Ubiquitous cortical sensitivity to visual information during naturalistic, audiovisual movie viewing

Hannah Small[1*], Haemy Lee Masson[2], Ericka Wodka[3,4], Stewart H. Mostofsky[3,4], Leyla Isik[1]

[1*]Department of Cognitive Science, Johns Hopkins University, Baltimore, 21218, MD, USA.
[2]Department of Psychology, Durham University, Durham, DH13LE, UK.
[3]Center for Autism Services, Science and Innovation, Kennedy Krieger Institute, Baltimore, 21211, MD, USA.
[4]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, 21287, MD, USA.

*Corresponding author(s). E-mail(s): hsmall2@jhu.edu;
Contributing authors: haemy.lee-masson@durham.ac.uk; wodka@kennedykrieger.org; mostofsky@kennedykrieger.org; lisik@jhu.edu;

Abstract

Both vision and language carry rich information useful for social understanding in the real world, yet the neural processing of these signals have been mostly studied separately. Even most prior work with naturalistic stimuli does not model the contributions of vision and language signals together. Here we combined established fMRI localizer experiments, which identify social interaction perception- and language-selective regions, with a fMRI movie-viewing paradigm in the same individual participants (n=34). To pinpoint how multi-modal signals contribute to movie responses, we densely labeled the movie using vision and language deep neural networks (DNNs) and use these to predict neural responses. We found that vision model (motion and image) embeddings of movie frames predict significant activity across the cortex, while language model (speech, word, and sentence) embeddings of the spoken language predict well only in portions of the STS. We find that the individually localized motion and social interaction regions are best explained by vision model embeddings. Language regions, on the other hand, are well predicted by speech, word, and sentence language model embeddings and, surprisingly, are as equally well predicted by vision model embeddings. In an analysis of the vision model's layer-wise and unit-wise predictivity, we find that the most predictive model units in social interaction and language regions are distinct from those in lower-level motion regions. Exploratory analyses suggest that the most predictive vision model units in social interaction and language regions contain social-semantic information conveyed by vision. Together, these results suggest that high-level visual information drives neural responses across cortex, even in language-selective regions, with varying integration of spoken language information across the STS.

1 Introduction

Humans effortlessly integrate vision and language signals in everyday social interactions. Both of these signals carry rich social information and often occur simultaneously. How do our brains process these simultaneous, socially-rich signals? Past work points to the superior temporal sulcus (STS) as a hub of social processing.

The STS is famously multi-modal, with many studies finding integrative processing along its length [1, 2]. Although the STS performs many functions, there is spatial and functional structure that seems to support distinct types of social perception and cognition [3, 4]. For example, portions of the posterior and anterior STS are specifically engaged when viewing visual displays of dynamic social interactions between other agents [5–8]. Recent work found that these regions are also sensitive to social interactions presented in an auditory modality [9], suggesting that these regions are involved in the multi-modal extraction of social information.

Social information can also be extracted from language input. Language processing is supported by a network of left-lateralized regions in the posterior temporal and frontal lobe that are selective for meaningful language over many non-linguistic tasks [10]. However, while these regions extract meaning from language input, recent work found that these regions do not respond more to social interactions conveyed by dialogue than they do non-social monologues [11]. Interestingly, right hemisphere homologues respond more to dialogue than monologue [11]. Critically, no study has compared social visual and language responses in the same participants.

Most of this past evidence is limited to studying brain responses to unimodal, controlled stimuli. Recent work using naturalistic movies aims to move beyond controlled unimodal experiments [12] to study responses in more ecologically valid contexts. Much of this work has relied on measuring correlated brain responses across participants [13, 14], leaving open questions about the specific features driving these correlated responses. One powerful method for linking specific features to neural responses is an encoding model approach [15–18], which fits a linear mapping between features and neural responses. However, while this encoding model approach is becoming increasingly common in naturalistic stimuli, it is still mostly limited to studies of isolated vision [19, 20] or language [21–28]. It remains unclear how these regions respond when processing simultaneous vision and language signals in an audiovisual setting.

To investigate how the brain processes simultaneous vision and language during social perception, we used established functional localizer experiments to identify social interaction perception- and language-selective regions in the same individual participants for the first time. We then examined their responses to a naturalistic, audiovisual movie. We densely labeled the movie with both vision and language deep neural network models and human annotations, which allowed us to link different movie features to neural responses with an encoding model approach. We found that social interaction and language regions are largely non-overlapping but during the movie they shared a similar functional profile, with visual features dominating prediction across the cortex. Both social interaction and language regions are also predicted by language features, but language regions are sensitive to more levels of language processing (speech, word, and sentence). Variance decomposition analyses showed that vision model predictivity in all regions is mostly driven by later layers, but the most predictive units within these layers differ across vision and language regions. Finally, exploratory analyses suggest that vision model predictivity in language regions overlaps with high-level social-semantic information conveyed by vision.

2 Results

2.1 Little similarity between vision and language model embeddings during natural movie

Each participant watched a naturalistic audiovisual movie: an episode of the BBC series Sherlock (Figure 1a), while their whole brain fMRI activity was recorded. To operationalize the visual and language features of the movie, we automatically extracted embeddings from vision and language deep neural network models (Figure 1b) and collected human annotations of visual, social, and language features in the movie. For vision models, we extracted embeddings from a motion energy model [29, 30] and from the seven layers of AlexNet [31], which have previously been shown to predict visual responses in high-level visual cortex [32] of the first frame of each TR (1.5s). For language, we extracted activations from all layers of a speech transformer model (HuBERT [33]), a word-level semantic model (word2vec [34]), and a sentence-level transformer model (sBERT; all-mpnet-base-v2, huggingface.co) of the spoken content of the episode from the movie transcript. In follow-up analyses, we ensured that the results held when using more modern vision and language models, SimCLR and GPT-2 (Supplemental Figure S7). Human annotated features included relevant aspects of social processing based on prior studies, including the presence of faces on screen, the presence of social

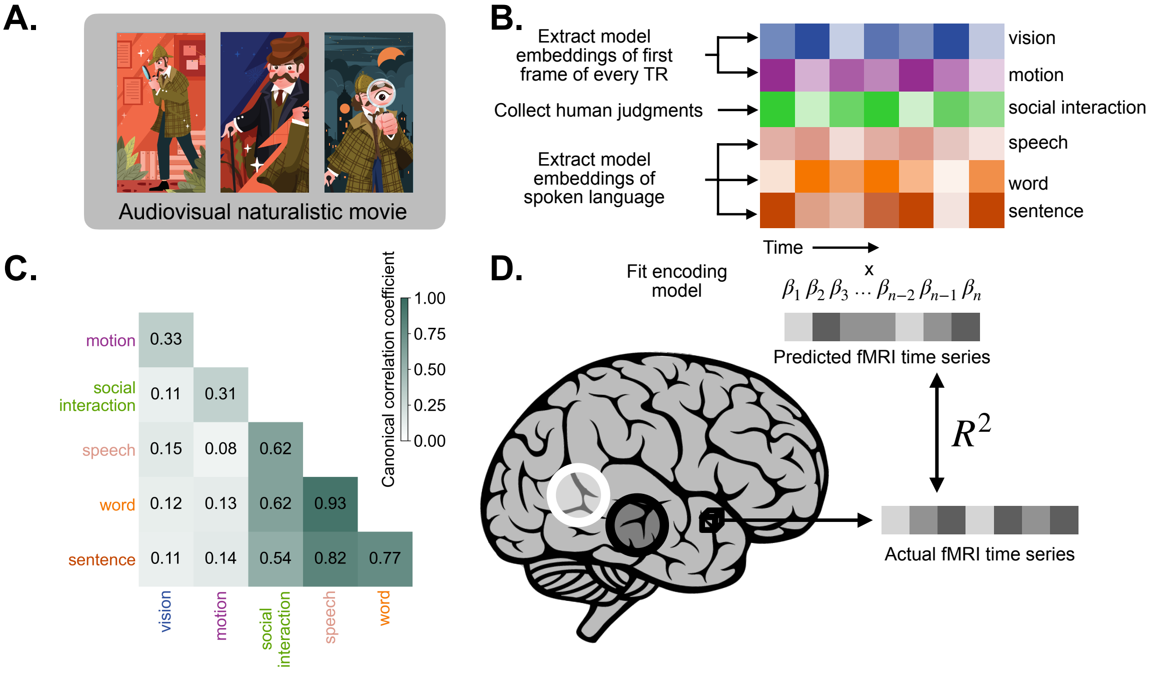

A.
Audiovisual naturalistic movie
B.
Extract model embeddings of first frame of every TR
Collect human judgments
Extract model embeddings of spoken language
vision
motion
social interaction
speech
word
sentence
Time
C.
motion 0.33
social interaction 0.11 0.31
speech 0.15 0.08 0.62
word 0.12 0.13 0.62 0.93
sentence 0.11 0.14 0.54 0.82 0.77
vision
motion
social interaction
speech
word
Canonical correlation coefficient
1.00
0.75
0.50
0.25
0.00
D.
Fit encoding model
x
$\beta_1\,\beta_2\,\beta_3 \cdots \beta_{n-2}\,\beta_{n-1}\,\beta_n$
Predicted fMRI time series
R^2
Actual fMRI time series

Fig. 1 A. We recorded fMRI activity while participants (n=34) watched an audiovisual, live-action episode of the BBC series Sherlock (figure shows a cartoon representation) B. We extracted features from the movie, including vision embeddings of the frames, language model embeddings of the spoken content, and human annotations of visual, auditory, and social features (see full list in Methods). C. Similarity between select feature spaces during a naturalistic movie. Each square shows the canonical correlation coefficient of a pair of feature spaces. The similarity of spaces from vision (AlexNet), speech (HuBERT), and sentence (sBERT) embeddings was measured layer-wise. The average across all layers is shown here. The full layer-wise similarity matrix is available in Supplemental Figure S2. D. For each participant, we fit a cross-validated encoding model using banded ridge regression to predict voxel-wise activity within an ISC mask (Figure 2a). Performance is evaluated by calculating coefficient of determination between the predicted and actual time series. We examine both whole brain performance and within social interaction (white outlines) and language (black outlines) selective regions.

interactions, valence and arousal annotations, as well as other automatically extracted low-level visual features like pixel values and hue (see full list in Methods Section 4.7).

