Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/7/22/eabe7547/DC1

Supplementary Materials for

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

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

*Corresponding author. Email: msabuncu@cornell.edu

Published 28 May 2021, *Sci. Adv.* **7**, eabe7547 (2021)
DOI: 10.1126/sciadv.abe7547

This PDF file includes:

HCP Movies

Table S1 summarizes the HCP movie-watching dataset split used for training and evaluating all models.

Table S1. HCP dataset split

Movie	Split	Stimulus-response pairs per subject
7T_MOVIE1_CC1_v2	Training/Validation	652
7T_MOVIE2_HO1_v2	Training/Validation	716
7T_MOVIE3_CC2_v2	Training/Validation	669
7T_MOVIE4_HO2_v2	Testing	699

Region of Interest (ROI) selection

ROIs were selected for each analysis based on the descriptions provided in the neuroanatomical supplementary results of the HCP MMP parcellation and an extensive literature review. For Figure 2 in the main text and Figure S9, ROIs were thus assigned to groups 1-5 according to Table S2.

Table S2. ROI categorization

Group	ROIs
1. Auditory	A1, LBelt, PBelt, MBelt, RI, STSda, STSva, A4, A5, TA2
2. Visual	V1, V2, V3, V3A, V3B, V3CD, V4, V4t, V6, V6A, V7, V8, DVT, LO1-3, PIT, FFC, VMV1-3, IPS1, MT, VVC
3. Multi-sensory + sensory bridges	STSdp, STSvp, STGa, STV, TPOJ1-3
4. Language	55b, SFL, PSL, 44, 45
5. Frontal	IFSa, IFSp, IFJa, IFJp, FEF

Dorsal and ventral visual stream ROIs as well as early and association auditory cortex ROIs in Figure 4 (main text) were derived from the explicit stream segregation and categorization described in the HCP MMP parcellation and are defined here for quick reference.

- Dorsal: V3A, V3B, V6, V6A, V7, IPS1
- Ventral: V8, VVC, PIT, FFC, VMV1-3
- MT+: MT, MST, V4t, FST
- Early auditory: A1, PBelt, MBelt, RBelt, RI
- Association auditory: A4, A5, TA2, STGa, STSdp, STSda, STSvp, STSva

All ROIs are shown in Figure S1.

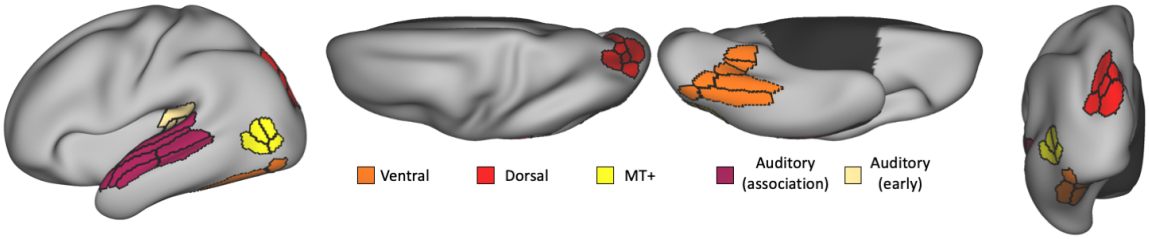

Fig. S1. Group segregation from the HCP MMP parcellation.

Estimating BOLD response delay

BOLD response delay was estimated using ROI-level encoding models due to their faster iteration times in comparison to voxel-wise encoding. The input to these models was the preprocessed stimuli as described for voxel-wise encoding with the same train-validation-test split, and the output was the evoked ROI-level fMRI response at different lags (1-7 seconds) from the stimulus. Thus, the output is a 360-D vector corresponding to the mean fMRI response in each ROI of the HCP MMP parcellation. The feature extractors were identical to those in the proposed voxel-wise auditory and visual models. However, instead of a convolutional response model, here, the response model comprised two fully connected layers with output dimensions of 512 and 360 with an exponential linear unit and linear activation respectively. All models were trained for 20 epochs with a batch size of 4 and a learning rate of 1e-4. Validation curves were monitored to ensure convergence. Prediction accuracy of each model was computed as

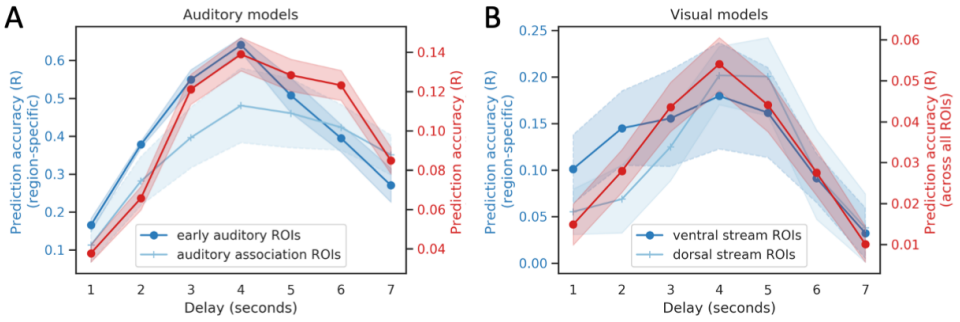

Fig. S2. ROI-based encoding performance for estimating delay. (A) depicts the estimated mean and standard error of the prediction accuracy (R) across various delays (1-7s) within the early auditory and association auditory group (blue) as well as across all ROIs (red), as obtained using the single epoch (1s) auditory model. (B) depicts the estimated mean and standard error of the prediction accuracy (R) for various delays (1-7s) within the primary and dorsal visual streams (blue) as well as across all ROIs (red), as obtained using the single frame visual model. Shaded regions depict the standard error in estimating mean across ROIs within each group. ROI categorization is described in the sub-section on ROI selection.

the mean Pearson correlation coefficient between the predicted and measured response across all ROIs, in the held-out movie dataset. Based on Figure S2, we estimated a response delay of 4 seconds, as this lag yielded the maximum prediction accuracy across all ROIs for both

auditory and visual ROI-level models. Further, even while restricting the prediction accuracy (R) to ROIs within different cortical areas (such as the early/association auditory areas or the dorsal/ventral visual stream), the optimal lag was consistently 4 seconds, suggesting that the difference in performance of 1-sec and 20-sec models in these regions (Figure 4) is not largely driven by differences in the hemodynamic response function (HRF).

Defining the stimulus-driven or "synchronous" cortex

We isolated voxels involved in stimulus-driven processing, termed "synchronous" or "stimulus-driven" voxels, by computing mean inter-group correlations over all training movies. Inter-group correlations were computed by splitting the entire group of subjects into two halves and computing correlations between the mean response time-course of each half (comprising 79 subjects) at every voxel. We employed a liberal threshold of 0.15 for this correlation value. Thus, the mask of "stimulus-driven" voxels included those voxels that achieved an inter-group correlation of 0.15 or above. We computed mean quantitative metrics over this mask in Figure 3E (main text) to compare different models.

