 1024, so that the multimodal model is not equipped with a higher-dimensional feature space than the maximum among unimodal models. We note that the response model has the same architecture across all six proposed models. Similarly, the audiovisual 20-s model uses the same feature extraction scheme as the visual 20-s and audio 20-s models but fuses the last hidden state output of the respective LSTM units by simple concatenation followed by a dense layer to reduce feature dimensionality to 1024 before feeding it into the response model.

Evaluation

We first evaluated the prediction accuracy of all models on the independent held-out movie by computing Pearson correlation coefficient (R) between the measured and predicted response at every voxel. Here, the "measured" response refers to the group-averaged response across the same group of 158 subjects on which the models were trained. Comparison among these models enables us to tease apart the sensitivity of individual voxels to input time scales and different sensory stimuli. Voxel-level correlation coefficients between the predicted and measured responses were averaged to summarize the prediction accuracy of each model in relevant cortical areas (Fig. 2, B to F). For this region-level analysis, ROIs were derived with a comprehensive multimodal parcellation of the human cortex (*23*), which was mapped onto the MNI 1.6-mm-resolution template. We note that ROIs were used only to interpret the results of the study and relate them to existing literature.

We emphasize that all performance metrics reported henceforth are based on voxel-level correlations. It is important to note that prediction accuracy at every voxel is bounded by the proportion of non–stimulus-related variance that reflects measurement noise or other factors. We thus also show the regional-level performance of all models against the reliability ("noise ceiling") of measured responses within those regions (Fig. 3).

Noise ceiling estimation

The reliability of the group-averaged response at each voxel is estimated from a short 84-s clip that was repeatedly presented at the end of all movie sessions. We compute an effective upper bound on our performance metric, i.e., the correlation coefficient, as the correlation between the measured fMRI response (group-mean) during different runs. We repeat this process six times (choosing pairs from four repeat measurements) to get a mean noise ceiling estimate per voxel, as shown in Fig. 3D. We divide the voxel-level prediction accuracy (R) by this noise ceiling to get noise-normalized prediction accuracy of all models in the left panels of Fig. 3 (A to C). We note that this noise ceiling is computed on the repeated video clip, which is distinct from the test movie on which the model performance metrics are computed. Direct comparison against this noise ceiling can be suboptimal, especially if the properties of the

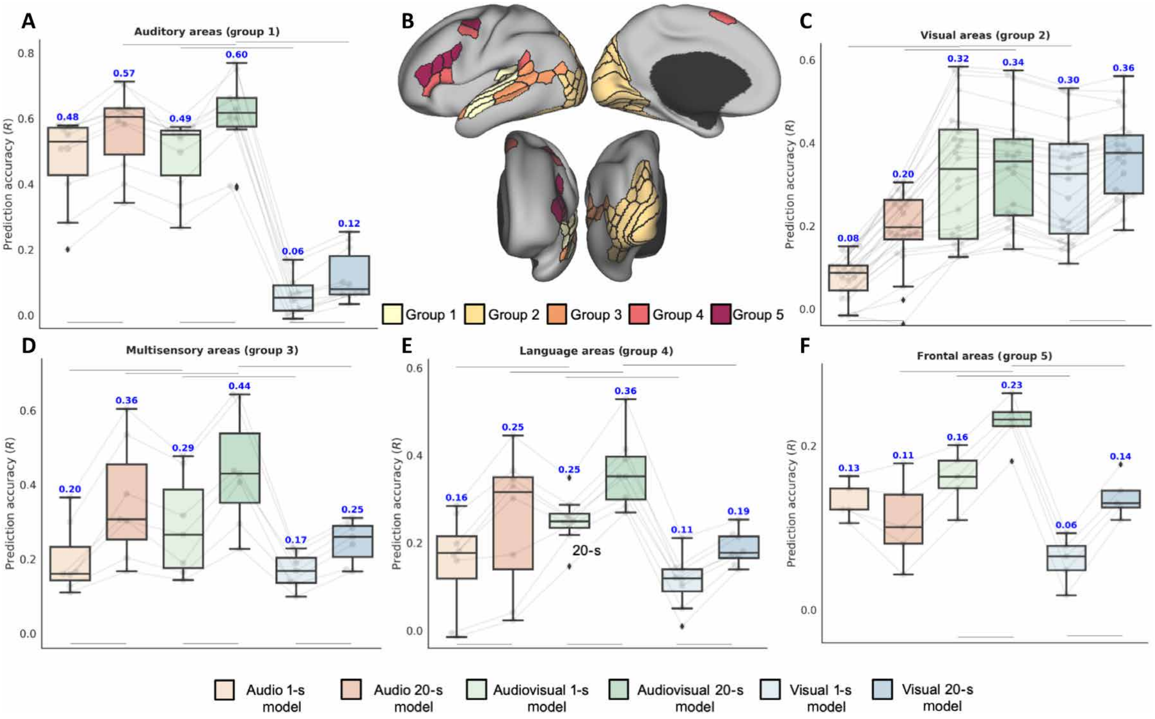

Fig. 2. Regional predictive accuracy for the test movie. (**A** and **C** to **F**) Quantitative evaluation metrics for all the proposed models across major groups of regions as identified in the HCP MMP parcellation (**B**). Predictive accuracy of all models is summarized across (A) auditory, (C) visual, (D) multisensory, (E) language, and (F) frontal areas. Box plots depict quartiles, and swarmplots depict mean prediction accuracy of every ROI in the group. For language areas (group 4), left and right hemisphere ROIs are shown as separate points in the swarmplot because of marked differences in prediction accuracy. Statistical significance tests are performed to compare 1-s and 20-s models of the same modality (three comparisons; results are indicated with horizontal bars below the box plots) or unimodal against multimodal models of the same duration (four comparisons; results are indicated with horizontal bars above the box plots) using the paired *t* test ($P < 0.05$, Bonferroni corrected) on mean prediction accuracy within ROIs of each group.