We measured the similarity of these feature spaces over the movie with Canonical Correlation Analysis (CCA) [35, 36]. This method finds the linear combinations of two feature spaces that maximizes their correlation, likely providing an overestimate of the correlation between multi-dimensional feature spaces. As expected, there was high similarity between the vision model embeddings (AlexNet and motion average correlation = 0.33; Figure 1c) and between the language model embeddings (speech, word, and sentence average correlation = 0.84; Figure 1c). Interestingly, there was little correlation between the vision and language feature spaces over the course of the movie (average vision-language correlation of 0.12; Figure 1c). This result held when using 5, 10, and 100 latent dimensions in CCA (average vision-language correlations of 0.09, 0.08, and 0.05, respectively). This low correlation suggests that in naturalistic contexts vision and language convey different information and facilitates future analyses investigating the brain responses to these signals [37, 38]. We also found that the human annotated social interaction feature was correlated with both the language model embeddings of the spoken content and, to a lesser extent, the motion model features (Figure 1c), emphasizing the multi-modal nature of social interaction recognition in naturalistic settings.

2.2 Visual features of the movie dominate prediction across cortex, while sensitivity to language varies

We linearly mapped the feature spaces (including deep neural network embeddings and human-annotated features) over the course of the movie to fMRI BOLD responses using banded ridge regression (Figure 1d). To reduce computational cost and limit analyses to voxels with stimulus-driven signal, we computed an intersubject correlation (ISC) mask and performed voxel-wise encoding within

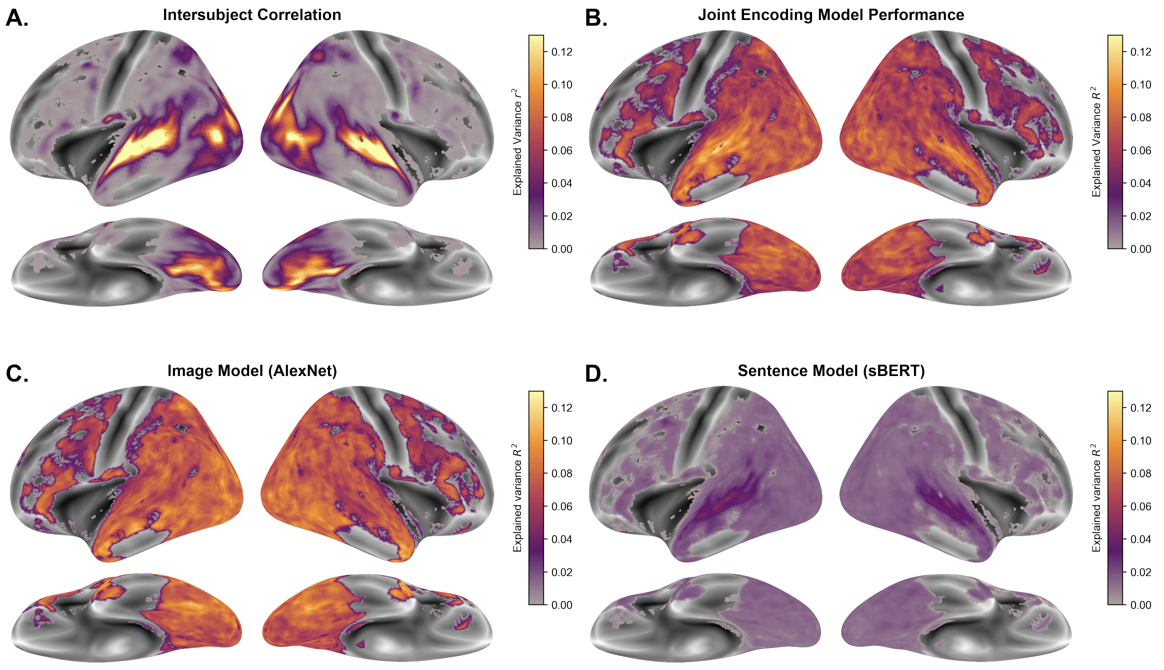


Fig. 2 A. The average intersubject correlation across participant. Note that we squared the correlation to be comparable to the explained variance (R^2) values of the encoding models . The map is thresholded to FDR q<0.001 (permutation test) B. The group-averaged performance of the joint feature encoding model, thresholded to FDR q<0.001 (signed permutation test). C) Group-averaged product measure map for the vision (AlexNet) and sentence (sBERT) features, thresholded to FDR q<0.001 (signed permutation test). The motion, speech, and word models explain less variance and are available in Supplemental Figure S3.

this mask (Figure 2) [39, 40]. The joint encoding model explained significant group-level variance (FDR corrected q<0.001) in all voxels and was higher than the ISC in most voxels (Figure 2ab).

We next examined the predictive contributions of all individual features using the product measure, an estimate of the unique contribution of each predictor in the joint encoding model [41]. Vision model (AlexNet) embeddings explained significant variance in the majority of voxels throughout the cortex at the group-level (FDR q<0.001; Figure 2c, see example individual participant in Supplemental Figure S3), and was the highest performing feature space. The next most predictive feature space was the embeddings of the sentence-level language transformer, which was most predictive in temporal regions (Figure 2d). Other high dimensional feature spaces also predicted significant variance (FDR q<0.001; Supplemental Figure S3), although much lower in magnitude. The motion model embeddings explained variance particularly in MT area and ventral visual stream. In contrast, the speech and word model embeddings explained variance in temporal regions, similar to the sentence level model, but with, in general, fewer significantly predicted voxels than the sentence-level model (FDR q<0.001; Supplemental Figure S3). No uni-dimensional features predicted significant variance in any voxels in a whole-brain group analysis.

2.3 Visual features predict neural responses in social interaction and language regions

To understand how brain responses to the movie vary across known functional regions of interest, participants also completed localizer experiments to identify social interaction- and language-selective regions, as well as a control visual region MT. As in past work, we identified group-constrained participant-specific regions of interest for social interaction within posterior and anterior portions of the STS [8] and language in posterior and anterior temporal and frontal regions [42]. In the temporal lobe, the social interaction and language regions are functionally (Supplementary Figure 5) and spatially non-overlapping (DICE coefficients: right posterior = 0.05 ± 0.05, right anterior = 0.04 ± 0.08, left posterior = 0.05 ± 0.08, left anterior = 0.03 ± 0.04). We next examined encoding model performance across the localized regions of interest. The full encoding model explained about 20-30%

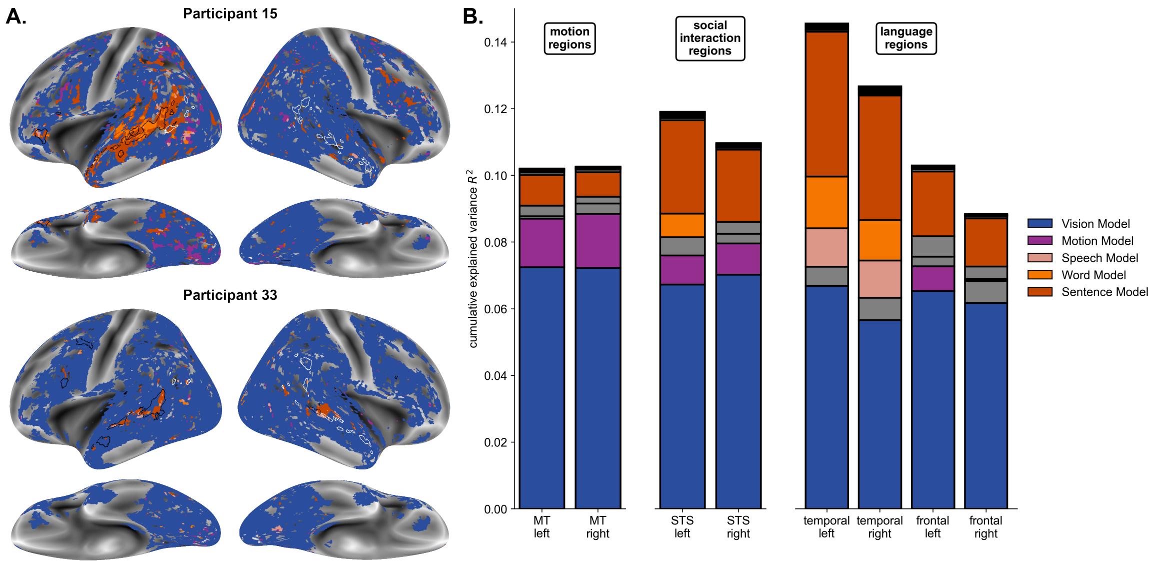


Fig. 3 A. Examples of individual vertex-wise preference maps of feature predictivity in the joint encoding model. Each vertex is colored according to the feature that best predicts its responses (has the highest product measure). Individually localized social interaction (white) and language (black) regions are overlaid on each preference map. All participant's preference maps are available in Supplemental Figure S4. B. Feature predictivity averaged across all participant regions. Each rectangle represents the variance explained by that feature space (all layers are added together when relevant) as measured by the product measure. Colored rectangles indicate a feature space that explains significantly (linear mixed-effect model (LME), Bonferroni corrected ps<0.05) more than zero variance (gray indicates non-significance). The responses of posterior and anterior social and language regions are very similar and averaged per participant before averaging across the group. Visualization and statistical comparisons of vision and combined language model predictivity is available in Supplemental Figure S7.