Model architectures and implementation

The base feature extraction networks and convolutional response model in Figure 1 had the architecture as detailed in Figure S3. The feature extraction networks are reminiscent of the feature pyramid network, which has shown significant improvements as a generic feature extractor across various applications. These networks comprise a parallel top-down pathway with lateral connections which grants them the ability to characterize both "what" and "where" in cluttered scenes, thereby enhancing object detection. We note that similar models with top-down and skip connections have been popular in vision research, since they can enrich low-level features with high-level semantics. The output of the feature extractor is fed into the convolutional response model to predict the evoked fMRI activation. This enables us to train both components of the network simultaneously in an end-to-end manner. Since the output response is differentiable with respect to network weights, the weights are adjusted via a first-order gradient-based optimization method to minimize the *mean squared error* between the predicted and target activation values across the entire brain.

For ResNet-50, we use activations of the last residual block of each stage, namely, res2, res3, res4 and res5 to construct our stimulus descriptions $\mathbf{s}$. From the VGG-ish network, we use the activations of each convolutional block, namely, conv2, conv3, conv4 and the penultimate dense layer fc2 (Pre-trained tensorflow/keras models for the visual and auditory backbone were available at `https://keras.io/applications` and `https://github.com/tensorflow/models/tree/master/research/audioset/vggish`, respectively). The first three sets of activations are refined through a top-down path to enhance their semantic content, while the last activation is concatenated into $\mathbf{s}$ directly (res4 activations are vectorized using global average pool). The top-down path comprises three feature maps at different res-

olutions with an up-sampling factor of 2 successively from the deepest layer of the bottom-up path. Each such feature map comprising 256/128 channels (in visual/auditory models respectively) is merged with the corresponding feature map in the bottom-up path (reduced to 256/128 channels by 1x1 convolutions) by element-wise addition. Subsequently, the feature map at each resolution is collapsed into a 256/128-dimensional feature vector through a global average pool operation and concatenated into $\mathbf{s}$, leading to a 1024-D and 512-D feature representation for the visual and auditory stimuli respectively. The aggregated features are then passed onto a CNN comprising the following feedforward computations: a fully connected layer to map the features into a vector space which is reshaped into a 1024-channel cuboid of size 6x7x6 followed by four 3x3x3 transposed convolutions (conv.T) with a stride of 2 and exponential linear unit activation function to up-sample the latter. Each convolution reduces the channel count by half with the exception of the last convolution which outputs the single-channel predicted fMRI response.

The 20-second models additionally comprised an LSTM layer to model the temporal propagation of features across the contiguous sequence of input frames and/or spectrograms. The LSTM module has driven success across varied sequence modeling tasks due to its ability to efficiently regulate the flow of information across cells through gating. The memory cell in LSTM is modulated by three gates, namely, the input, forget and output gates. We note that the LSTM layer did not change the dimensionality of the input features so that equitable comparisons can be made against 1-sec models. The Audiovisual-1sec model concatenated features obtained from the base visual (1024-D) and audio (512-D) feature extraction networks, reduced their combined dimensionality to the higher value among the two (1024-D) by passing through a bottleneck dense layer followed by the same convolutional response model. The Audiovisual-20sec model additionally incorporated modality-specific LSTM networks prior to feature concatenation.

Implementation:
We note that all 6 models have roughly the same order of trainable parameters in the range of 242M-362M. All parameters were optimized using Adam with a learning rate of 1e-4. Auditory and visual models were trained for 50 epochs with unit batch size. The stimulus as well as subject whose fMRI response is used as the target in the loss ("mean squared error") are randomly sampled over each step of the training but kept consistent across models. We found this method to work better than using the group-averaged response as target, presumably because this sampling provides information about both the cross-subject mean and the variance of response. Given the noise characteristics at each voxel, we hypothesize that this enables the model to focus on regions that can be well predicted with the given stimulus. Validation curves were monitored for all models to ensure convergence.

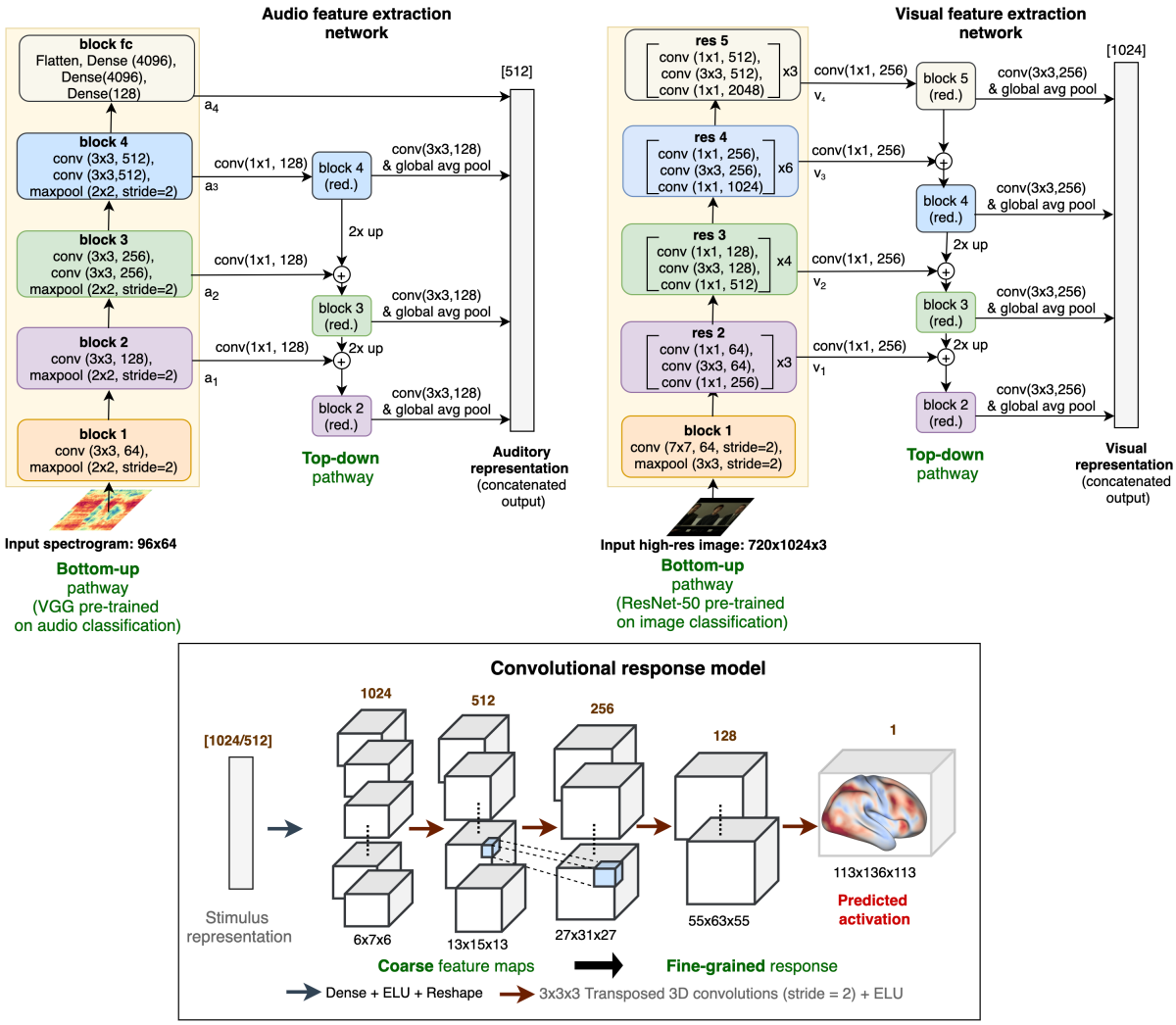

Fig. S3. Implementation details for the audio (top left) and visual (top right) feature extraction networks as well as the convolutional response model (bottom). All layers and blocks outside the yellow rectangle (bottom-up pathway) are trained from scratch. The blocks inside the yellow rectangular window are initialized with networks pre-trained on image or sound recognition. Further, ResNet-50 is frozen during the training of all encoding models, whereas VGG is fine-tuned. The sequence of operations within each block are defined from top to bottom, while the number of repetitions for each sequence within the block are indicated with the multiplicative symbol on the right.