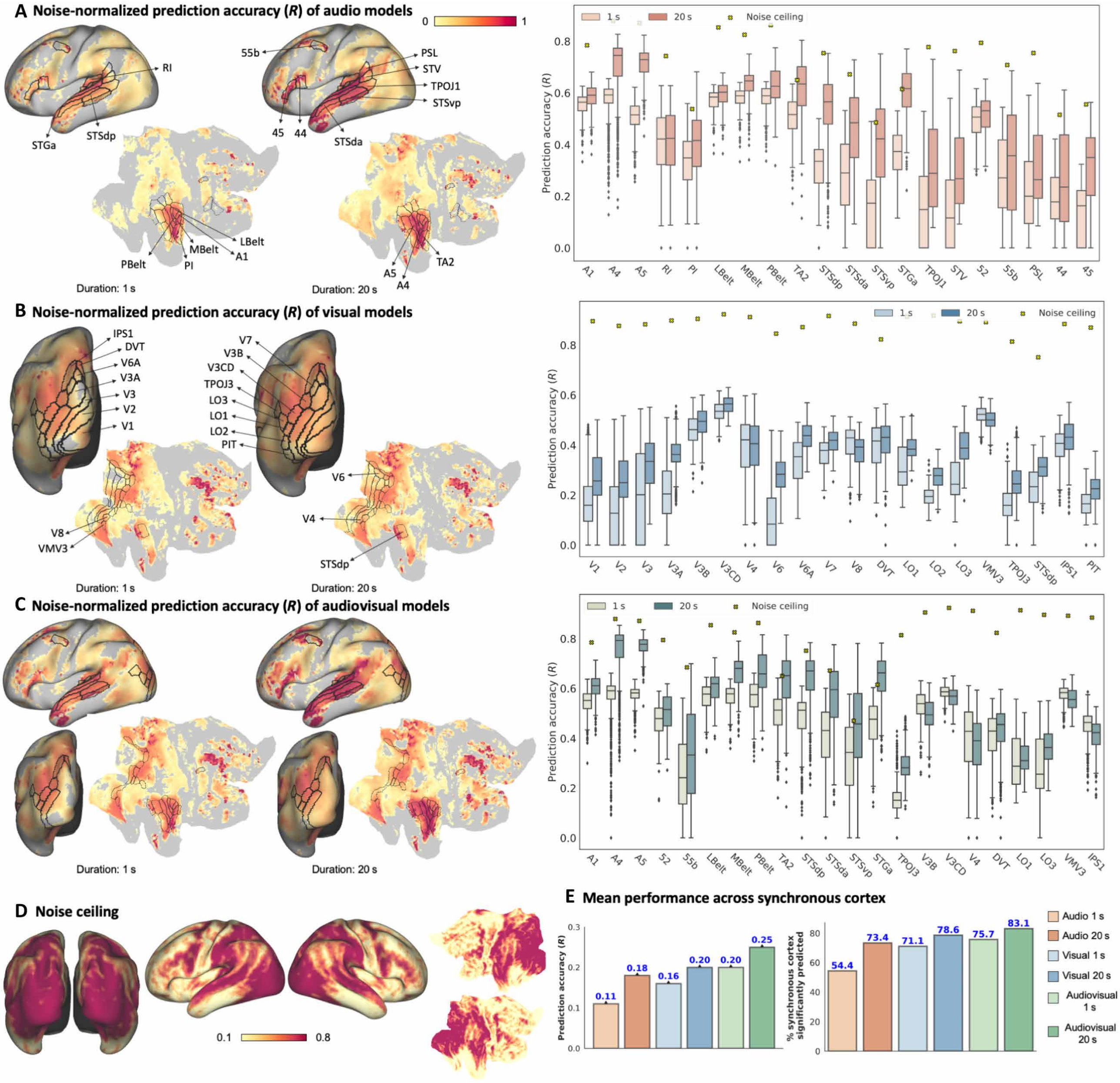

Fig. 3. Model prediction accuracy in standard brain space. Left panel depicts the predictive accuracy of unimodal (**A** and **B**) and multimodal (**C**) models over the whole brain in the test movie. Colors on the brain surface indicate the Pearson correlation coefficient between the predicted time series at each voxel and the true voxel's time series normalized by the noise ceiling (**D**) computed on repeated validation clips. Only significantly predicted voxels [$P < 0.05$, False Discovery Rate (FDR) (*59*) corrected] are colored. ROI box plots depict the un-normalized correlation coefficients between the predicted and measured response of voxels in each ROI and the respective noise ceiling for the mean. (**E**) Percentage of voxels in stimulus-driven cortex that are significantly predicted by each model and mean prediction accuracy across the stimulus-driven cortex.

group-averaged response vary drastically across the two stimulus conditions. We address this limitation during model evaluation against data from a held-out independent group of subjects by computing a more suitable upper bound, which is achievable by a group-level encoding model (fig. S8; see the Supplementary Materials for more details). As we demonstrate in Results (figs. S8 and S9), the trend and spatial distribution of model performance against noise ceiling remain unchanged across the model evaluation and noise ceiling estimation methods.

RESULTS

Multisensory inputs and longer time scales lead to the best encoding performance with significant correlations across a large proportion of the stimulus-driven cortex

To gain quantitative insight into the influence of temporal history and multisensory inputs on encoding performance across the brain, we computed the mean prediction accuracy in five groups of regions defined as per the HCP multi-modal parcellation (MMP) parcellation (*23*), namely, (i) auditory regions comprising both early and association areas, (ii) early visual and visual association regions, (iii) known multisensory sites and regions forming a bridge between higher auditory and higher visual areas, (iv) language-associated regions, and (v) frontal cortical areas. As our research concerns stimulus-driven processing, only ROIs belonging to the "stimulus-driven" cortex were included in the above groups (table S2; see the Supplementary Materials for the definition of "stimulus-driven" cortex).

Groups 1 and 2, which are associated with a single modality (auditory or visual), do not show any marked improvement from audiovisual multisensory inputs and are best predicted by features of their respective sensory stimulus (Fig. 2, A and C). The performance boost with multisensory inputs is more pronounced in groups 3, 4, and 5, which are not preferentially associated with a single modality but are involved in higher-order processing of sensory stimuli (Fig. 2, D to F). Furthermore, temporal history of the stimulus yields consistent improvement in prediction performance in almost all groups of regions, albeit to different extents. Improvements in groups 3, 4, and 5 agree well with the idea that higher-order sensory processing as well as cognitive and perceptual processes, such as attention and working memory, are hinged upon the history of sensory stimuli; therefore, accumulated information benefits response prediction in regions recruited for these functions. Furthermore, both auditory and visual association cortices are known to contain regions that are responsive to sensory information accumulated over the order of seconds (*24*). This potentially explains the significant improvement observed for long–time scale encoding models compared with their short–time scale counterparts in these sensory cortices (Fig. 4). Together, the audiovisual 20-s model integrating audiovisual multisensory information over longer time scales yields maximum prediction accuracy (*R*) and highest percentage (~83%) of significantly predicted voxels across the stimulus-driven cortex (Fig. 3E), suggesting that the audiovisual 20-s model can adequately capture complementary features of each additional facet (multisensory stimuli/temporal information) of the sensory environment.

Longer time scales improve encoding performance, particularly in higher-order auditory areas

As a movie unfolds over time, the dynamic stream of multimodal stimuli continuously updates our neural codes. Evidence from neuroimaging experiments suggests that different brain regions integrate information at different time scales; a cortical temporal hierarchy is reported for auditory perception where early auditory areas encode short time scale events while higher association areas process information over longer spans (*25*). This temporal gradient of auditory processing is well replicated within our study. Comparison of 1-s and 20-s models allows us to distinguish brain regions that process information at shorter time scales from those that rely on longer dynamics. There is a small, albeit significant, contribution of longer time scale inputs on prediction correlations in regions within early auditory cortex, such as A1, LBelt, PBelt, MBelt, and restroinsular cortex (RI) (Figs. 3A and 4A), in line with previous reports suggesting short temporal receptive windows (TRWs) of early sensory regions (*25*). Shorter integration windows are in agreement with the notion that these regions facilitate rapid processing of the instantaneous incoming auditory input. In contrast, response in voxels within auditory association ROIs lying mainly in the superior temporal sulcus or along the temporal gyrus (A4, A5, STSda, STSva, STSdp, STSvp, STGa, and TA2) is seen to be much better predicted with longer time scales (Figs. 3A and 4A). Cumulatively across association ROIs, the audio 20-s model yields a highly significant improvement in prediction accuracy (~50%) over the audio 1-s model, in comparison to a smaller improvement (~5%) across early auditory ROIs.