of an estimated cross-subject noise ceiling (Figure 3b shows the cumulative total variance explained; see performance of cross-subject model in Supplemental Figure S7a), similar to previous language encoding models predicting responses to auditory narratives [43].

To understand the relative predictive contributions of each feature space across the brain, we created vertex-wise preference maps by coloring each voxel (projected to the surface) by the feature with the highest product measure. We performed these analyses at the individual level to avoid blurring feature selectivity due to anatomical differences across subjects. Most voxels were best predicted by vision (AlexNet) embeddings, although there were voxels best explained by language model embeddings, especially in the temporal lobe and STS (Figure 3a). In some participants (Supplemental Figure S4) these correspond to the language regions identified with controlled stimuli (black outlines), as well as the social interaction perception regions (white outlines). Strikingly, across participants, the vision model consistently dominates prediction across cortex, while language models predict best in more specific portions, mostly in the temporal lobe.

A similar trend was found in ROI analyses (Figure 3b) with vision model (AlexNet) predictivity exceeding the combined language model (speech+word+sentence) predictivity in control region MT and social interaction regions (Supplemental Figure S7b). Surprisingly, the vision model embeddings also explain significant variance even in temporal language regions, where vision model embeddings were just as predictive as the combination of all language model embeddings (speech+word+sentence) (Figure 3), suggesting an important role of visual features in driving responses in these regions. Further, in frontal language regions, vision model embeddings were more predictive than the combined language model embeddings (Supplemental Figure S7b). Follow-up work with more advanced vision and language models (SimCLR, a vision transformer trained using a contrastive image objective without any language input, and GPT-2, a transformer trained to predict the next word) showed a similar trend to the smaller models used in our main analyses, with significantly more variance explained by vision compared to language models in all regions, including language regions (Supplemental Figure S7c).

Vision and language models predicted variance in all regions to significantly varying degrees (Supplemental Table 1). The sentence-level model explained significant variance in all regions of interest, including a small but significant portion of variance in MT (Figure 3b). However, speech and word-level models explained significant variance only in bilateral temporal language regions,

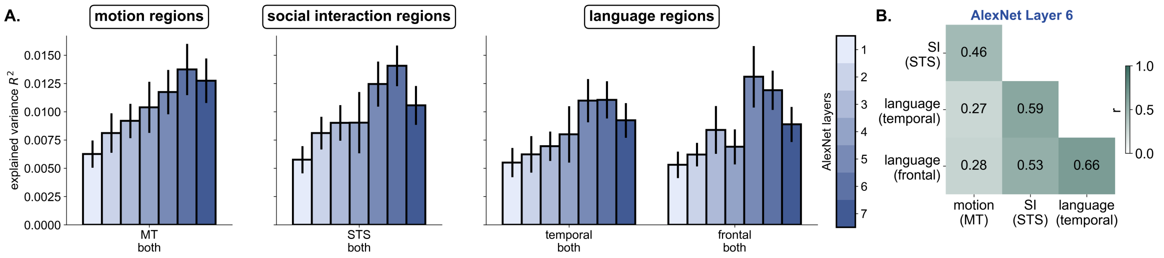


Fig. 4 A. Average product measure of each layer of AlexNet from the encoding model. The responses in posterior and anterior and left and right regions are very similar and are averaged per participant before averaging across the group B. Pairwise correlations of beta weights from AlexNet layer 6, the highest performing AlexNet layer in the encoding model. Again, posterior and anterior and left and right regions are averaged per participant before computing the pairwise correlations. (separated left and right hemispheres available in Supplementary Figure 8)

highlighting a unique role of these regions for processing multiple levels of information about spoken language. Interestingly, motion embeddings explained significant variance in only one language region (left frontal), but they explained significant variance in all motion and social interaction regions (Figure 3b).

Given the significant predictivity of vision (AlexNet) and sentence models, we looked at how well individual voxels in this region were predicted by each model (Supplementary Figure 6). While most voxels in social interaction and language regions were well-predicted by one model and not the other, there were some that were predicted by both, suggesting the presence of some multi-modal voxels in these regions (note that the product measures sums to the total explained variance, so it is unlikely to have a voxel maximally predicted by both model embeddings).

2.4 High-level visual features drive activity across vision and language regions

To investigate the surprising predictivity of vision model embeddings in language regions, we asked whether this predictivity was driven by early or later layer features of the model. We examined the separate product measures of all layers of AlexNet. There were no differences between the layer-wise predictivity patterns in social interaction or language regions (linear mixed-effects model (LME), Bonferroni corrected ps>0.05), with all regions being best predicted by later layers of AlexNet (shown averaged across hemispheres in Figure 4, no difference between left and right hemisphere, see Supplemental Figure S8a). The language regions' vision model layer-wise predictivity was not significantly different from motion-selective MT's (LME, Bonferroni corrected ps>0.05), suggesting high-level visual features drive responses in both vision and language regions.

To further investigate how the vision model predicts responses across regions, we next asked whether similar patterns of model units were used to predict each region's responses. We computed the regional pairwise correlations of the fitted weights of AlexNet layer 6 (the best performing layer across regions) from the joint encoding model. We found that there were significantly higher correlations between the fitted weights in social interaction and language regions than between the fitted weights in motion MT and social interaction or language regions (Figure 4). This pattern is consistent across both hemispheres (Supplemental Figure S8; statistics in Supplemental Table S5), suggesting that there is similar visual feature tuning in social interaction and language regions that is distinct from the tuning in motion regions.

2.5 Most predictive vision model units in social interaction and language regions are aligned with social-semantic features

Finally, we conducted exploratory analyses to understand which aspects of the vision model embeddings are driving the predictivity across the different regions of interest. We fit an encoding model with just the vision and language model embeddings, excluding the non-predictive uni-dimensional features for interpretability. From this fitted model, we extracted the 25 highest weighted units from AlexNet layer 6 (the best performing layer in most regions) in each region in each participant. To interpret those units, we measured their similarity to the uni-dimensional annotated movie features using CCA. We found that across participants, the vision model units' feature correlations did not