Regularized linear regression: deep convolutional features

We also trained group-level encoding models using a linear response model since this constitutes the dominant state-of-the-art approach to neural encoding. To enable a fair comparison against the proposed 1-sec uni-modal models, we extract hierarchical features from the same layers of the ResNet-50 and VGG-ish architectures as employed by the proposed models. The only difference here is the lack of a top-down pathway (since it is not a part of the pre-trained network but is trained with random initialization on the neural response prediction task), which prevents the refinement of coarse feature maps before aggregation. Pooling the outputs of different layers channel-wise using the global average pooling operation (namely $\{v_1, v_2, v_3, v_4\}$ for the visual model and $\{a_1, a_2, a_3, a_4\}$ for the audio model in Figure S3) leaves us with and 1024 and 3840 features to present to the auditory and visual models, respectively. Further, to compare against the longer-duration 20-sec models, we adopted two approaches: (1) we simply concatenated the stimulus features extracted for each second (as described above) over T-second windows with T ranging from 1 to 20 seconds and presented these aggregated features to the linear response model; alternatively, (2) we reduced the dimensionality of the aggregated features to a fixed length (set to 128) as in (1) using principal component analysis run on the training data. We added this comparison to rule out the fact that the temporal trend in performance of linear models is simply driven by a higher-dimensional feature space. We note that even after dimensionality reduction, the components retained at least 80% of the explained variance in all cases. Audio-visual encodings with linear response models were obtained similarly by simply fusing the respective audio and visual hierarchical features through concatenation before linear regression. We apply l_2 regularization on the regression coefficients and adjust the optimal strength of this penalty through cross-validation on the training data using log-spaced values in $\{1\mathrm{e}{-14}, 1\mathrm{e}14\}$ for each model. We report performance of the best models in Figure S4(A). Note that unlike the WordNet models, we found that optimizing a single regularization penalty α common across all voxels outperformed independent voxel-wise fitting with bootstrap in this case. Thus, we only present the results for the former. We note here that the convolutional response model in our proposed approach (instead of a fully-connected approach) allowed us to keep the learnable parameters manageable, facilitating joint optimization/fine-tuning of the feature extractor and response models. The consistently superior performance of the proposed models against linear regression-based approaches strongly suggests that there is merit in end-to-end learning for encoding responses to dynamic, multi-sensory stimuli.

Regularized linear regression: WordNet features

Another popular approach in voxel-wise forward encoding beyond primary sensory cortices is the semantic category encoding model that is based on high-level semantic features [38]. This approach relies on labels that indicate the presence of semantic object and action categories in each movie frame. In this analysis, we employed WordNet labels that were provided as part of the HCP movie-watching data pipeline. The semantic labels were manually assigned by the Gallant lab team using the WordNet semantic taxonomy and subsequently converted to Word-

Net synsets to build an 859-D semantic representational space (corresponding to 859 WordNet synset names). Following [38], we fitted l_2 regularized linear regression models (known as ridge regression) to find weights corresponding to different input features for every voxel. The regularization parameter, α was optimized independently for each voxel by testing among 10 log-space values in [1, 1000]. The optimal alpha is obtained by averaging across 15 bootstrapped held-out sets. In addition to fitting models with WordNet features extracted 4s prior to the measured neural response, we developed longer timescale linear models by concatenating the WordNet features extracted for each second (as described above) over T-second windows with T ranging from 1 to 20 seconds and presented these aggregated features to the bootstrapped regularized regression model. Figure S4 (B) demonstrates the performance of WordNet models across different groups of regions as a function of T, and (C) depicts the voxel-level prediction accuracy (R) of the best performing WordNet model that stacks features from 4-12s (at an interval of 1s) prior to the encoded cortical response. While simple and interpretable, the WordNet models clearly under-perform in terms of prediction accuracy (R) in comparison to the models proposed in the present study.

Ablation study

To determine the influence of different architectural components on prediction performance of the proposed models, we performed an ablation study to investigate the individual contributions of (i) non-linearities in the response model, (ii) hierarchical (multi-scale) feature maps, (iii) fine-tuning audio sub-network (VGG) and (iv) LSTM. We selectively removed each of the components from the respective model and compared the resulting performance against the proposed 1-sec and 20-sec models that employ all (i)-(iii) and (i)-(iv) components respectively. We note that the model without LSTM (iv) uses concatenated features instead of employing recurrence. Due to computational constraints, we could not train a model that feeds 20-sec concatenated features directly to the convolutional response model since this raises the number of parameters substantially. Instead, we map the concatenated feature input to a 1024-D and 512-D feature space for visual and audio models respectively using a fully connected layer. We note that this also ensures a more equitable comparison against the proposed 20-sec models that use LSTMs by enforcing that the representations fed into the response models in both cases are of the same dimension. We follow the same protocol for training these models as used for training the proposed models. There are several interesting observations to make from this ablation analysis (Figure S4D). (i) First, we find that encoding models with a frozen VGG network that is not updated during training incur a loss in performance compared to the proposed model where VGG layers are trainable during neural response prediction. This clearly demonstrates the advantages of altering these pre-trained models and suggests that fine-tuning is both feasible and beneficial in improving neural response prediction. (ii) Next, we find that prediction performance deteriorates after removing the non-linearities in both the Audio-1sec and Visual-1sec models. In the context of the Visual-1sec model with a frozen pre-trained backbone (ResNet-50) and coupled with (i), this observation further highlights that it is possible to