Longer time scales lead to significantly better predictions in the dorsal visual stream and medial temporal complex

The distinct association of dorsal visual stream with spatial localization and action-oriented behaviors and ventral visual stream with object identification is well documented in the literature (*26*). Another specialized visual area is the medial temporal (MT^+) complex,

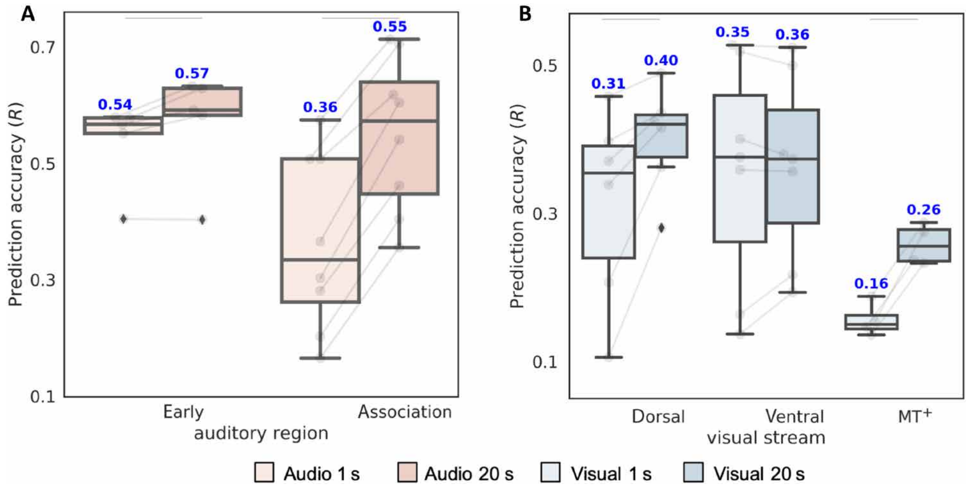

Fig. 4. Influence of temporal history on encoding performance. (**A**) Mean predictive performance of audio 1-s and audio 20-s models in early auditory and association auditory cortex ROIs. A major boost in encoding performance is seen across auditory association regions with the 20-s model. (**B**) Mean predictive performance of visual 1-s and visual 20-s models across ROIs in the dorsal, ventral, and MT^+ regions. Dorsal stream and MT^+ ROIs exhibit a significant improvement with visual 20-s model, but no effect is observed for the ventral stream. Box plots are overlaid on top of the beeswarm plot to depict quartiles. Horizontal bars indicate significant differences between models in the mean prediction accuracy within ROIs of each stream using the paired *t* test ($P < 0.05$).

which has been shown to play a central role in motion processing. The functional division between these streams thus suggests a stronger influence of temporal dynamics on responses along the dorsal pathway and MT^+ regions. To test this hypothesis, we contrast the encoding performance of visual 1-s and visual 20-s models across the three groups by averaging voxel-wise correlations in their constituent ROIs. In accordance with the dorsal/ventral/MT^+ stream definition in the HCP MMP parcellation, we use the following ROIs for analysis: (i) dorsal—V3A, V3B, V6, V6A, V7, and IPS1; (ii) ventral—V8, ventral visual complex (VVC), posterior inferotemporal (PIT) complex, fusiform face complex (FFC), and ventromedial visual areas 1, 2, and 3; and (iii) MT^+—MT, medial superior temporal (MST), V4t, and the fundus of the superior temporal (FST). Figure 4B demonstrates the distribution of mean correlations over these ROIs for different models and streams. Our findings suggest that temporal history, as captured by the visual 20-s model, can be remarkably beneficial to response prediction across the dorsal visual stream (~30% improvement over visual 1-s model) and the MT^+ complex (~62% improvement over visual 1-s model), in agreement with our a priori hypothesis. Furthermore, in our experiments, no marked improvement was observed for the ventral visual stream, indicating a nonsignificant influence of temporal dynamics on these regions.

Auditory and visual stimuli features jointly approach the noise ceiling in multisensory areas

Examining prediction accuracy against response reliability allows us to quantify how far we are from explaining predictable neural activity. A high fraction of the stimulus-driven cortex (~83%) is predictable with a longer time scale input and joint audiovisual features. Notably, areas extending anteriorly and posteriorly from the primary auditory cortex such as the posterior STS, STGa, and TA2 achieve prediction correlations close to the noise ceiling with the audiovisual 20-s model (Fig. 3C), suggesting that DNN representations are suited to encode their response.

Performance in auditory regions is much closer to the noise ceiling than visual regions. Understanding audition and vision in the same space further allows us to appreciate the differences between these modalities. While this may suggest that audition is perhaps a simpler modality to model, the differences could also result from a bias of the dataset. A more diverse sampling of acoustic stimuli in the training set could allow the model to generalize better in auditory regions. Furthermore, in contrast to auditory stimulation where all subjects hear the same sounds, visual stimulation can elicit highly varied responses dependent on gaze location. This variability could plausibly make group-level visual encoding a more difficult task.

Joint encoding models tease apart the modal sensitivity of voxels throughout the sensory cortex

Neural patterns evoked by movies are not simply a conjunction of activations in modality-specific cortices by their respective unisensory inputs; rather, there are known cross-modal influences and regions that receive afferents from multiple senses (*27*). Can we interrogate a joint encoding model to reveal the individual contribution of auditory and visual features in encoding response across different brain regions? To address this question, we shuffled inputs of either modality along the temporal axis during inference. We measured test performance of the trained audiovisual model on predictions generated by shuffling inputs of one modality while keeping the other one intact. This distortion at test time allows us to identify areas that are preferentially associated with either visual or auditory modality. We hypothesized that regions encoding multisensory information will incur loss in prediction accuracy upon distortion of both auditory and visual information. Furthermore, unisensory regions will likely be adversely affected by distortion of either auditory or visual information but not both. To test this hypothesis, we further developed a sensory sensitivity index that directly reflects the sensitivity of individual brain regions to information about auditory or visual stimuli (see the Supplementary Materials for details). For this examination, we used the audiovisual 1-s model to avoid potential confounds associated with temporal history, although analysis of the audiovisual 20-s model showed similar results. Figure 5 demonstrates the result of this analysis on sensory-specific regions and regions known for their involvement in multisensory integration. The benefit from (nondistorted) multisensory inputs to the prediction correlations of the audiovisual model is most seen in posterior STS, STGa, and sensory-bridge regions such as the temporal-parietal-occipital junction and superior temporal visual area. Another region that seems to be using features of