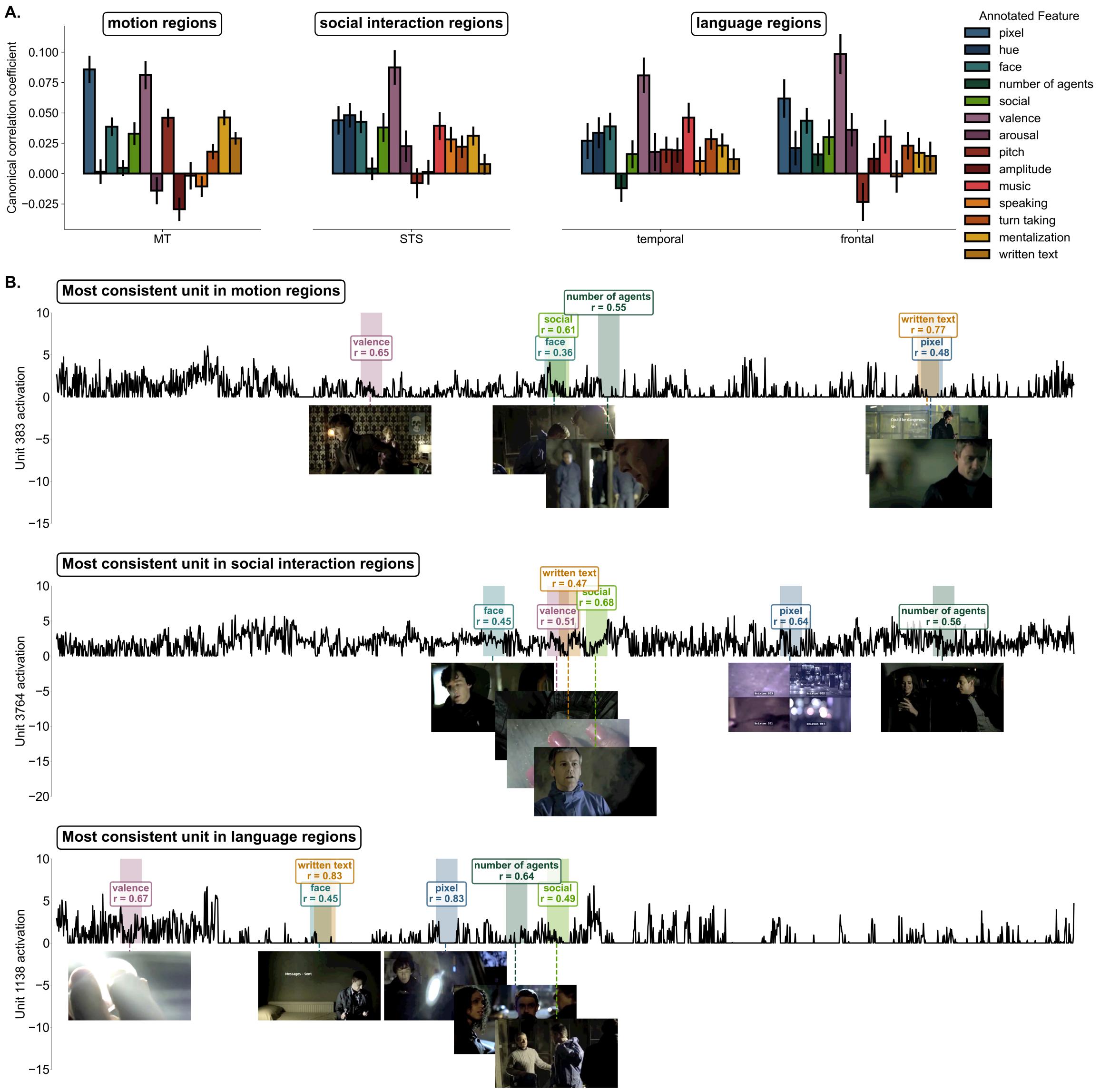


Fig. 5 A. Correlations between the top 25 highest weighted units from the most predictive vision model layer, AlexNet layer 6, and annotated one dimensional features. Each bar represents the participant-averaged similarity between the AlexNet units and features during the movie. B. Visualizations of the most consistently highly weighted AlexNet layer 6 units across participants in localized motion (MT, unit present in top 2.5% weighted units of 31 participants, unit top), social interaction (STS, unit present in top 2.5% weighted units of 16 participants, middle), and language (combined temporal and frontal, unit present in top 2.5% weighted units of 18 participants, bottom) regions. The black line shows one unit's activations throughout the movie and the highlighted boxes shows the maximal correlation (Pearson's r) with a given feature. Correlations were calculated with a sliding window of 40 TRs. The frame with the highest unit activation during that window is displayed for each feature.

significantly differ between social interaction and language regions, but both were significantly different from MT, where there was a higher correlation with low-level pixel-level information (LME, ps<0.001). The highest weighted units in social interaction and language regions were most aligned with valence. Other correlated features included a mix of social features, such as social interaction, faces, speaking, and turn-taking, and low-level visual features, like pixel and hue (Figure 5a).

We then visualized the most consistently highly weighted vision units across participants. We extracted the top 2.5% highest weighted units from each participant's fitted model (from AlexNet layer 6) and found the unit that most consistently occurred across participants for each region. We measured each unit's similarity (Pearson's r) to social features of interest and the presence of written text using a sliding window through the time course. This revealed that in each region, the same unit correlates well with different features at different times in the movie. Notably, in the language regions the same unit was correlated with written text in one portion, where a frame with a text message on the screen drives the highest activation, but was also highly correlated with number of agents in another portion, where a frame showing three people drives the highest unit activation (5c). These results tentatively suggest that the visual responses in language regions are not only driven by the presence of written language, but also social-semantic features conveyed by vision.

3 Discussion

In this work, we make several novel contributions towards understanding multi-modal social processing in the human brain. We found that individual participant-defined social interaction- and language-selective regions' responses to an audiovisual movie are well-predicted by uncorrelated vision and language model embeddings, suggesting that both regions integrate multi-modal information during social processing. Surprisingly, language regions were equally predicted by visual information in the movie as they were by spoken language content and this mixed selectivity was found even at the voxel-level in some cases. Follow-up analysis of the layer-wise responses showed that high-level visual information drives responses in both vision and language regions and analysis of the unit-wise responses showed that different units contribute to the predictivity in vision and language regions. Interpreting these units with hand-annotated features revealed that social perception and language regions were more sensitive to social-semantic information conveyed by vision and motion regions were as sensitive to low-level visual information.

We found multi-modal processing in regions of the STS/temporal lobe, in line with past work showing that multi-modal encoding models predict these regions better than uni-modal models [38, 44, 45]. We further showed that language regions were accurately predicted by both vision and language model embeddings, suggesting that visual information drives responses as much as the spoken language content does. The language network has been argued to respond specifically and selectively to meaningful language (for a review see [10]), but this does not preclude the possibility that language regions are affected by meaningful non-verbal information when such visual signals are available [46]. For example, our work is consistent with findings showing significant intersubject correlation in localized language regions during viewing of silent but semantically meaningful movies [47, 48]. We extended beyond previous work by linking specific visual features to language region responses in naturalistic stimuli. Specifically, we found correlation between the highest weighted vision model units in language regions and social-semantic features of the movie. The extent to which this information is being used for online language processing is an open question for future research.

We also showed for the first time that social interaction perception and language regions, as identified with standard controlled experiments, are largely non-overlapping, spatially and functionally. However, in the movie, responses across social interaction and language regions showed more functional overlap, consistent with other work showing more distributed responses to naturalistic stimuli [21, 43, 49]. Our findings in the controlled and naturalistic experiments point to an intimate relationship between these regions. However, while both social interaction and language regions were well predicted by high-level visual and sentence-level language information, there were differences in their response profiles during the movie. Motion drove responses in social interaction regions but not in language regions, underscoring the importance of visual motion processing in social interaction regions [50, 51] and suggesting that motion integration is a critical difference between social interaction and language regions. We also find that right social interaction regions and bilateral frontal language regions were exclusively sensitive to sentence level information, but not speech and word information. Only bilateral temporal language regions were sensitive to speech, word, and sentence information. This is consistent with previous intracranial work showing that temporal region activity is better predicted by speech embeddings while frontal regions are better predicted by language embeddings [28]. Our results are consistent with a recently proposed "soft processing hierarchy" within the temporal lobe, where language regions communicate via acoustic, speech, and language features [52]. In this hierarchy, increasingly abstract language features are the basis of communication with other higher-order cortical regions, perhaps regions like the social interaction regions, which we find are sentence-sensitive.

While past work emphasizes the semantic convergence of vision and language in brains and models [53–55], we investigated their interaction when they are presented simultaneously in a naturalistic, audiovisual setting. Interestingly, we found that vision and language neural network embeddings are only very weakly correlated over the course of a naturalistic movie. This may be surprising given the tendency of features to co-vary in naturalistic contexts and recent work emphasizing the alignment between vision and language in brain [53–55] and in model representations of images [56]. However, our results illustrate an additional perspective [38], that simultaneous vision and spoken language are not necessarily aligned in naturalistic settings. How language, which communicates rich information about people and things that may or may not be physically present, interacts with high-level visual information is an exciting open question. The predictivity of visual features, which do not converge with spoken language information, highlights the importance of using multi-modal data in studies of vision-language convergence.

This work is a critical first step in understanding naturalistic, multi-modal social processing, but there are some limitations. First, the stimulus was a commercial movie, with editing decisions to develop a compelling narrative [37]. These include sharp cuts between faces, close-ups, music, and so on, which all create a more enjoyable, stimulating experience but introduce difficulties when interpreting responses to the movie. While we employed analytical methods to ensure that our results are not due to trivial correlations in the movie features, there is nevertheless the possibility that the directed nature of the movie increases the similarity between visual features and other aspects of the movie. Second, while deep neural networks are a performant means to automatically extract vision and language information, they are also "black boxes" with highly correlated layers, making it difficult to strongly interpret layer-wise differences. We sought to overcome this with our dense movie annotations to understand how network predictivity may be driven by interpretable features of interest. Finally, this work is unique because we localized social interaction and language regions in the participants, however, understanding multi-modal, naturalistic stimuli likely engages additional networks which need to rapidly communicate with one another. In particular, the nearby theory-of-mind (ToM) network [3], which supports mental state processing in both text and visual movies [39, 48, 57–62] is especially relevant for processing multi-modal, naturalistic stimuli. Investigating how the ToM network interacts with social perception and language regions during naturalistic processing is an important future direction.

Overall, our work points to intimate interactions between the visual and language systems supporting social processing in the human brain. The surprising predictivity of visual signals across cortex, including the language network, highlights the importance of studying cognitive processing in multi-modal, natural contexts. This work also makes important methodological contributions by leveraging deep neural network models in a multi-modal encoding model approach, and presents new, exciting opportunities to analyze naturalistic, audiovisual fMRI movie data in a feature driven manner.