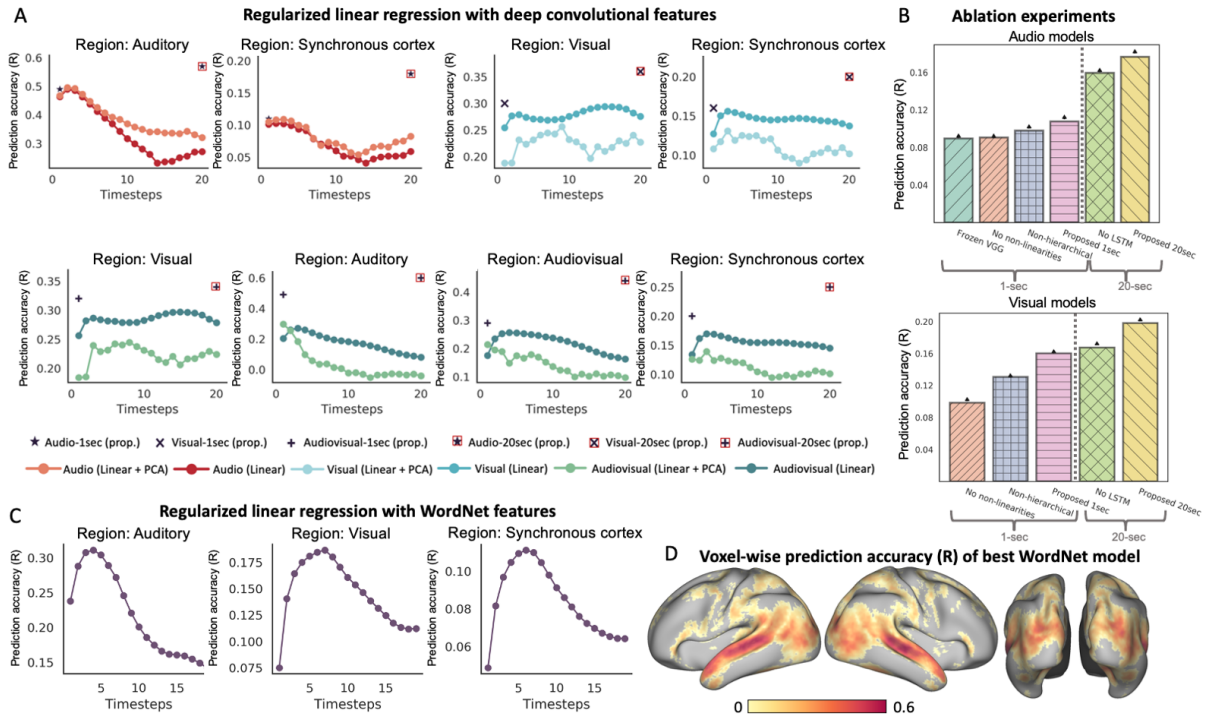

Fig. S4. Performance of linear response models and baselines. (A) shows the region-averaged prediction accuracy of linear response models using deep convolutional features. (B) shows results of the ablation study and highlights the importance of different components of the proposed model architecture. (C) shows the region-averaged prediction accuracy of linear response models using semantically rich WordNet features and (D) shows the cortical map of the prediction accuracy (R) for the best WordNet model. The x-axis in (A) and (C) depicts the length of the windows (in seconds) over which the stimulus features are concatenated and y-axis shows the mean Pearson correlation coefficient between the predicted and measured responses across the stimulus-driven voxels.

develop models of human sensory processing that are quantitatively more precise in matching brain activity than task-driven neural networks. (iii) Next, we assessed the benefit of using hierarchical feature maps over selecting the single best-performing layer for each model (audio or visual) based on cross-validation. For both audio and visual models, we find that features from the last layer (i.e., a_4 and v_4, respectively) yield the highest mean prediction accuracy (R) across the synchronous cortex. However, although the convolutional response model architecture is common across these encoding models, it is important to note that this analysis is still plagued by confounds such as the different dimensionality of feature spaces across different layers that feed into the response model. The best performing single-layer encoding model, however, still performs worse than the hierarchical approach. (iv) Finally, while the encoding models with concatenated features outperform the 1-sec models, the performance still falls short against the accuracy obtained by the proposed 20-sec models employing LSTM. We believe this noticeable difference arises from the ability of LSTMs to efficiently capture long-term dependencies and reconcile the recent input history ('memory') with the immediate context (current frame).

Computing significance estimates

The statistical significance of individual voxel predictions (Figure 3) was computed as the p-value of the obtained sample *correlation coefficient* for the null hypothesis of uncorrelation (i.e., true correlation coefficient is zero) under the assumptions of a bivariate normal distribution. We employed the false-discovery procedure of Benjamini & Hochberg (1995) to control for multiple comparisons under assumptions of dependence. For statistical comparison of model performance within each group of regions in Figure 2 (main text), we performed the paired t-test on ROI-level average performance metrics and corrected for multiple comparisons among models (Bonferroni).

Sensory-sensitivity index

Distorting the input to the audio-visual model at test time allows us to interrogate the sensory-sensitivity of different brain regions. We developed a sensory-sensitivity index of each ROI based upon predictive performance of the model with distorted inputs, as shown in Figure 5. Let SV_r and SA_r denote the mean prediction accuracy of the model in region r after shuffling (temporally) the input order of the visual and auditory stimuli, respectively. The sensory-sensitivity index for region r is then defined as $s_r = \frac{SA_r - SV_r}{SA_r + SV_r}$. Note that positive values of this index indicate that region r incurs a greater loss in predictivity upon distortion of visual information than auditory information, suggesting a higher visual sensitivity for this voxel. Similarly, negative values signal towards a higher auditory sensitivity.

Stimuli for synthetic contrasts

Synthetic contrasts were generated to study the generalization of our models to new experimental paradigms (Figure 6). We focus on predicting task-based contrasts for three semantic

categories, namely, *faces*, *places* and *speech*, since these are the most well-studied categories in the context of their distinct functional signatures. The stimuli for visual contrasts were derived from the HCP Working Memory paradigm, which combines category specific representation tasks (including faces and places) and working memory tasks. After excluding grayscale images, we were left with 102, 77, 97 and 103 images for the categories of faces, places, body parts and tools, respectively. Since these are static image without any dynamic content, we employed the Visual-1sec model to derive the visual contrasts (Figure 6(C),(D)).

Stimuli for the speech and non-speech contrast were extracted from large popular datasets for these categories. Speech stimuli were extracted from a human speech-utterance dataset comprising short audio clips of interviews recorded on YouTube [60]. Non-speech stimuli were extracted from another large dataset comprising short clips of environmental sounds [61]. We randomly extracted ~ 100 minutes of audio waveforms from these datasets for both categories. The stimuli were processed for mel-spectrogram extraction in the same manner as the HCP audio-visual movies. Since the non-speech stimuli only comprised contiguous clips of roughly $3-5$ second duration, we employed the Audio-1sec model to obtain the speech contrast (Figure 6(B)).

Perturbation analysis with 20-sec models

To address the influence of temporal continuity and short-term memory (past inputs) on the predictions of 20-sec models, we conducted a perturbation analysis by distorting the input context seen by these models at inference time using two shuffling experiments:

- Shuffled (different segment): In this experiment, we keep the last frame of every 20-second input segment and replace the preceding 19 frames with contiguous frames of a randomly selected 19-sec input clip within the test movie. This input perturbation thus largely maintains a temporal continuity and highlights the influence of past inputs or short-term memory on response predictions.
- Shuffled (same segment): Under this experimental set-up, we randomly shuffle the first 19 frames of the same input clip at inference time while keeping the last frame the same. This obliterates the temporal continuity of the input clip without changing the overall content that is fed into the encoding model.

We repeated both shuffling experiments 10 times and report the average performance of each model under these two perturbation methods across different ROIs in Figure S5. As can be seen from the figure, both input perturbations cause a drop in model performance, albeit to different degrees. Interestingly, the Audio-20sec model seems to rely on the temporal continuity of the input more heavily than the Visual-20sec model, as evidenced by the much sharper drop in performance for the former model under same segment re-shuffling. The consistent deterioration of model performance under these control experiments is thus another indication that the 20-sec models exploit recent input history ('memory') while computing response predictions.