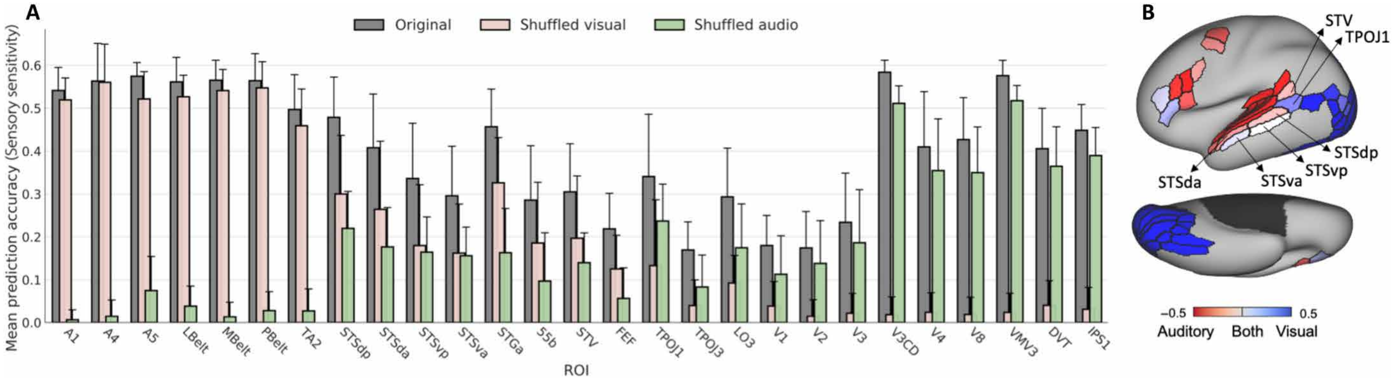

Fig. 5. Sensitivity of ROIs to different sensory inputs. (**A**) Predictive accuracy (*R*) of audiovisual encoding model with and without input distortions, (**B**) Sensory sensitivity index of different brain regions as determined using performance metrics under input distortion (see the Supplementary Materials for details). Regions dominated by a single modality are shown in darker colors, whereas light-colored regions are better predicted by a combination of auditory and visual information. Red indicates auditory-dominant regions, whereas blue indicates visual dominance.

both modalities, albeit to a lesser extent, is the frontal eye field, whose recruitment in audiovisual attention is well studied (*28*).

Classically, multisensory integration hubs are identified as regions that show enhanced activity in response to multisensory stimulation as opposed to presentation of either unisensory stimuli based on some statistical criteria (*29*). Accordingly, the posterior STS is consistently described as a multisensory convergence site for audiovisual stimuli (*9*, *27*, *29*, *30*). Its role in audiovisual linguistic integration has also been well studied in the literature (*28*). Other multisensory integration sites reported extensively in prior literature include the temporoparietal junction (*9*, *27*, *28*) and superior temporal angular gyrus (*31*). Our findings above lend strong support for the multisensory nature of all these regions.

Encoding models as virtual neural activity synthesizers

Next, we sought to characterize whether encoding models can generalize to novel task paradigms. By predicting neural activity for different visual categories from the category-specific representation task within the HCP working memory paradigm, we generated synthetic functional localizers for the two most common visual classes: faces and places. Specifically, we predict brain response to visual stimuli, comprising faces, places, tools, and body parts from the HCP task battery. We use the predicted response to synthesize contrasts (FACES-AVG and PLACES-AVG) by computing the difference between mean activations predicted for the category of interest (faces or places, respectively) and the average mean activations of all categories at each voxel (Fig. 6). The predicted and measured contrasts are thresholded to keep the top 5, 10, or 15% most activated voxels. We report the Dice overlap between the predicted and measured contrasts for each of these threshold values to quantify the agreement between these cortical maps. We also computed the Dice overlap of the predicted contrast for each experiment against all 86 measured task fMRI (tfMRI) contrasts provided as part of the HCP task battery to assess the identifiability of the synthetic contrast.

We observe a notable overlap between the synthetic and measured group-level contrasts. We find that the synthetic contrasts for FACES-AVG and PLACES-AVG are identifiable in that the synthetic contrast exhibits the highest agreement with the measured contrast of the same contrast condition. Furthermore, our findings are consistent with the well-known cortical specificity of neuronal activations for processing of faces and places. Both the synthetic and measured face contrasts are consistent with previously identified regions for face-specific processing, including the fusiform face area (corresponds to FFC in Fig. 6), the occipital face area in lateral occipital cortex (overlaps with the PIT complex in HCP MMP parcellation), and regions within the temporo-parieto-occipital junction and STS (*32*, *33*). Among these, the selective role of the fusiform face area in face processing has been most consistently and robustly established. Another region known to respond more strongly to faces than other object categories, namely, posterior STS, has been previously implicated in processing of facial emotions (*32*).

Similarly, both synthetic and measured place contrasts highlight cortical regions thought to be prominent in selective processing of visual scenes. These include the parahippocampal areas (PHA1–3), retrosplenial cortex (POS1 in HCP MMP parcellation), and the transverse occipital sulcus (TOS), which comprises the occipital place area (OPA) (*34*).

Cortical areas related to speech processing are similarly found using our models by contrasting activations predicted for speech stimuli against nonspeech stimuli such as environmental sounds (Fig. 6B, see the Supplementary Materials for more details). The synthetic contrast shows increased activation in language-related areas of the HCP MMP parcellation such as 55b, 44, and the superior frontal language area with left lateralization, in accordance with previous language fMRI studies (*35*). In addition, areas tuned for voice processing in STS (*36*) are also highlighted. The synthetic map also shows highest correlation with "speech" on neurosynth term-based meta-analysis (*37*) and overlaps considerably with the speech association template on the platform. Together, these experiments illustrate the potential of encoding models to simulate contrasts and reconcile contrast-based studies with naturalistic experiments.

Additional analyses

In prior studies, neural response prediction is done via regularized regression, where the signal at each voxel is modeled as a weighted sum of stimulus features with appropriate regularization on the regression weights. Following earlier works, we also train l_2-regularized regression models using features derived from hierarchical convolutional networks trained on image or sound recognition such as those used in the proposed models, as well as semantic category features labeled using the WordNet semantic taxonomy similar to (*38*). The latter are typically used for mapping the semantic tuning of individual voxels across the cortex. Our models consistently outperform the baselines, further illustrating the benefits of the proposed methodology (fig. S4, A to C; see the Supplementary Materials for more details). In addition, we also performed ablation studies to understand the influence of different network components, namely, the "nonlinear" response model as well as the "hierarchical" feature extractor on model prediction performance, and found that both components improve performance, although their relative contribution is stronger in visual encoding models than auditory models (fig. S4D; see the Supplementary Materials for more details). The superior predictive performance of our models in comparison to the classical approach along with our ablation studies suggests that an interplay of end-to-end optimization with a nonlinear response model can jointly afford improved generalization performance.

To test the generalizability of the models beyond the subject population they were trained on, we further compared the predictions of all models against the group-averaged response of a held-out group within HCP comprising 20 novel subjects distinct from the 158 individuals used in the training set, on the same independent held-out movie. The noise ceiling for this group was computed as the correlation coefficient between the mean measured response for the independent test movie across all 158 subjects in the training set and the group-averaged response computed over the 20 new subjects. This metric captures the response component shared across independent groups of subjects and thus reflects the true upper bound achievable by a group-level encoding model. As shown in fig. S8 (see the Supplementary Materials for more details), the models can accurately predict neural responses as measured with respect to the group mean of the held-out subjects, with the audiovisual 20-s model performance even approaching noise ceiling in some regions, particularly the higher-order auditory association regions and multisensory sites such as the posterior STS. The predictivities across the cortical surface are consistent with the performance metrics reported for the training subject population in Fig. 3. Last, by comparing model predictions against neural responses at the single subject level for subjects from the held-out group, we further