4 Methods

Data and code for replicating analyses are available at OSF (https://osf.io/by8pw/) and GitHub (https://github.com/Isik-lab/ubiquitous-vis).

4.1 fMRI experimental procedure and data collection

Participants (n=39, neurotypical, ages 19-38, 17 female) watched the first 45 minutes of the first episode of the BBC series Sherlock. We chose this stimulus because it has been validated in prior literature [39, 63] and contains a narrative that depends on rich social understanding. Participants were told to pay attention as if they were watching a television show they were interested in. They were told that they would need to summarize what they watched afterwards, but they did not need to memorize the episode. The episode was split into two runs. The first run was 23 minutes and 39 seconds long and the second run was 25 minutes and 45 seconds long, exactly the same as previous studies using this stimulus [39, 63]. All experiments were approved by the Johns Hopkins School of Medicine IRB. All experiments were performed in accordance with the IRB and informed consent was obtained from every participant.

All participants also completed a social interaction localizer where they passively watched videos of point light walkers that were either engaged in social actions or performing independent actions [5]. Individual videos ranged in length from 3 to 8 s, and three videos were presented in each 16-s block. Each run consisted of eight blocks of each condition and two 16-s fixation blocks presented at the middle and end of each run, for a total time of 160 s per run. Stimulus conditions were presented in a palindromic order. This experiment was split over the course of three runs.

A subset of participants (n=26) also completed a language localizer where they passively listened to auditory stimuli of intact and degraded speech while a dot was in the center of the screen [64].

This audio was from various sources (The Moth podcast, TED talks, celebrity interviews) and was between 16 and 18 seconds long (32 total clips, each with an intact and degraded version). The degraded audio was created by filtering the audio clips and adding noise so that the audio clips were unintelligible but it was still clear that a human was talking. In each run, there were 16 stimulus blocks of 18 seconds interspersed with 5 fixation blocks of 14 seconds. (see full details and text of the audio material at [64]). They completed 2 runs of this task.

All stimuli were presented in the scanner using the Psychophysics Toolbox (http://psychtoolbox.org/), displayed on a monitor inside the scanner room, which the participants viewed with an angled mirror in the scanner. Audio was delivered through in-ear Sensimetrics headphones. Before the scan, we conducted an audiovisual test to select a volume at which the participant could hear and understand the stimulus audio over the sounds of the fMRI scanner. We used this volume for the entire scan.

Neuroimaging data were collected at the Kennedy Krieger Institute on a Philips dStream Achieva 3T TX scanner with a 32-channel head coil. A high-resolution T1-weighted anatomic image (multiecho MPRAGE) was collected at each scan [repetition time (TR), 8,100ms; echo time (TE), 3.7 ms; WFS (pix) / BW (Hz), 2.255/192.7; timing interval (TI), 751.10 ms; flip angle, 8; field of view (FOV) 212x170mm; matrix size, 212x170mm; slice thickness, 1 mm; 150 near-axial slices]. Functional data were collected using a T2*-weighted multi-echo EPI sequence sensitive to blood oxygen level-dependent (BOLD) contrast [TR, 1,500 ms; TE, 30 ms; effective echo spacing, 0.69 ms; multiband (MB) factor, 4; WFS (pix) / BW (Hz),33.696/12.9; BW in EPI freq dir., 1742.2; flip angle, 52; FOV, 216x216 mm, matrix, 108x107mm; slice thickness, 2mm; 60 near-axial slices].

4.2 fMRI preprocessing

All neuroimaging data was converted to BIDS format using the dcm2bids software [65]. Anatomicals were defaced before further preprocessing using pydeface [66] and quickshear [67].

The text of the following sections (Preprocessing of B0 inhomogeneity mappings, Anatomical data preprocessing, Functional data preprocessing) was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts unchanged. It is released under the CC0 license.

Preprocessing of B0 inhomogeneity mappings

A total of 1 fieldmaps were found available within the input BIDS structure for this particular subject. A B0-nonuniformity map (or fieldmap) was estimated based on two (or more) echo-planar imaging (EPI) references with topup [68](FSL 6.0.5.1:57b01774).

Anatomical data preprocessing

A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset.The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection [69], distributed with ANTs 2.3.3 [70], and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast [71]. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1 [72]), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle [73]. Volume-based spatial normalization to two standard spaces (MNI152NLin2009cAsym, MNI152NLin6Asym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [74] [TemplateFlow ID: MNI152NLin2009cAsym], FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model [75][TemplateFlow ID: MNI152NLin6Asym].

Functional data preprocessing

For each of the 7 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated by aligning and averaging 1 single-band references (SBRefs). Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 6.0.5.1:57b01774 [76]). The estimated fieldmap was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.7s (0.5 of slice acquisition range 0s-1.4s) using 3dTshift from AFNI [77]. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration [78]. Co-registration was configured with six degrees of freedom. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions [79]) and Jenkinson (relative root mean square displacement between affines [76]). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by [79]). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor [80]). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. [80] in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's aseg segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each [81]. Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): fsaverage. Grayordinates files [82] containing 170k samples were also generated using the highest-resolution fsaverage as intermediate standardized surface space. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels [83]. Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.8.1[84], mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

Data was smoothed with a 3mm FWHM kernel for subsequent localizer and encoding model analyses. Data was smoothed with a 6mm FWHM kernel for computing the intersubject correlation mask, which is in the recommended smoothing range [40].

All further analyses were performed in volume space, thresholded, and were then projected to the fsaverage inflated surface for visualization purposes only. We projected the activity from along the thickness of the cortex. A line was drawn between each vertex on the white matter surface mesh to the corresponding vertex on the inflated mesh and then sampled at 50 equal distances along the line, with each sample taking the value of the nearest voxel. These were then averaged for that vertex. This was done with nilearn's vol_to_surf() function [84].

Data was smoothed with a 3mm FWHM kernel using Nilearn [84] for subsequent localizer and encoding model analyses.

4.3 Participant Exclusion

In total, 5 participants were excluded from further analysis. One participant was excluded for falling asleep during the localizer tasks. One participant was excluded for falling asleep in the scanner during both runs of the Sherlock episode. Three other participants were excluded due to technical errors during scanning: two participants did not hear the audio in both ears and one participant's neural responses were not correctly time-locked to the episode.

After exclusions, we had data from 34 participants (19-35, 17 female) who completed watching Sherlock and the point-light walkers experiment and 25 (19-35, 10 female) of those also completed the language experiment.

4.4 Analysis of Localizer Experiments

Generalized Linear Models (GLM) were run using Nilearn's run_glm() function [84]. The fMRI BOLD time series was mean-scaled per run prior to fitting the GLM and computing the first level contrasts. Each task model included regressors for all conditions along with confounds from fMRIPrep output, including the 6 rigid-body transformations, framewise displacement, and the aCompCor components from the combined white matter and CSF masks [80] and the associated discrete cosine bases for high-pass filtering (cutoff of 128s).

We identified motion, social interaction, and language selective voxels from various contrasts of the controlled stimuli within previously established parcels. For each contrast, we selected the top 5% most selective voxels across the corresponding parcels in left and right hemispheres. From the point-light walker task we identified social interaction selective voxels within posterior and anterior superior temporal sulcus (STS) using the contrast of interacting-independent dyads. We used a mask of STS and the temporal parietal junction (TPJ) [3], which we split into posterior and anterior portions, as in prior work [8]. We identified motion selective voxels within the FreeSurfer anatomical MT mask using the contrast of interacting & independent dyads. From the language task, we identified language selective voxels within previously identified temporal language regions [42] using the contrast of intact speech-degraded speech.

4.5 Measuring overlap between regions

We measured the proportion of overlap between the social interaction and language regions with the DICE coefficient [85], calculated as the ratio of the number of voxels of overlap between the two sets of voxels and their sum. To test this in the most conservative way, we constrained voxel selection for both sets of voxels to the STS.

4.6 Second Level Analyses of Localizer Experiments

Group maps were created using Nilearn's second level modeling with a smoothing FWHM of 3.0mm. Statistical testing was done with one-sided Benjamini-Hochberg FDR correction [84].