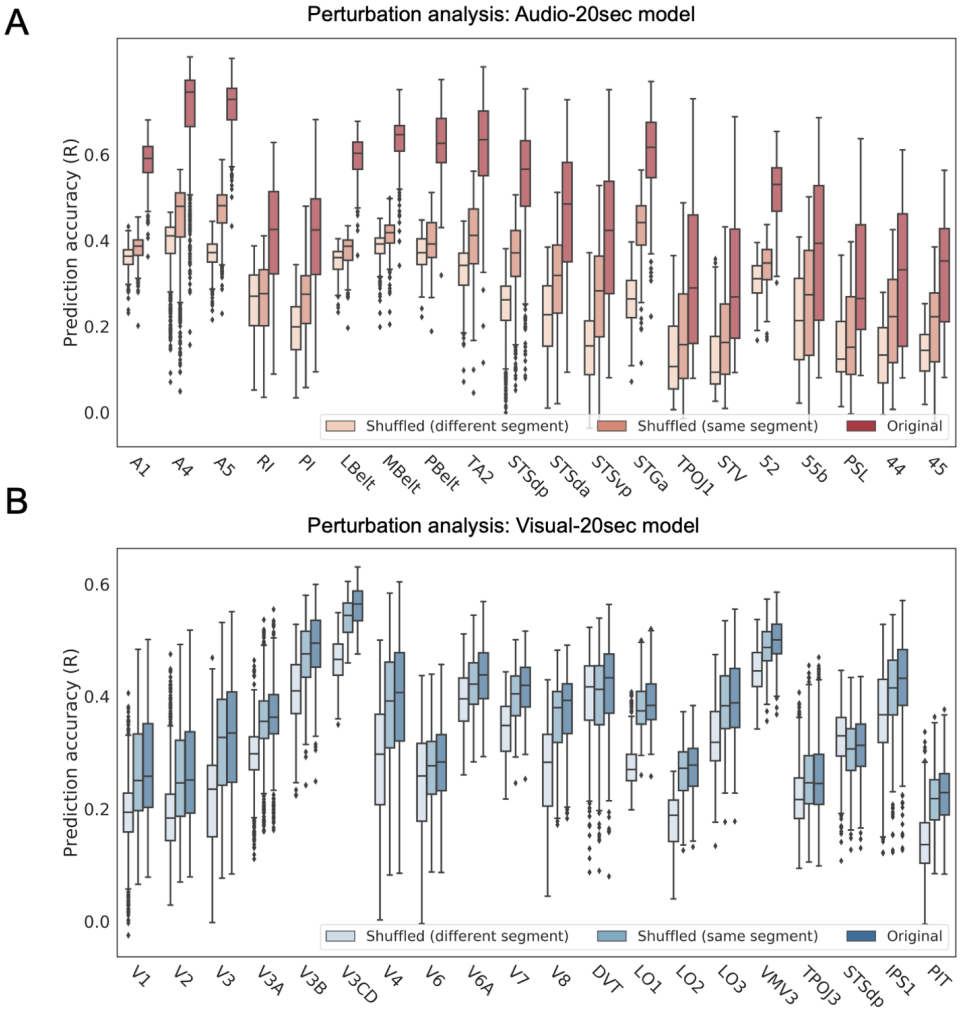

Fig. S5. Perturbation analysis with Audio-20sec (A) and Visual-20sec (B) models. ROI box plots depict the un-normalized correlation coefficients between the predicted and measured response of voxels in each ROI using original or distorted 20-sec input clips at inference time.

Performance improvement and autocorrelation decay

In the past, processing timescales in the brain have been probed using several different means [46]. In one of the proposed approaches, the decay time of temporal autocorrelation is used as a proxy measure to understand the variation in processing timescales across different brain regions. With this approach, it was shown that decay times increased progressively along the temporal hierarchy. Following this line of work, we estimated the autocorrelation decay time constant (π) for each voxel by fitting an exponential, $A\exp\{-t/\pi\}$, to the autocorrelation function (autocorrelation computed at different lags). The exponential model was first independently fit for each movie run and each voxel and the estimated π were subsequently averaged across runs to obtain one decay time constant per voxel. Here, we were primarily interested in understanding whether there is any relationship between the performance improvement of the 20-sec model over 1-sec model, ΔR, computed as the difference between the prediction accuracies of the Audiovisual-20sec and Audiovisual-1sec at every voxel, and the temporal autocorrelation properties of that voxel. We hypothesized that in voxels with longer processing timescales, the autocorrelation would persist for longer durations (resulting in larger π) and the longer timescale model (20-sec) would yield more substantive improvement over the 1-sec model. As shown in Figure S6, we observed a significantly positive correlation between performance improvement and the autocorrelation decay time constant (r = 0.49 and 0.50 across voxels in auditory and visual regions as defined in Table S2), in line with our hypothesis. This suggests that the benefit of employing the 20-sec model, as quantified in terms of performance improvement, is indeed more remarkable in regions with longer processing timescales.

Surface visualization

All input fMRI data, as well as response predictions in this study are volume-based. In order to be consistent with prior research on encoding models that employ surface visualizations, we created surface versions of volumetric predictability and synthetic contrast maps, as shown in Figures 3, 5 and 6. We employed the 3D trilinear mapping method from connectome workbench that computes the result on each vertex based on linear interpolation from voxels on each side of the vertex (`https://www.humanconnectome.org/software/workbench-command`). However, since volume to surface mappings are an approximation, we only employ this conversion for visualizations. All reported metrics are computed on volumes only on a per-voxel basis.

Qualitative analysis

To gain qualitative insights into the predictions of the most accurate model (Audiovisual-20sec) on the held-out movie, we plot the predicted as well as measured response time-series of the voxel with 'median' prediction accuracy (R) in the best performing ROI of each group (Figure S7). The latter corresponds to A4, V3CD, STSdp, IFSp and Area 45 for the auditory, visual, multi-sensory, frontal and language groups respectively.

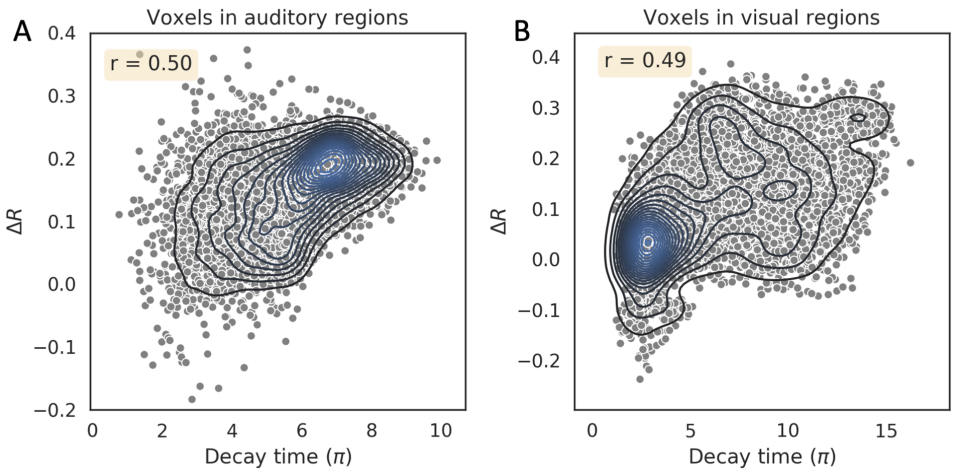

Fig. S6. Performance boost of the 20-sec model over 1-sec model is higher in voxels with longer autocorrelation decay times. (A) & (B) depict the performance improvement (ΔR) against decay time constants for voxels associated with auditory and visual regions, respectively (Table S2). The r value indicates the Pearson correlation coefficient between the two quantities. Each dot in the scatterplot represents an individual voxel. Bivariate kernel density estimates are overlaid on top of the scatterplot as contours to depict the probability distribution of observations.