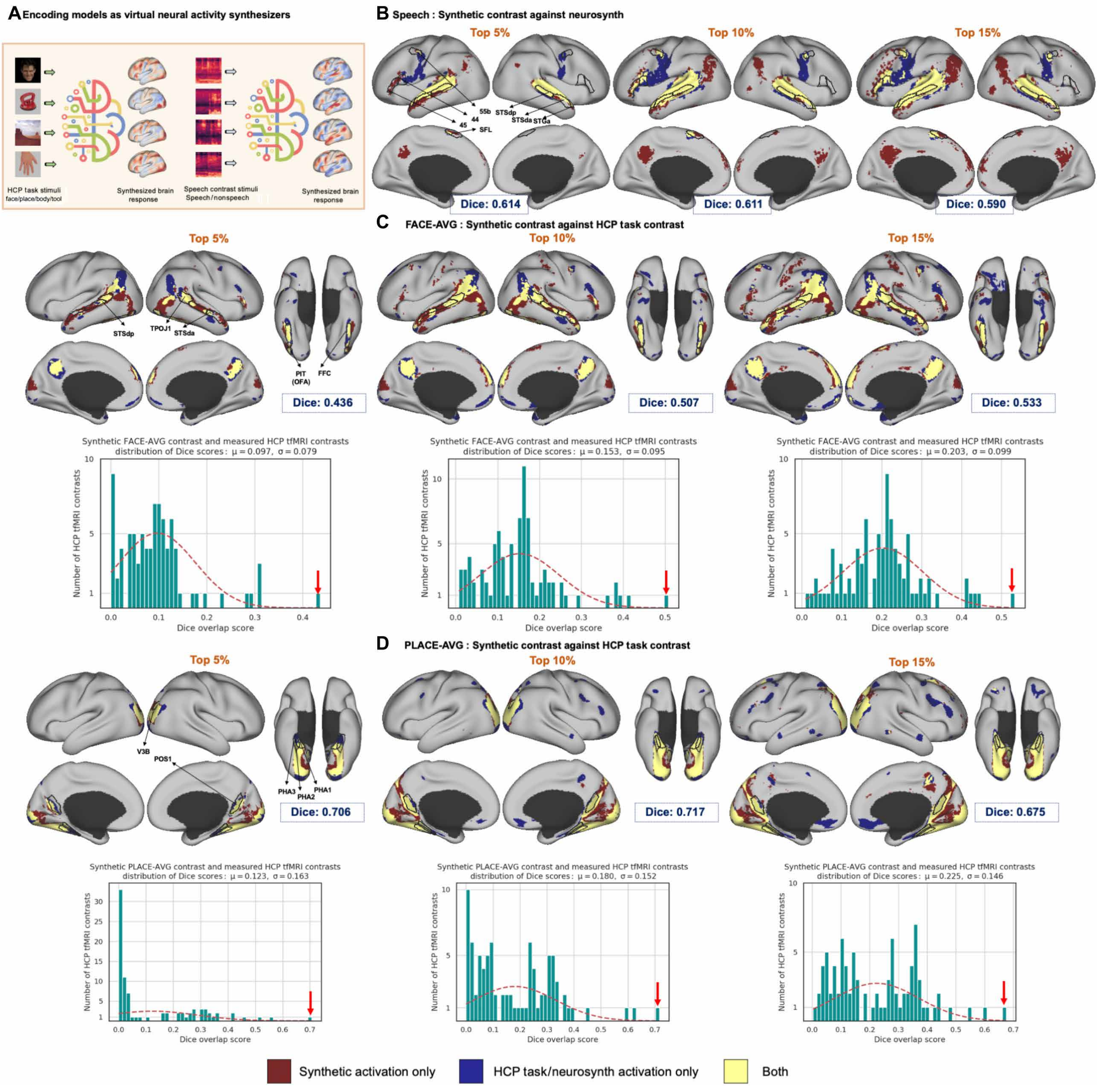

Fig. 6. Encoding models as virtual brain activity synthesizers. (**A**) Synthetic contrasts are generated from trained encoding models by contrasting their "synthesized" (i.e., predicted) response to different stimulus types. (**B**) Comparison of the synthesized contrast for "speech" against the speech association template on neurosynth, both thresholded to keep the top 5, 10, or 15% most activated vertices. (**C**) and (**D**) compare the synthesized contrasts for "faces" and "places" against the corresponding contrasts derived from HCP tfMRI experiments, both thresholded to keep the top 5, 10, or 15% most activated vertices. Vertices activated in only synthetic or predicted contrast maps are shown in red and blue colors, respectively, whereas yellow indicates the overlap. Corresponding Dice scores are displayed alongside the surface maps. Distributions of Dice overlap scores between the synthetic map and all 86 HCP tfMRI contrast maps are shown as histograms at each threshold level. Red arrow points to the Dice overlap between the synthetic contrast and HCP tfMRI contrast for the same condition. In all cases, the synthetic contrast exhibits the highest agreement with the tfMRI contrast that it was generated to predict.

demonstrate that the audiovisual 20-s model can also successfully capture the response component that individual subjects share with the population (fig. S10; see the Supplementary Materials for details).

DISCUSSION

Free viewing of dynamic audiovisual movies enables an ecologically valid analysis of a collective set of functional processes at once, including temporal assimilation and audiovisual integration in addition to momentary sensory-specific processing. Perception, under such stimulation, thus recruits sensory systems as well as areas subserving more sophisticated cognitive processing. Building quantitatively accurate models of neural response across widespread cortical regions to such real-life, continuous stimuli thus requires an integrated modeling of these disparate computations on sensory inputs. Here, we have presented six DNN-based encoding models with varying sensory and temporal information about the audiovisual stimulus. Subsequently, we queried the role of input history and different sensory information on prediction performance across individual regions of the cortex. We have shown that exploiting the richness of the stimulus along the time axis and sensory modalities substantially increases the predictive accuracy of neural responses throughout the cortex, so far as approaching the noise ceiling for voxels in some known multisensory sites, such as the posterior STS (*9*, *27*, *29*, *30*).

Auditory and visual scenes are the principal input modalities to the brain during naturalistic viewing. Yet, existing encoding models ignore their interactions. We use a common strategy in multimodal machine learning settings, namely, feature fusion, to jointly model auditory and visual signals from the environment. We find that minimizing the prediction error is a useful guiding principle to learn useful joint representations from an audiovisual stimulation sequence and demonstrate that models that consume multimodal signals concurrently, namely, audiovisual 1-s and audiovisual 20-s, can not only predict the respective unimodal cortices slightly better but also lead to remarkable improvements in predicting response of multisensory and frontal brain regions (Fig. 2). Furthermore, we show that multimodal neural encoding models not only boost performance in large areas of the cortex relative to their unimodal counterparts (Figs. 2 and 3E) but also shed light on how neural resources are spatially distributed across the cortex for dynamic multisensory perception (Fig. 5). The predictivity of different sensory inputs for neural response, as evaluated on independent held-out data, can facilitate reverse inference by identifying the sensory associations of different brain regions, providing clues into the multisensory architecture of the cortex. By comparative analysis of predictive performance in different regions across models (Fig. 2) as well as perturbation analysis within the multimodal model (Fig. 5), we identify a number of regions that are consistently sensitive to both auditory and visual information, most notably the superior temporal sulcus and some frontal regions. Regions within the inferior frontal cortex have been implicated in the processing of visual speech, guiding sensory inferences about the likely common cause of multimodal auditory and visual signals, as well as resolving sensory conflicts (*39*). Prior research has also implicated an extensive network of inferior frontal and premotor regions in comprehending audiovisual speech, suggesting that they bind information from both modalities (*40*). While unveiling the causal sequence of events for a mechanistic understanding of multisensory perception is not possible with the proposed approach, our findings align well with commonly held theories of sensory fusion, which suggests that unisensory signals are initially processed in segregated regions and eventually fused in regions within the superior temporal lobe, occipital-temporal junction, and frontal areas (*27*). This proposition is corroborated by our experiments as response prediction in these regions is best achieved by a combination of both sensory inputs (Figs. 3 and 5).