For second level ROI analyses, we extracted responses (beta weights) for each condition of the controlled stimuli experiments. Separate data subsets were used for voxel selection and extracting responses. When the data to select the voxels and extract responses came from different experiments, we used all available runs of each experiment for both voxel selection and response extraction. When the data for voxel selection and response extraction came from the same experiment, we used a subset of the runs (2 runs for point light walker experiment, 1 run for language experiment) to select the voxels and the remaining run to extract condition responses. We repeated this process so that condition responses were extracted from each available run once and then averaged the responses over these splits. Statistical testing was performed using linear mixed-effect modeling with the 'Lmer' function from the pymer4 package [86]. For each pairwise condition comparison, we selected the relevant subset of the data containing only the two conditions being compared. We modeled the beta weight as a function of condition grouped by participant as a random effect, which allows individual variability and controls for repeated measures (see formula below). P-values for each pairwise condition comparison were extracted from the model estimates and corrected for multiple comparisons using Bonferroni correction within each region.

$$\text{Beta weight} \sim 1 + \text{Condition} + (1|\text{Participant})$$

4.7 Analysis of Naturalistic Movie

4.7.1 Deep Neural Network Model Embeddings of Vision and Language Signals

To operationalize the vision signal of the naturalistic movie, we extracted the embeddings of the first frame of every TR from the 7 layers of AlexNet [31]. AlexNet is a convolutional neural network trained to classify images whose layers have been shown to map onto the visual system hierarchy [32]. We extracted activations after the ReLU and max pooling steps, similar to previous work [41]. Specifically, we extracted activations from MaxPool2d-2-3, MaxPool2d-2-6, ReLU-2-8, ReLU-2-1, MaxPool2d-2-13, ReLU-2-16, and ReLU-2-19. In follow-up analyses, we use a more advanced self-supervised self-supervised Vision Transformer (ViT) models trained on the YFCC15M (Yahoo Flickr Creative Commons) dataset, which is a subset of the YFCC100M dataset with English titles and descriptions [87]. The model was trained using a view-based self-supervised SimCLR-style contrastive vision objective [88] (referred to as SimCLR). We used the trained ViT-B versions [89] and extracted embeddings from each of the 12 layers. As in prior work [38, 51], we used the DeepJuice package [90] to extract these embeddings. To reduce computational demand and avoid overfitting, we reduced the dimensionality of these embeddings using sparse random projection, which projects a high dimensional feature space into a lower dimensionality feature space while preserving the pairwise Euclidean distance between points. The dimensionality of the lower dimensional space is determined using the Johnson-Lindenstrauss lemma and an epsilon specifying the amount of tolerated distortion [91]. Using the standard epsilon value of 0.1 and our sample size of 1921 time points, the Johnson-Lindenstrauss lemma outputs a target dimensionality of 6480 projections. These projections are randomly generated as a sparse matrix of nearly orthogonal dimensions. The feature spaces are projected onto this matrix using the dot product, resulting in a 1921 x 6480 dimensional feature space. This pipeline has been used in several recent papers to speed up model fitting and to avoid overfitting [38, 92**?**]. We did not reduce the dimensionality of layers with less than 6480 dimensions because this upsampling could introduce noise and sparsity which might unfairly decrease the encoding model performance for those layers.

To operationalize the language signal, we extracted speech, word, and sentence embeddings from the episode. First, we time-stamped the beginning and end of each word of the transcript using the forced aligner gentle (https://github.com/strob/gentle) as a first pass and then manually checking the alignment using Praat software [93]. Before extracting any embeddings, the British spelling was modified to American spelling to ensure that we could extract embeddings for as many words as possible.

To collect speech embeddings, we input the audio of every sentence into HuBERT, a speech transformer model. Similar to previous work [94], we enforced unidirectional embeddings by extracting representations at the end of the sentence with a stride of 20 ms. We downsampled these embeddings by taking the embeddings every 1.5s (to match our TR).

To collect word-level embeddings, we extracted representations for every time-stamped word from the word2vec model [34]. Word2vec is a shallow, 2-layer neural network trained to predict the context around words. It produces a vector that captures the semantic relationships among the words in the vocabulary. As in prior work [21, 53, 95–98], we down-sampled the embeddings using a 3-lobe Lanzcos filter convolved with a continuous-time representation function that is zero everywhere except for at the middle timestamp of each word, where it is equal to infinitesimal-duration spike. This down-sampling procedure assumes that the neural response is the sum of responses to each word [98].

To collect the sentence-level embeddings, we extracted the representations for every time-stamped sentence of the transcript from every layer of the sentence transformer model sBERT (all-mpnet-base-v2, huggingface.co). This is a pretrained BERT model [99] fine-tuned on 1B sentence pairs using a contrastive objective. Given a sentence, it is trained to predict which other sentence it is semantically paired with. We extracted the activations from all 12 transformer layers. We downsampled each layer's embedding using the same downsampling procedure as the word embeddings, but using the time-stamp of the end of the last word of the sentence. In follow-up analyses, we also extracted embeddings in the same way from GPT-2 [100]

4.7.2 Other automatically extracted and hand-annotated features

We also included automatically extracted features that were previously used to model Sherlock [39]. These included the pitch and amplitude of the episode's audio, the hue and pixel values averaged across the screen, and whether there was a face on screen. For this study, we also extracted motion energy features for every TR using pymoten [30]. This model uses a pyramid of different spatial and temporal Gabor filters at different frequencies, motion directions, xy positions, and sizes to estimate motion energy [29].

We used previously collected [63] binary features of whether music is playing and whether there are written words on the screen. We used valence and arousal ratings for every 4.5 second video segment, which were rated by a combined 113 participants on a 1-9 Likert scale [101]. There were about 55 ratings per video segment. We used the group mean valence and arousal ratings as our valence and arousal features.

We also included social features which were previously generated by two human raters and then averaged to get the final feature values. One feature specified whether a video segment contains social interactions (=1) or not (=0). A social interaction was defined as actions or communication between two or more individuals that are directed at and contingent upon each other. Additional social features included whether a person speaks in the scene (=1) or not (=0) and whether a person infers the mental states of others (=1) or not (=0) [39]. Additionally for this study, two raters counted the number of agents (1,2,3,4+) and whether conversational turn-taking occurred in every TR (=1) or not (=0).

4.7.3 Measuring Feature Space Similarity

We measured the similarity between each feature space extracted from the movie with Canonical Correlation Analysis [35, 36]. Given two feature spaces of any dimension, this analysis finds latent dimensions of each feature space with maximal correlation between them. For each pair of feature spaces, we projected each feature space to 1, 5, 10, or 100 latent dimension(s) that were maximally correlated across both feature spaces. We used L2 regularized CCA as implemented in the CCA-Zoo python library [102], testing 5 log-spaced regularization parameters from 10^-5 to 0 with nested cross validation (5-fold outer loop and 5-fold inner loop). Similar to the subsequent encoding model analyses later on, we grouped the signal into windows of 20 TRs (30 seconds) before splitting into train/test sets in both the outer and inner cross-validation loops to account for temporal autocorrelation in the naturalistic signal. The final similarity score for the pair of feature spaces was the correlation between the projections of the two feature spaces' test data onto their corresponding latent dimensions, averaged across the 5 outer cross validation folds.

4.7.4 Intersubect Correlation Mask

All encoding model analyses were restricted to voxels with reliable stimulus-driven activity, measured using intersubject correlation (ISC), as in prior work [39]. Intersubject correlation measures the shared neural responses across participants, which is taken to be stimulus-driven as long as the neural responses are time-locked to the stimulus [40]. For every voxel within the MNI whole brain mask in each participant, each of the two fMRI BOLD time series runs from Sherlock was denoised by linearly detrending and regressing out the 6 rigid body head motion parameters, framewise displacement, and the first 5 aCompCor components from the combined white matter and CSF masks [80] and the associated discrete cosine bases for high-pass filtering (cutoff of 128s). Each run was then trimmed, and z-scored before being combined into one time series. For each participant, we calculated the Pearson coefficient between their time series and the time series averaged over every other participant. This resulted in an ISC value for every voxel in every participant. As recommended [40], we took the mean of Fisher z-transformed ISC values and then took the inverse Fisher z- transform of the averaged value.

To identify voxels significantly driven by the stimulus at the group level, we ran a sign permutation test with 10,000 permutations. Each participant's ISC values were randomly assigned a positive or negative sign and then the mean ISC was computed. This was repeated 10,000 times to generate the empirical null distribution of ISC values for every voxel. We used the 'permutation_isc' function from the BrainIAK python package to perform the permutation testing. All encoding model analyses were performed only in voxels with significant ISC (FDR corrected p-value<0.001) and correlations greater than 0.001. The resulting ISC mask entirely contains the parcels used to find motion, social interaction, and language selective voxels.

4.7.5 Linking movie features to neural responses with voxel-wise encoding models

We fit a linear mapping between the feature spaces of the naturalistic movie and the voxel-wise fMRI BOLD series within the ISC mask using banded ridge regression, implemented in the himalaya package [41]. Banded ridge regression allows each feature space to learn separate L2-regularization hyperparameters, which better accounts for differently sized feature spaces, prevents overfitting, and is especially useful for analyzing responses to naturalistic stimuli, where feature spaces are correlated [41]. This optimization allows the model to perform feature-space selection, as poorly predictive or redundant feature spaces can be ignored in the model using the regularization parameter. Prior to fitting the encoding models, data from both runs were trimmed to only include the parts showing the episode, and then combined into one time series of 1921 TRs. We used 5-fold nested cross validation to fit the feature weights and select the regularization hyperparameters per feature space per voxel. The himalaya python package was used for fitting this model.

The joint model used all layers from AlexNet, all layers from HuBERT, word2vec, and all layers from sBERT and other previously annotated and automatically extracted features, each as their own feature space.

To account for temporal autocorrelation in the movie and fMRI data, we grouped the signal into windows of 20 TRs (30 seconds) before splitting into train/test sets in both the outer and inner cross-validation loops. To account for hemodynamic delays across cortex, all feature spaces were duplicated with time shifts of 1.5, 3, 4.5, 6, and 7.5 seconds, as in prior work [21, 103]. All feature spaces were z-scored within the train set and the train mean and standard deviation were used to normalize the test set. The fMRI data was also z-scored within the train and test set, to match the scale of the feature spaces. We also included two nuisance regressors representing whether a time point was from the first or second scanning run of the episode, to control for mean differences between the runs. We also added the same fMRIprep confound regressors as used in the GLM analyses (see above). The weights assigned to these regressors were zeroed for model predictions on the held-out test set for every cross-validation fold.