Group-level prediction accuracy: held-out set

To test the generalizability of the models, we further compared model predictions against the group-averaged response of a held-out group within the HCP dataset comprising 20 novel subjects distinct from the 158 individuals used in the training set, on the same independent held-out movie.

Noise ceiling estimation: For the held-out group, we obtain the noise ceiling by considering variability across subjects. Here, the noise ceiling 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 upper bound achievable by a group-level encoding model. We employ this noise ceiling for comparison against the prediction accuracy of the model on the held-out group of subjects (Figure S8).

The models accurately predicted cortical responses evoked by the *independent* test movie as measured in the *independent* subject population (Figure S8, S9), with the best performing model (Audiovisual-20sec) even achieving close to perfect predictivity relative to the "noise ceiling" in certain multi-sensory sites such as the posterior STS (Figure S8(A), (G)). Here, the noise ceiling was computed as the correlation coefficient between the mean neural response in 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

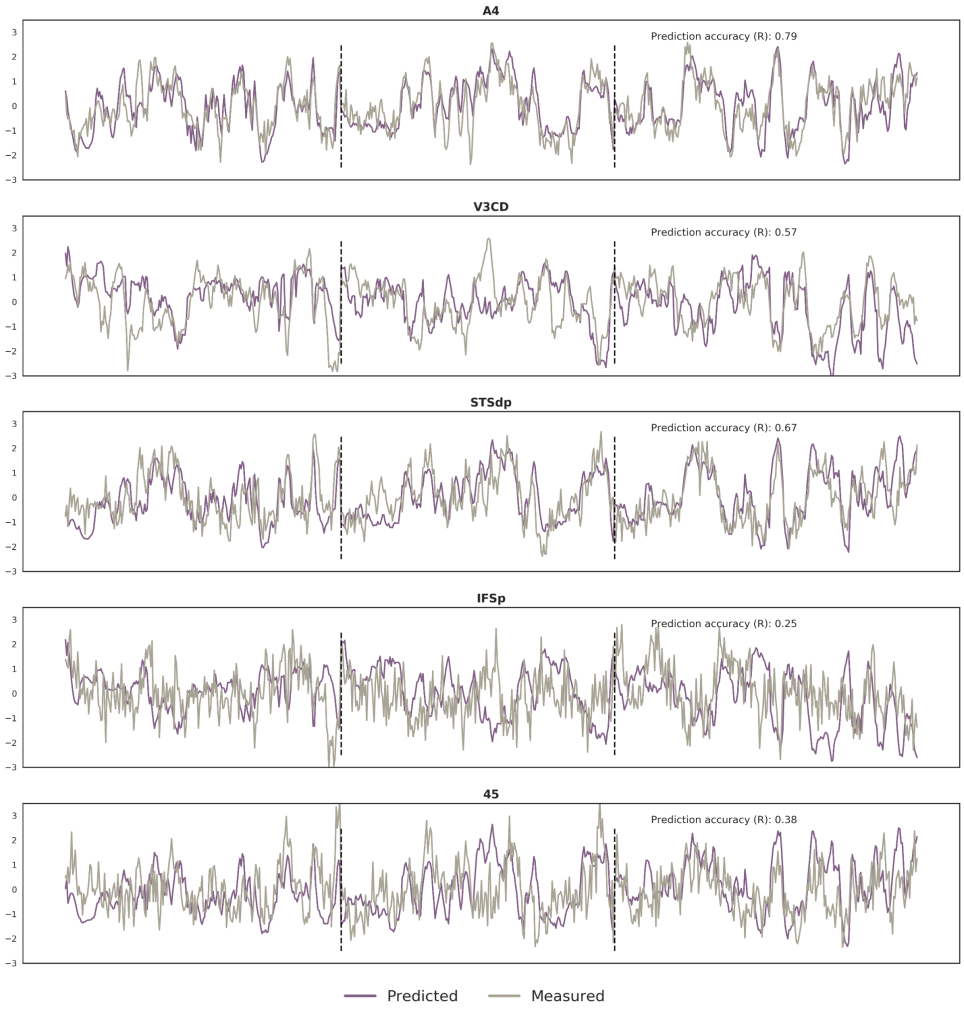

Fig. S7. Predicted and measured response time-series of the 'median' predictive accuracy (R) voxel across ROIs of different functional groups. Vertical dashed lines mark the boundary of clip segments in the held-out movie.

shared across independent subject populations and thus reflects the upper bound achievable by a group-level encoding model. These results clearly indicate that inclusion of temporal history and multi-sensory information pushes the prediction accuracies closer to their upper bound, as also evidenced by a higher slope of the linear model fit on their corresponding data points. Further, voxels that truly approach the noise ceiling are predominantly associated with the auditory group of regions as broadly characterized within the HCP MMP parcellation. Interestingly, we find that this regional distribution of predictivity against noise ceiling holds even for subject-specific responses and not just the group-averaged responses, as described in the next section and shown in Figure S10.

Subject-level prediction accuracy: held-out set

For each participant in our independent subject group (N = 20), we computed the correlation coefficient (R) between the predictions of the best performing model (Audiovisual-20sec) and the subject-specific fMRI response corresponding to the independent movie. We further contrast this cortical map of prediction performance against another map computed as the voxel-wise correlation coefficient between the mean neural response across all 158 training subjects and the respective subject-specific response on the independent movie. The latter places an upper bound on the predictivity of each voxel as achievable by any group-level model. Here, we present the results for 5 subjects with mean prediction accuracy (un-normalized) within the stimulus-driven cortex in the *i*th percentile with $i \in \{0.01, 25, 50, 75, 99.9\}$. The results (Figure S10) suggest that the model can successfully capture the response component that individual subjects share with the population.

Correcting with inter-group synchrony

Since the present study focuses on population-wide predictive models, another upper bound on performance estimates that naturally comes to mind is one based on inter-subject or inter-group synchrony in cortical activity on the independent test movie. We computed split-half correlations between the mean response time-course of each group on the test movie. To compare the prediction accuracy against ISC, we divided the prediction accuracy of the best predictive model, i.e., the Audiovisual-20sec model by this synchrony-based noise-ceiling to get the synchrony-normalized prediction accuracy, shown in Figure S11. A stronger shift towards values approaching unity indicates that the model is able to capture stimulus-driven activity highly accurately across large regions of the cortex.