A linear response model with pretrained and nontrainable feature extractors, while simple and interpretable, imposes a strong constraint on the feature-response relationship. The underlying assumption is that neural networks optimized for performance on behaviorally relevant tasks are mappable to neural data with a linear transform. We designed a flexible model, capable of capturing complex nonlinear transformations from stimulus feature space to neural space, leading to more quantitatively accurate models that are better aligned with sensory systems. Even better accounts of cortical responses are then obtained by interlacing dynamic, multimodal representation learning with whole-brain activation regression in an end-to-end fashion. Using these rich stimulus descriptions, we demonstrated a widespread predictability map across the cortex that covers a large portion (~83%) of the stimulus-driven cortex (Fig. 3, C and E), including association and some frontal regions. While ISCs in these regions are frequently reported (*12*, *41*), suggesting their involvement in stimulus-driven processing, response predictability in these areas had remained elusive so far. Furthermore, the cortical predictivity is maintained even as we compare model predictions against neural responses of held-out subjects (figs. S8 and S10), suggesting that the proposed models are capable of successfully capturing the shared or stimulus-driven response component. These results provide compelling evidence that DNNs trained end to end can learn to capture the complex computations underlying sensory perception of real-life, continuous stimuli.

We further demonstrated that encoding models can form an alternative framework for probing the time scales of different brain regions. While primary auditory and auditory belt cortex (comprising A1, PBelt, LBelt, and Mbelt) as well as the ventral visual stream benefit only marginally from temporal information, there is a remarkable improvement in prediction performance in auditory and visual association and prefrontal cortices, most notably in superior temporal lobe, visuomotor regions within the dorsal stream such as V6A, temporal parietal occipital junction, and inferior frontal regions. The improvement in prediction performance with the 20-s input is consistently seen for both unimodal and multimodal models. It is important to acknowledge that directly comparing the prediction accuracies of static (1-s) and recurrent (20-s) models to infer processing time scales of different brain regions has its limitations. First, this analysis can be confounded by the slow hemodynamic response as performance improvement may be driven in part by the slow and/or spatially varying dynamics. On the basis of our analysis with ROI-level encoding models, the latter seems like a less plausible explanation (fig. S2; see the Supplementary Materials for details). Furthermore, we performed additional analyses to understand the relationship between performance improvement in individual voxels and their autocorrelation properties and found a strong correspondence between the two, suggesting that the distribution of performance improvement across the cortex broadly agrees well with processing time scales (fig. S6; see the Supplementary Materials for details).

Predictions from long–time scale models are based on temporal history as provided in stimulus sequences and not just the instantaneous input. Modeling dynamics within these sequences appropriately is crucial to probe effects of temporal accumulation. RNNs have internal memories that capture long-term temporal dependencies relevant for the prediction task, in this case encoding brain response, while discarding task-irrelevant content. We compare this modeling choice against a regularized regression approach on stimulus features concatenated within T-second clips, with T ranging between 1 and 20 (fig. S4; see the Supplementary Materials for details). The inferior performance compared with our proposed models as well as a nonincreasing performance trend against T for these linear models indicate that accumulation of temporal information by simply concatenating stimulus features over longer temporal windows is insufficient; rather, models that can efficiently store and access information over longer spans, such as RNNs with sophisticated gating mechanisms, are much more suitable for modeling neural computations that unfold over time.

Because activations of units within RNNs depend not only on the incoming stimulus but also on the current state of the network as influenced by past stimuli, they are capable of holding short-term events into memory. Adding the RNN module can thus be viewed as augmenting the encoding models with working memory.

Investigating time scales of representations across brain regions by understanding the influence of contextual representations on language processing in the brain, as captured by LSTM language models for instance, has become a major research focus recently (*42*). In these language encoding models for fMRI, past context has been shown to be beneficial in neural response prediction, surpassing word embedding models. However, models that explain neural responses under dynamic natural vision while exploiting the rich temporal context have not yet been rigorously explored with human fMRI datasets. In a previous study with awake mice, recurrent processing was shown to be useful in modeling the spiking activity of V1 neurons in response to natural videos (*43*). In dynamic continuous visual stimulation fMRI paradigms, a common practice is to concatenate multiple delayed copies of the stimulus to model the hemodynamic response function as a linear finite impulse response (FIR) function (*38*). However, because the feature dimensionality scales linearly with time steps, this approach is limited to HRF modeling and is not feasible to capture longer dynamics of the order of tens of seconds. Another approach is to use features from neural networks trained on video tasks, such as action recognition (*6*). However, these encoding models are constrained to capture one aspect of dynamic visual scenes and are likely useful to predict neural responses in highly localized brain regions. Most studies in visual encoding remain limited to static stimuli and evoked responses in relatively small cortical populations.

Our brain has evolved to process "natural" images and sounds. Recent evidence has shown that sensory systems are intrinsically more attuned to features of naturalistic stimuli, and such stimuli can induce stronger neural responses than task-based stimuli (*44*). Here, we demonstrate that encoding models trained with naturalistic data are not limited to modeling responses of their constrained stimuli set. Instead, by learning high-level concepts of sensory processing, these models can also generalize to out-of-domain data and replicate results of alternate task-bound paradigms. While our models were trained on complex and cluttered movie scenes, we tested their ability to predict response to relatively simple stimuli from the HCP task battery, such as faces and scenes (Fig. 6). The remarkable similarity between the predicted and measured contrasts in all cases suggests that "synthetic" brain voxels, predicted by the trained DNNs, correspond well with the target voxels they were trained to model. We thus provide evidence that these encoding models are capsulizing stimulus-to-brain relationships extending beyond the experimental world they were trained in. On the other hand, classical fMRI experiments, for instance, task contrasts, do not generalize outside the experimental circumstance they were based on. This preliminary evidence suggests that encoding models can serve as promising alternatives for circumventing the use of contrast conditions to study hypotheses regarding the functional specialization of different brain regions. Embedded knowledge within these descriptive models of the brain could also be harnessed in other applications, such as independent neural population control by optimally synthesizing stimuli to elicit a desired neural activation pattern (*45*).