On every fold of the outer loop, the train set went through 5-fold inner loop regularization hyperparameter selection. Candidate hyperparameters (1000, as in prior work [103]) were sampled from a Dirichlet distribution and scaled by 25 log-spaced values between 10^-5 and 10^{10}. The best performing hyperparameters, together with the estimated feature weights, were used to predict the fMRI response in the held-out test set of the outer loop. We measured prediction accuracy using the coefficient of determination R^2, averaging over the 5 folds to get the final value for each voxel in each participant.

4.7.6 Variance Decomposition

We quantified the predictive contribution of each feature space compared to the rest of the feature spaces using the product measure [104, 105], which accounts for the correlation of the feature spaces. We use the product measure instead of the commonly used variance partitioning because it is more efficient for a large number of feature spaces [41]. The model prediction accuracy and product measure for each feature space was averaged across the 5 cross-validation folds.

Group maps of all encoding model results were created by smoothing each individual brain's results with a Gaussian kernel (fwhm=3.0mm) and then averaging each voxel. Statistical testing was performed on each voxel using a signed-rank permutation test, and one sided Benjamini-Hochberg FDR correction with FWER of 0.05. The map was thresholded in volume space before projection to the surface for visualization.

For preference maps, the voxel-wise product measures of each feature were projected to the surface for each participant using the same procedure already described. On the surface level data, we assigned every vertex whichever feature had the highest product measure there.

For the ROI analyses, we averaged the feature product measures of voxels within individually-defined ROIs. If any voxel had a negative joint model performance, we set all feature product measures to zero to ensure we are only examining meaningful product measure values. This was to offset one limitation of the product measure metric, that a product measure can be positive even if the overall model performance is negative [41]. These cases are difficult to interpret so we chose to exclude them.

Statistical testing was performed using linear mixed-effect modeling with the 'Lmer' function from the pymer4 package [86]. To compare the response profiles across regions, we tested for an interaction between feature space and region with participant as a random effect, which allows individual variability and controls for repeated measures. For each region pair within each hemisphere, linear mixed-effect models with and without an interaction term were compared with likelihood ratio tests (LRT). (see formula for model including the interaction term below).

$$\text{Product Measure} \sim 1 + \text{Feature Space} + \text{Region} + \text{Feature Space} \times \text{Region} + (1|\text{Participant})$$

Statistical testing for comparing each product measure to zero was done by regressing the product measure on feature space grouped by participant as a random effect (see formula below). We extracted p-values for each feature space from the model estimates and adjusted them for multiple comparisons with Bonferroni correction within each region (cite).

$$\text{Product Measure} \sim 0 + \text{Feature Space} + (1|\text{Participant})$$

4.7.7 Model-mediated response similarity between regions

We compared the response similarity between regions by computing the pairwise regional correlations using the fitted weights of AlexNet layer 6 (the best performing layer). We do this for each participant and then average across participants for each pairwise region comparison. In this way, we evaluate the response similarity of two regions in the context of a specific feature space. This method is conceptually similar to Model Connectivity [106] and intersubject model-based connectivity [24, 52].

For each hemisphere, we evaluated pairwise response similarity between regions by regressing the response similarity on region pairs grouped by participant as a random effect, which allows individual variability and controls for repeated measures (see formula below). We simplified these pairwise comparisons by including posterior and anterior social interaction regions as one region. We assessed the significance of each pairwise comparison using estimated marginal means (emmeans package) and adjusted the values for multiple comparisons using Bonferroni correction.

$$\text{Product Measure} \sim 1 + \text{Region Pair} + (1|\text{Participant})$$

4.7.8 Cross-subject voxel-wise encoding models

To situate our results in relation to the total amount of explainable variance, we estimated a cross-subject noise ceiling. The movie was presented only once so we could not estimate a within-subjects split-half correlation, Intersubject correlation is also not a suitable noise ceiling because the assumption of anatomical correspondence across different brains can artificially lower the correlations. Instead, for every participant, we used banded ridge regression to predict voxel-wise responses to the naturalistic movie with brain responses from every other participant's brain. We defined two feature spaces, one within the entire ISC mask and one that was anatomically constrained to the union of the left and right MT, STS, and language parcels. The same regression procedure (including nuisance regressors for every participant) as carried out as in the main analysis (see Section 4.7.5). Since our features in this model (other participant brain responses) are not hypothesized to be driving fMRI responses in the target participant (instead they should also be responding in temporally similar ways), we only fit two time delays: 0 and 1.5 seconds. We took the average of the R^2 across the 5 folds as each participant's cross-subject encoding score.

Noise ceilings for ROI analyses were calculated by averaging over the voxel-wise cross-subject encoding performance per participant-defined ROI and then averaging over participants.

4.7.9 AlexNet and sBERT Unit Interpretation

To interpret vision and language predictivity across regions, we fit an encoding model using the AlexNet, motion, HuBERT, word2vec, and sBERT feature spaces, excluding the non-predictive unidimensional feature spaces for use in interpretation later. We then extracted the top 25 most highly weighted AlexNet layer 6 and sBERT layer 10 units from each participant's fitted encoding model. In these exploratory analyses we average the beta weights over left and right and posterior and anterior regions for simplicity before determining the highest weighted units. We then measured the similarity between these units and the unidimensional feature spaces using regularized CCA, with identical methods as before (4.7.3). We obtained a similarity per feature space per region per participant and then averaged over participants.

Statistical testing was performed using linear mixed-effect modeling with the 'Lmer' function from the pymer4 package [86]. To compare the feature tuning across regions, we tested for an interaction between feature space and region with participant as a random effect, which allows individual variability and controls for repeated measures. For each region pair, linear mixed-effect models with and without an interaction term were compared with likelihood ratio tests (LRT). (see formula for model including the interaction term below).

$$\text{Canonical correlation coefficient} \sim 1+\text{Feature Space}+\text{Region}+\text{Feature Space}\times\text{Region}+(1|\text{Participant})$$

In an additional exploratory analysis, we identified the AlexNet units that most consistently occurred in top 2.5% highest weighted units in each region over all the participants (we compiled over temporal and frontal language regions for this analysis). For each of these most consistently highly predictive units, we interpreted their activity in two ways. First, we correlated the selected unit's entire time course during the movie with each of the unidimensional features. Second, we computed a sliding window correlation (window size = 40 TRs) of specific social features of interest and the written text feature. We plotted the portions that were the most highly correlated with each feature of interest over the entire movie, along with the frame that drove the highest unit activation within that window.

References

[1] Hein, G. & Knight, R. T. Superior temporal sulcus–It's my area: or is it? *Journal of Cognitive Neuroscience* **20**, 2125–2136 (2008).

[2] Redcay, E. The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism. *Neuroscience & Biobehavioral Reviews* **32**, 123–142 (2008). URL https://www.sciencedirect.com/science/article/pii/S0149763407000747.

[3] Deen, B., Koldewyn, K., Kanwisher, N. & Saxe, R. Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex (New York, N.Y.: 1991)* **25**, 4596–4609 (2015).

[4] Pitcher, D. & Ungerleider, L. G. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences* **25**, 100–110 (2021).

[5] Isik, L., Koldewyn, K., Beeler, D. & Kanwisher, N. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences* **114**, E9145–E9152 (2017). URL https://www.pnas.org/doi/abs/10.1073/pnas.1714471114.

[6] Walbrin, J., Downing, P. & Koldewyn, K. Neural responses to visually observed social interactions. *Neuropsychologia* **112**, 31–39 (2018). URL https://www.sciencedirect.com/science/article/pii/S0028393218300800.

[7] Walbrin, J., Mihai, I., Landsiedel, J. & Koldewyn, K. Developmental changes in visual responses to social interactions. *Developmental Cognitive Neuroscience* **42**, 100774 (2020). URL https://www.sciencedirect.com/science/article/pii/S1878929320300220.

[8] McMahon, E., Bonner, M. F. & Isik, L. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology* **33**, 5035–5047.e8 (2023). URL https://www.cell.com/current-biology/abstract/S0960-9822(23)01373-8.

[9] Landsiedel, J. & Koldewyn, K. Auditory dyadic interactions through the "eye" of the social brain: How visual is the posterior STS interaction region? *Imaging Neuroscience* **1**, 1–20 (2023). URL https://doi.org/10.1162/imag_a_00003.

[10] Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience* 1–24 (2024). URL https://www.nature.com/articles/s41583-024-00802-4.

[11] Olson, H. A., Chen, E. M., Lydic, K. O. & Saxe, R. R. Left-Hemisphere Cortical Language Regions Respond Equally to Observed Dialogue and Monologue. *Neurobiology of Language* **4**, 575–610 (2023). URL https://doi.org/10.1162/nol_a_00123.

[12] Redcay, E. & Moraczewski, D. Social cognition in context: A naturalistic imaging approach. *NeuroImage* **216**, 116392 (2020).