Influence of motion

fMRI measurements are prone to various sources of noise, including spurious head motion and physiological artifacts, which may vary in systematic ways with the variables of interest in any study. While the fMRI data was pre-processed with motion correction, the effects of motion cannot be fully eliminated and need to be further accounted for. Motion confounds have been

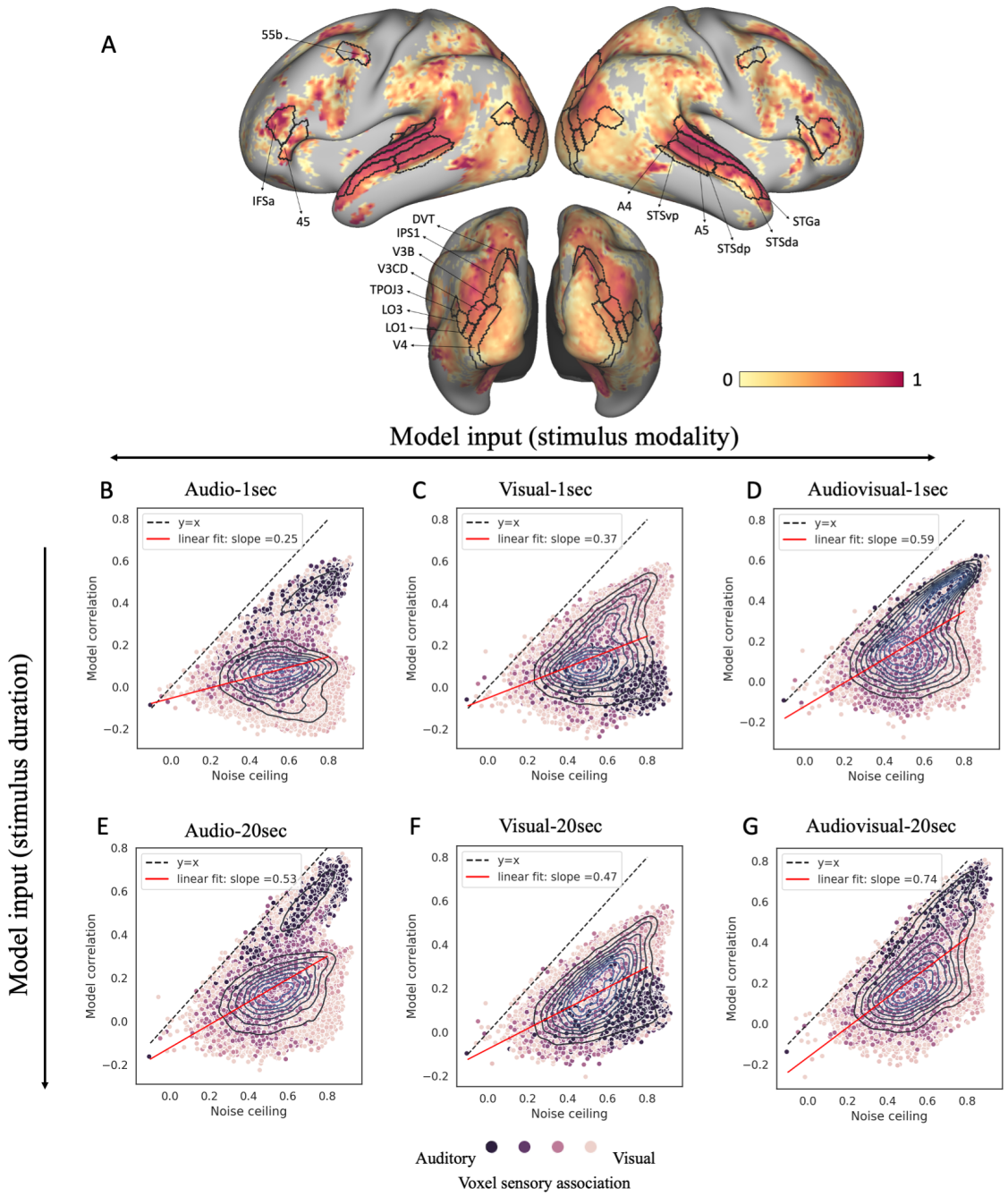

Fig. S8. Model performance on held-out group of subjects. (A) Pearson correlation coefficient (R) between the model predictions and group-averaged response of an independent subject group comprising 20 subjects, on the held-out test movie, normalized by the voxel-specific noise ceiling. (B) Predictivity against the noise ceiling for all voxels with high "synchrony" across training movies (>0.5) (see Supplementary Information for details). This gives a total of 52,954 highly "synchronous" voxels that are colored based on their association with auditory and visual groups. This hue assignment of each voxel was derived from the coloration of the corresponding ROI in the multi-modal HCP parcellation. Each dot in the scatterplot represents an individual voxel. Bivariate kernel density estimates are overlaid on top of the scatterplot as contours to depict the probability distribution of observations (prediction accuracy/noise ceiling pair at every voxel).

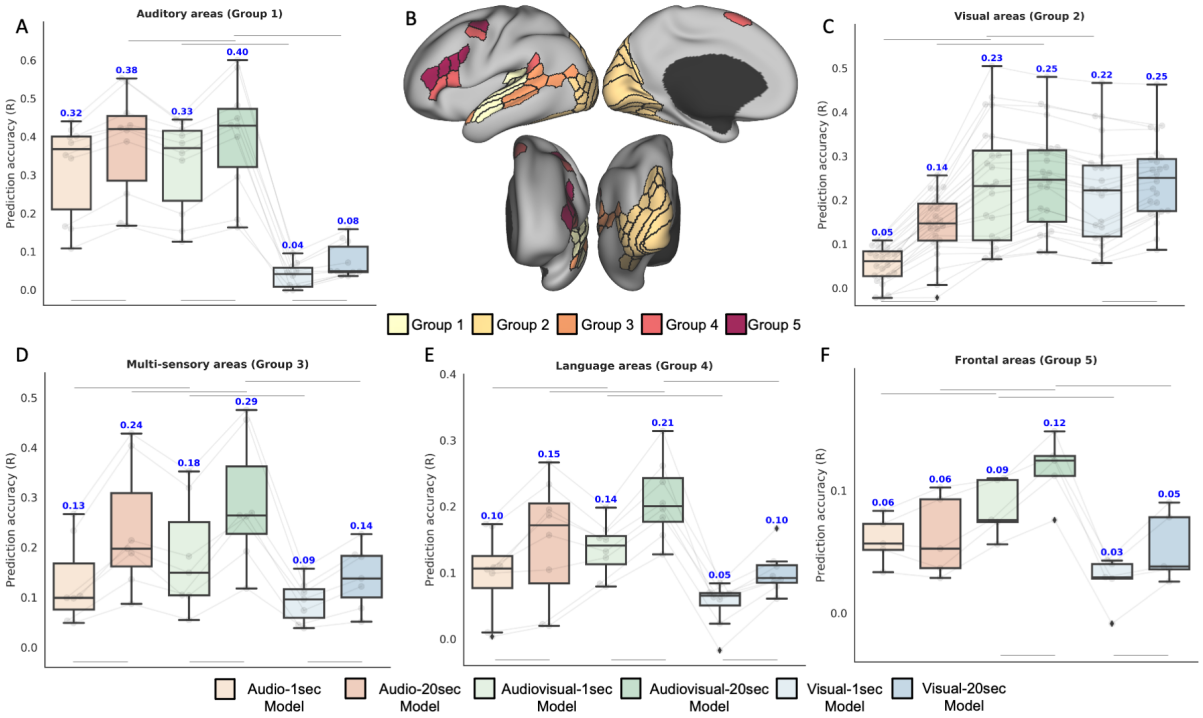

Fig. S9. Quantitative evaluation metrics for all the proposed models on the independent *held-out* population comprising 20 novel subjects. (A),(C)-(F) depict prediction accuracy (R) for all the proposed models across major groups of regions as identified in the HCP MMP parcellation (B). Predictive accuracy of all models is summarized across (A) auditory, (C) visual, (D) multi-sensory, (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 the prediction accuracy. Statistical significance tests (results indicated with horizontal bars) are performed to compare 1-sec and 20-sec models of the same modality (3 comparisons) or uni-modal against multi-modal models of the same duration (4 comparisons) using paired t-test (p-value < 0.05, Bonferroni-corrected) on mean prediction accuracy within ROIs of each group.