With purely data-driven exploration of fMRI recordings under a hypothesis-free naturalistic experiment, our models replicate the results of previous neuroimaging studies operating under controlled task-based regimes. Our analysis lends support to existing theories of perception that suggest that primary sensory cortices build representations at short time scales and lead up to multimodal representations in posterior portions of STS (*25*). Encoding performance in these regions is consistently improved with longer time scales and multisensory information. We reasoned that regions that are sensitive to multimodal signals and/or longer stimulus dynamics could be distinguished by interrogating the performance of these models on unseen data. To date, encoding models have been rarely used in this manner to assess integration time scales or sensory sensitivity of different brain regions. Classically, processing time scales have been probed using various empirical strategies, for example, by observing activity decay over brief stimulus presentations or by comparing autocorrelation characteristics of resting-state and stimulus-evoked activity (*46*). Furthermore, multisensory regions are identified via carefully constructed experiments with unimodal and multimodal stimulus presentations, followed by analysis of interaction effects using statistical approaches (*27*). Here, we suggest that encoding models can form an alternate framework to reveal clues into these functional properties that can be rigorously validated with future investigation. As with interpreting the results of any predictive model, one should, however, proceed with caution. Sounds are generated by events; this implies that sound representations implicitly convey information about actions that generated them. Similarly, visual imagery provides clues into auditory characteristics, such as the presence or absence of speech. Thus, it is difficult to completely disentangle the individual contributions of auditory and visual features to prediction performance across cortical regions. Similarly, longer time scale inputs can lead to a more robust estimate of the momentary sensory signal, potentially confounding the interpretations of TRWs. Furthermore, scanner noise can affect changes in BOLD signals across the auditory cortex, and several studies have reported that the phenomenon is exacerbated at high field strengths (*47*). Brain function requires additional attentional resources and increased listening effort under the presence of scanner noise, and this may affect the processing of visual input as well, for example, by affecting fixation locations and/or prioritizing attentional deployment for auditory stimuli. Scanner noise can also reduce the sensitivity to stimuli of interest ("movies") by causing non–stimuli-associated activations across the auditory cortex that may interfere in nontrivial ways with stimuli-induced activations. We do not expect this to affect the prediction correlations since the influence of scanner noise is expected to be independent of the stimuli characteristics; nonetheless, this is an important caveat of the proposed audio-based encoding models that hinders their ability in explaining neural responses outside the scanner. Here, notwithstanding the limitations, we contend that these models can, nonetheless, serve as powerful hypothesis generation tools.

The methodological innovations in this study must also be considered in light of their limitations. Because of high dimensionality of features in early layers of the ResNet-50 architecture for high-dimensional visual inputs, we use pooling operations on these feature maps. Thus, low-level visual features, such as orientations, are compromised. The consequent unfavorable outcome is a low predictive performance in V1. Because of a limited computational and memory budget, we could not experiment with fine-tuning the visual subnetwork in this study; in the future, with large-scale collection of naturalistic fMRI datasets that represent a more extensive sampling of the stimulus space, we anticipate that data-fitted or fine-tuned models may surpass the baseline established by pretrained goal-driven networks and may enable us to inch closer to a complete model of the human visual cortex. These models might even provide inspiration in the form of inductive biases or regularization for representation learning on diverse perceptual tasks (*48*). Furthermore, because different subjects can focus on different parts of the stimulus, group-level models can also blur out the precise object orientation information. This is particularly relevant for complex naturalistic stimuli such as movies. In the future, incorporating eye gaze data into these models can be an interesting exploration. Furthermore, because of computational constraints, the proposed model is only able to examine the effects of stimuli up to 20 s in the past. However, previous research with naturalistic stimuli has shown that some brain regions maintain memory of the order of minutes during naturalistic viewing (*49*). Existing evidence also suggests that neural activity is structured into semantically meaningful and coherent events (*25*). Capturing long-range context in encoding models can be a challenging, yet fruitful endeavor yielding potentially novel insights into memory formation.

There are also inherent differences between proposed neural network models and biological networks. DNNs fail to capture known properties of biological networks such as local recurrence; however, they have been found to be useful for modeling neural activity across different sensory systems. At present, feed-forward DNNs trained on recognition tasks constitute the best predictors of sensory cortical activations in both humans and nonhuman primates (*2*). In light of this observation, a recent study proposed that very deep feed-forward only CNNs (for example, ResNet-50 as used in this study for visual feature extraction) might implicitly be approximating "unrolled" versions of recurrent computations of the ventral visual stream (*50*). Object recognition studies on nonhuman primates have also hinted at a functional correspondence between recurrence and deep nonlinear transformations (*51*). Although the functional significance of intraregional recurrent circuits in core object recognition is still under debate, mounting evidence suggests they may be subserving recognition under challenging conditions (*51*, *52*). Thus, investigation of more neurobiologically plausible models of the cortex that innately model intraregional recurrent computations should be explored in the future, especially in relation to their role in visual recognition.

While the present study focuses on shared stimulus-driven brain signals across a subject population, the quest to understand interindividual variability in neural responses remains an important direction forward that promises exciting scientific discoveries linking brain activity to behavior and novel clinical applications (*53*). Over the last decade, these interindividual differences have been shown to result from differences in attentional control and engagement (*54*) and, perhaps, from differences in interpretation (*55*), emotional valence/arousal (*56*), as well as intrinsic individual traits and behavior (*57*). The use of studying interindividual variability under naturalistic stimulation from a clinical perspective has already been highlighted in several prior studies using variants of the ISC framework that attempt to extricate the stable and idiosyncratic stimulus-evoked response component in subjects with neuropsychiatric disorders from the shared stimulus-driven response that is consistent across a control subject population (*58*). In the future, we expect that this direction of using naturalistic paradigms to study interindividual differences will complement the approach of encoding models (at single-subject or group level) in understanding how the brain processes sensory signals from its complex environment and how individual differences in this processing are linked to individual behaviors.

Comprehensive descriptive models of the brain need comprehensive accounts of the stimulus. In this study, using a novel group-level encoding framework, we showed that "reliable" cortical responses to naturalistic stimuli can be accurately predicted across large areas of the cortex using multisensory information over longer time scales. Because our models were trained on a large-scale, multisubject, and open-source dataset, we believe these results could provide an important point of reference against which encoding models for naturalistic stimuli can be assayed in the future. The continued interplay of artificial neural networks and neuroscience can pave the way for several exciting discoveries, bringing us one step closer to understanding the neural code of perception under realistic conditions.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/22/eabe7547/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. G. Varoquaux, R. A. Poldrack, Predictive models avoid excessive reductionism in cognitive neuroimaging. *Curr. Opin. Neurobiol.* **55**, 1–6 (2019).
2. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
3. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
4. H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, Z. Liu, Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* **28**, 4136–4160 (2018).
5. U. Güçlü, M. A. J. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
6. U. Güçlü, M. A. J. van Gerven, Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* **145**, 329–336 (2017).
7. A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
8. A. J. King, G. A. Calvert, Multisensory integration: Perceptual grouping by eye and ear. *Curr. Biol.* **11**, R322–R325 (2001).
9. J. Driver, T. Noesselt, Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron* **57**, 11–23 (2008).
10. J. Miller, Divided attention: Evidence for coactivation with redundant signals. *Cogn. Psychol.* **14**, 247–279 (1982).
11. S. Sonkusare, M. Breakspear, C. Guo, Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
12. U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
13. M. Schönwiesner, R. J. Zatorre, Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 14611–14616 (2009).
14. D. Schwartz, M. Toneva, L. Wehbe, Inducing brain-relevant bias in natural language processing models, in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 8 to 14 December 2019.
15. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, M. Jenkinson; WU-Minn HCP

Consortium, The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).