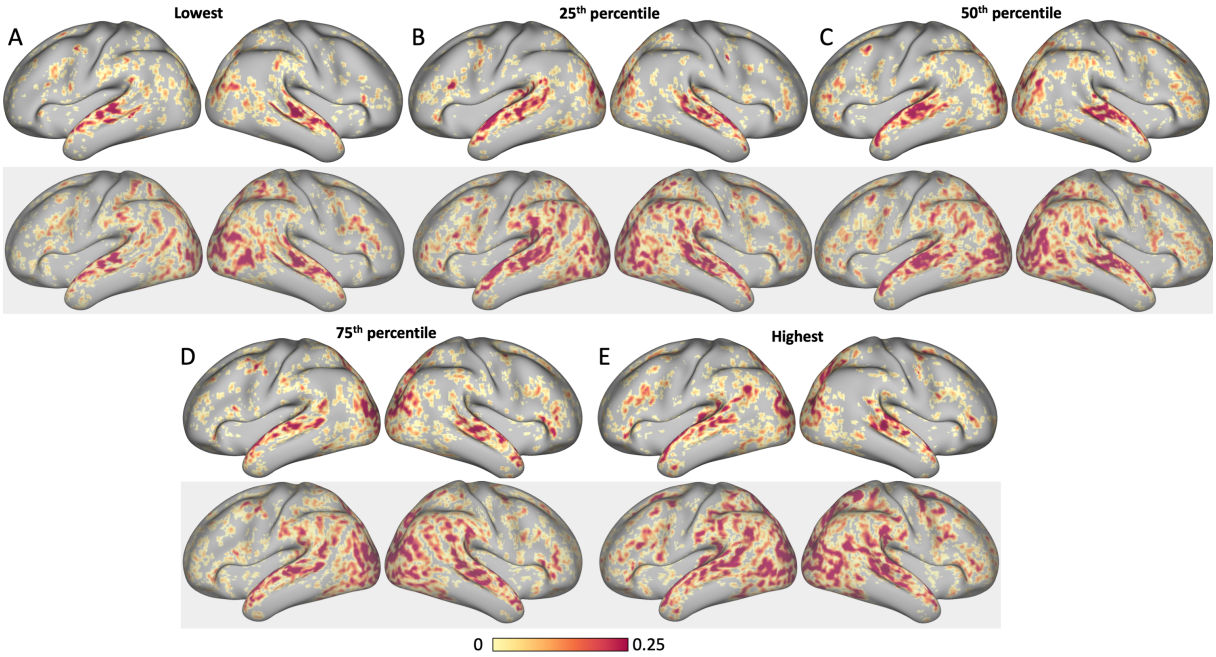

Fig. S10. Comparison of voxel-level prediction accuracies (R) against subject-specific noise ceiling for 5 representative subjects from the held-out set. The subjects were chosen such that their mean prediction accuracy (unnormalized) within the stimulus-driven cortex lied in the ith percentile with $i \in \{0.01, 25, 50, 75, 99.9\}$. Surface maps with white background in (A)-(E) depict raw correlation coefficients between model (Audiovisual-20sec) predictions and subject-specific response on the held-out movie whereas maps on gray background indicate the respective subject-specific noise ceiling. Only significantly correlated voxels ($p<0.05$, FDR corrected) are colored on the surface.

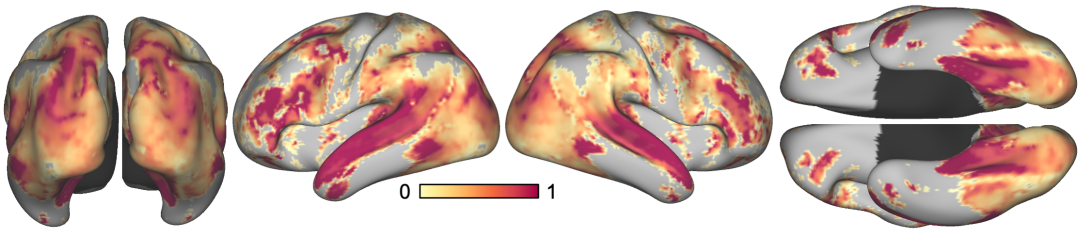

Fig. S11. Synchrony-normalized prediction accuracy (R) of the Audiovisual-20sec model

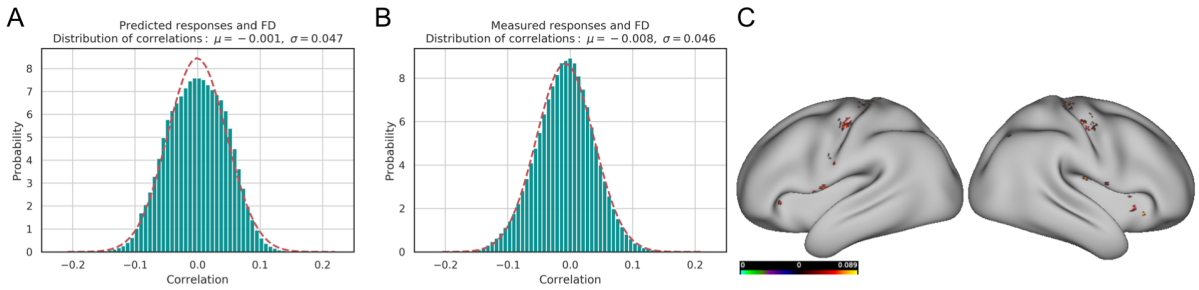

Fig. S12. Addressing the influence of motion on measured and predicted responses. (A) and (B) depict the distribution of the Pearson correlation coefficient of FD with the predicted responses of the Audiovisual-20sec model and measured responses across the whole brain respectively. Surface maps in (C) depict the raw correlation coefficients between FD and the measured responses. Only statistically significant voxels ($p < 0.05$, FDR corrected) are colored on the surface.

reported in prior studies that use neuroimaging data as a "predictor" for different behavioral states or as clinical biomarkers. In our study, the inputs are natural images and the "predicted" variable (fMRI response) is the one prone to motion artifacts. In this study, we developed group-level predictive models of whole-brain cortical activity. One could expect to see the influence of motion in predictions if there was a systematic correlation between motion signals across subjects (so that the signal could persist post averaging), which would suggest that average subject motion tracks the stimulus characteristics. To address this issue, we examined the Pearson correlation coefficients between the predicted/measured response of each voxel and the framewise displacement across the independent test movie clips. The framewise displacement was computed as described in Power et al.,[62] from the averaged motion estimates across subjects on the independent test movie.

$$FD(t) = \sum |d(t-1) - d(t)| + 50\frac{\pi}{180}\sum |r(t-1) - r(t)| \quad (1)$$

where *d* denotes translation distances $\{x, y, z\}$ and *r* denotes rotation angles $\{pitch, yaw, roll\}$. As shown in Figure S12, the correlation coefficients are centered around zero with a very small standard deviation ($\sim$0.05). Importantly, upon computing the p-value of the obtained sample correlation coefficients for the null hypothesis of uncorrelation (under the assumptions of a bivariate normal distribution), we observed that none of these correlations were significant for the predicted responses and only very few voxels (shown on the cortical surface below) were found to exhibit statistically significant correlations between measured responses and FD ($p < 0.05$, FDR corrected).

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).