16. A. T. Vu, K. Jamison, M. F. Glasser, S. M. Smith, T. Coalson, S. Moeller, E. J. Auerbach, K. Uğurbil, E. Yacoub, Tradeoffs in pushing the spatial resolution of fMRI for the 7T Human Connectome Project. *Neuroimage* **154**, 23–32 (2017).
17. S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. W. Wilson, CNN architectures for large-scale audio classification, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 131–135.
18. A. S. Bregman, *Auditory Scene Analysis* (MIT Press, 2001).
19. T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 936–944.
20. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 770–778.
21. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical image database, in *2009 Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
22. S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, YouTube-8M: A large-scale video classification benchmark. arXiv:1609.08675 (2016).
23. M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, D. C. Van Essen, A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
24. U. Hasson, E. Yang, I. Vallines, D. J. Heeger, N. Rubin, A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
25. C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, K. A. Norman, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
26. M. A. Goodale, A. D. Milner, Separate visual pathways for perception and action. *Trends Neurosci.* **15**, 20–25 (1992).
27. G. A. Calvert, Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cereb. Cortex* **11**, 1110–1123 (2001).
28. T. Raij, K. Uutela, R. Hari, Audiovisual integration of letters in the human brain. *Neuron* **28**, 617–625 (2000).
29. M. S. Beauchamp, Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* **3**, 93–113 (2005).
30. M. S. Beauchamp, B. D. Argall, J. Bodurka, J. H. Duyn, A. Martin, Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nat. Neurosci.* **7**, 1190–1192 (2004).
31. G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. R. Williams, P. K. McGuire, P. W. R. Woodruff, S. D. Iversen, A. S. David, Activation of auditory cortex during silent lipreading. *Science* **276**, 593–596 (1997).
32. N. Kanwisher, G. Yovel, The fusiform face area: A cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2109–2128 (2006).
33. I. Tavor, M. Yablonski, A. Mezer, S. Rom, Y. Assaf, G. Yovel, Separate parts of occipito-temporal white matter fibers are associated with recognition of faces and places. *Neuroimage* **86**, 123–130 (2014).
34. S. Nasr, N. Liu, K. J. Devaney, X. Yue, R. Rajimehr, L. G. Ungerleider, R. B. H. Tootell, Scene-selective cortical regions in human and nonhuman primates. *J. Neurosci.* **31**, 13771–13785 (2011).
35. J. A. Frost, J. R. Binder, J. A. Springer, T. A. Hammeke, P. S. Bellgowan, S. M. Rao, R. W. Cox, Language processing is strongly left lateralized in both sexes. Evidence from functional MRI. *Brain* **122** (Pt. 2), 199–208 (1999).
36. P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312 (2000).
37. T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, T. D. Wager, Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
38. A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
39. Y. Cao, C. Summerfield, H. Park, B. L. Giordano, C. Kayser, Causal inference in the multisensory brain. *Neuron* **102**, 1076–1087.e8 (2019).
40. S. M. Wilson, I. Molnar-Szakacs, M. Iacoboni, Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cereb. Cortex* **18**, 230–242 (2008).
41. I. P. Jääskeläinen, K. Koskentalo, M. H. Balk, T. Autti, J. Kauramäki, C. Pomren, M. Sams, Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *Open Neuroimag. J.* **2**, 14–19 (2008).
42. S. Jain, A. Huth, Incorporating context into language encoding models for fMRI, in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, Canada, 3 to 8 December 2018.
43. F. H. Sinz, A. S. Ecker, P. G. Fahey, E. Y. Walker, E. Cobos, E. Froudarakis, D. Yatsenko, X. Pitkow, J. Reimer, A. S. Tolias, Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *bioRxiv*, 452672 (2018).
44. J. Schultz, K. S. Pilz, Natural facial motion enhances cortical responses to faces. *Exp. Brain Res.* **194**, 465–475 (2009).
45. P. Bashivan, K. Kar, J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
46. J. Chen, U. Hasson, C. J. Honey, Processing timescales as an organizing principle for primate cortex. *Neuron* **88**, 244–246 (2015).
47. J. E. Peelle, Methodological challenges and solutions in auditory functional magnetic resonance imaging. *Front. Neurosci.* **8**, 253 (2014).
48. F. H. Sinz, X. Pitkow, J. Reimer, M. Bethge, A. S. Tolias, Engineering a less artificial intelligence. *Neuron* **103**, 967–979 (2019).
49. U. Hasson, J. Chen, C. J. Honey, Hierarchical process memory: Memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
50. Q. Liao, T. A. Poggio, Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv:1604.03640 (2016).
51. K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, J. J. DiCarlo, Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
52. D. Wyatte, D. J. Jilk, R. C. O'Reilly, Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.* **5**, 674 (2014).
53. E. S. Finn, E. Glerean, A. Y. Khojandi, D. Nielson, P. J. Molfese, D. A. Handwerker, P. A. Bandettini, Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *Neuroimage* **215**, 116828 (2020).
54. J. J. Ki, S. P. Kelly, L. C. Parra, Attention strongly modulates reliability of neural responses to naturalistic narrative stimuli. *J. Neurosci.* **36**, 3092–3101 (2016).
55. M. Nguyen, T. Vanderwal, U. Hasson, Shared understanding of narratives is correlated with shared neural responses. *Neuroimage* **184**, 161–170 (2019).
56. L. Nummenmaa, E. Glerean, M. Viinikainen, I. P. Jääskeläinen, R. Hari, M. Sams, Emotions promote social interaction by synchronizing brain activity across individuals. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 9599–9604 (2012).
57. E. S. Finn, P. R. Corlett, G. Chen, P. A. Bandettini, R. T. Constable, Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nat. Commun.* **9**, 2043 (2018).
58. Z. Yang, J. Wu, L. Xu, Z. Deng, Y. Tang, J. Gao, Y. Hu, Y. Zhang, S. Qin, C. Li, J. Wang, Individualized psychiatric imaging based on inter-subject neural synchronization in movie watching. *Neuroimage* **216**, 116227 (2020).
59. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* **57**, 289–300 (1995).
60. A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **60**, 101027 (2020).
61. K. J. Piczak, *ESC: Dataset for Environmental Sound Classification* (Harvard Dataverse, 2015).
62. J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320–341 (2014).

Acknowledgments

Funding: This work was supported by NIH grants R01LM012719 (M.R.S.), R01AG053949 (M.R.S.), R21NS10463401 (A.K.), R01NS10264601A1 (A.K.), the NSF NeuroNex grant 1707312 (M.R.S.), the NSF CAREER 1748377 grant (M.R.S.), and Anna-Maria and Stephen Kellen Foundation Junior Faculty Fellowship (A.K.). **Author contributions:** M.K. and M.R.S. conceptualized the work. M.K., M.R.S., A.K., G.H.N., and K.J. gave input on analyses. M.K. and G.H.N. performed the main analyses. K.J. gave input on data preprocessing and data considerations. All authors contributed to the drafting and revision of the main manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All experiments in this study are based on the open-access Human Connectome Project movie-watching database. The dataset is publicly available for download through the HCP website (https://humanconnectome.org/). Throughout this study, we used 7T fMRI data from the "Movie Task fMRI 1.6 mm/59 k FIX-Denoised" package within the HCP dataset. The network implementation and analysis codes are available at https://github.com/mk2299/MultimodalEncoding/. Additional data related to this paper may be requested from the authors.

Submitted 11 September 2020
Accepted 12 April 2021
Published 28 May 2021
10.1126/sciadv.abe7547

Citation: M. Khosla, G. H. Ngo, K. Jamison, A. Kuceyeski, M. R. Sabuncu, Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Sci. Adv.* **7**, eabe7547 (2021).

Science Advances

Cortical response to naturalistic stimuli is largely predictable with deep neural networks

Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu

Sci Adv **7** (22), eabe7547.
DOI: 10.1126/sciadv.abe7547

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.