[13] Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science* **303**, 1634–1640 (2004). URL https://www.science.org/doi/10.1126/science.1089506. Publisher: American Association for the Advancement of Science.

[14] Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **28**, 2539–2550 (2008).

[15] Wu, M. C.-K., David, S. V. & Gallant, J. L. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience* **29**, 477–505 (2006).

[16] Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).

[17] Wehbe, L. *et al.* Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE* **9**, e112575 (2014). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112575.

[18] Dupré la Tour, T., Visconti di Oleggio Castello, M. & Gallant, J. L. The Voxelwise Encoding Model framework: A tutorial introduction to fitting encoding models to fMRI data. *Imaging Neuroscience* **3**, imag_a_00575 (2025). URL https://doi.org/10.1162/imag_a_00575.

[19] Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).

[20] Jung, H. *et al.* Action features dominate cortical representation during natural vision (2025). URL https://www.biorxiv.org/content/10.1101/2025.01.30.635800v1. Pages: 2025.01.30.635800 Section: New Results.

[21] Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016). URL https://www.nature.com/articles/nature17637.

[22] Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**, e2105646118 (2021). URL https://pnas.org/doi/full/10.1073/pnas.2105646118.

[23] Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* **25**, 369–380 (2022). URL https://www.nature.com/articles/s41593-022-01026-4.

[24] Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science* **2**, 745–757 (2022). URL https://www.nature.com/articles/s43588-022-00354-6.

[25] Toneva, M. & Wehbe, L. *Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)*, Vol. 32 (Curran Associates, Inc., 2019). URL https://papers.nips.cc/paper_files/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html.

[26] Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Communications Biology* **5**, 1–10 (2022). URL https://www.nature.com/articles/s42003-022-03036-1.

[27] Tuckute, G. *et al.* Driving and suppressing the human language network using large language models. *Nature Human Behaviour* **8**, 544–561 (2024). URL https://www.nature.com/articles/s41562-023-01783-7.

[28] Goldstein, A. *et al.* A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour* **9**, 1041–1055 (2025). URL https://www.nature.com/articles/s41562-025-02105-9.

[29] Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and Image Science* **2**, 284–299 (1985).

[30] Nunez-Elizalde, A. O., Deniz, F., Dupré la Tour, T., Visconti di Oleggio Castello, M. & Gallant, J. L. pymoten: scientific python package for computing motion energy features from video (2021). URL https://github.com/gallantlab/pymoten.

[31] Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks*, Vol. 25 (Curran Associates, Inc., 2012). URL https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[32] Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* **152**, 184–194 (2017). URL https://www.sciencedirect.com/science/article/pii/S1053811916305481.

[33] Hsu, W.-N. *et al.* HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (2021). URL http://arxiv.org/abs/2106.07447.

[34] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space (2013). URL http://arxiv.org/abs/1301.3781.

[35] Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936). URL https://doi.org/10.1093/biomet/28.3-4.321.

[36] Knapp, T. R. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin* **85**, 410–416 (1978).

[37] Grall, C. & Finn, E. S. Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience* **17**, 598–608 (2022).

[38] Small, H., Masson, H. L., Mostofsky, S. & Isik, L. Fumero, M. *et al.* (eds) *Vision and language representations in multimodal AI models and human social brain regions during natural movie viewing.* (eds Fumero, M. *et al.*) *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, Vol. 285 of *Proceedings of Machine Learning Research*, 69–84 (PMLR, 2024). URL https://proceedings.mlr.press/v285/small24a.html.

[39] Lee Masson, H. & Isik, L. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* 118741 (2021).

[40] Nastase, S. A., Gazzola, V., Hasson, U. & Keysers, C. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience* **14**, 667–685 (2019). URL https://doi.org/10.1093/scan/nsz037.

[41] Dupré La Tour, T., Eickenberg, M., Nunez-Elizalde, A. O. & Gallant, J. L. Feature-space selection with banded ridge regression. *NeuroImage* **264**, 119728 (2022). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811922008497.

[42] Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology* **104**, 1177–1194 (2010). URL https://www.physiology.org/doi/10.1152/jn.00032.2010.

[43] Kumar, S. *et al.* Shared functional specialization in transformer-based language models and the human brain. *Nature Communications* **15**, 5523 (2024). URL https://www.nature.com/articles/s41467-024-49173-5. Publisher: Nature Publishing Group.

[44] Khosla, M., Ngo, G. H., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances* **7**, eabe7547 (2021). URL https://www.science.org/doi/10.1126/sciadv.abe7547. Publisher: American Association for the Advancement of Science.

[45] d'Ascoli, S., Rapin, J., Benchetrit, Y., Banville, H. & King, J.-R. TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction (2025). URL http://arxiv.org/abs/2507.22229. ArXiv:2507.22229 [cs].

[46] Ivanova, A. A. *et al.* The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. *Neurobiology of Language (Cambridge, Mass.)* **2**, 176–201 (2021).

[47] Sueoka, Y. *et al.* The Language Network Reliably "Tracks" Naturalistic Meaningful Nonverbal Stimuli. *Neurobiology of Language* **5**, 385–408 (2024). URL https://doi.org/10.1162/nol_a_00135.

[48] Paunov, A. M. *et al.* Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language* **3**, 413–440 (2022). URL https://doi.org/10.1162/nol_a_00071.

[49] Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *The Journal of Neuroscience* **31**, 2906–2915 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3089381/.

[50] McMahon, E. & Isik, L. Seeing social interactions. *Trends in Cognitive Sciences* **27**, 1165–1179 (2023). URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(23)00248-6. Publisher: Elsevier.

[51] Garcia, K., McMahon, E., Conwell, C., Bonner, M. & Isik, L. *Modeling dynamic social vision highlights gaps between deep learning and humans* (2025). URL https://openreview.net/forum?id=wAXsx2MYgV.

[52] Samara, A., Zada, Z., Vanderwal, T., Hasson, U. & Nastase, S. A. Cortical language areas are coupled via a soft hierarchy of model-based linguistic features (2025). URL https://www.biorxiv.org/content/10.1101/2025.06.02.657491v1. Pages: 2025.06.02.657491 Section: New Results.

[53] Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience* **24**, 1628–1636 (2021). URL https://www.nature.com/articles/s41593-021-00921-6.

[54] Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence* **5**, 1415–1426 (2023). URL https://www.nature.com/articles/s42256-023-00753-y.

[55] Doerig, A. *et al.* Visual representations in the human brain are aligned with large language models (2024). URL http://arxiv.org/abs/2209.11737.

[56] Huh, M., Cheung, B., Wang, T. & Isola, P. The Platonic Representation Hypothesis (2024). URL http://arxiv.org/abs/2405.07987.

[57] Miao, Z. *et al.* Common and Distinct Neural Correlates of Social Interaction Perception and Theory of Mind (2025). URL https://www.biorxiv.org/content/10.1101/2024.12.19.628993v2.

[58] Wolf, I., Dziobek, I. & Heekeren, H. R. Neural correlates of social cognition in naturalistic settings: A model-free analysis approach. *NeuroImage* **49**, 894–904 (2010). URL https://www.sciencedirect.com/science/article/pii/S1053811909009628.

[59] Saxe, R. & Powell, L. J. It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science* **17**, 692–699 (2006). URL https://doi.org/10.1111/j.1467-9280.2006.01768.x.

[60] Jacoby, N., Bruneau, E., Koster-Hale, J. & Saxe, R. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage* **126**, 39–48 (2016). URL https://www.sciencedirect.com/science/article/pii/S1053811915010472.

[61] Ferstl, E. C. & von Cramon, D. Y. What Does the Frontomedian Cortex Contribute to Language Processing: Coherence or Theory of Mind? *NeuroImage* **17**, 1599–1612 (2002). URL https://www.sciencedirect.com/science/article/pii/S1053811902912474.

[62] Lee Masson, H., Chang, L. & Isik, L. Multidimensional neural representations of social features during movie viewing. *Social Cognitive and Affective Neuroscience* **19**, nsae030 (2024).

[63] Chen, J. *et al.* Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience* **20**, 115–125 (2017). URL https://www.nature.com/articles/nn.4450.

[64] Scott, T. L., Gallée, J. & Fedorenko, E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience* **8**, 167–176 (2017). URL https://www.tandfonline.com/doi/full/10.1080/17588928.2016.1201466.

[65] Boré, A., Guay, S., Bedetti, C., Meisler, S. & GuenTher, N. Dcm2Bids (2023). URL https://github.com/UNFmontreal/Dcm2Bids.

[66] Gulban, O. F. *et al.* poldracklab/pydeface: v2.0.0 (2019). URL https://doi.org/10.5281/zenodo.3524401.

[67] Schimke, N. & Hale, J. *Quickshear defacing for neuroimages* (USENIX Association, San Francisco, CA, 2011). URL https://www.usenix.org/conference/healthsec11/quickshear-defacing-neuroimages.

[68] Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage* **20**, 870–888 (2003). URL https://www.sciencedirect.com/science/article/pii/S1053811903003367.

[69] Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* **29**, 1310–1320 (2010).

[70] Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* **12**, 26–41 (2008).

[71] Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging* **20**, 45–57 (2001).

[72] Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).

[73] Klein, A. *et al.* Mindboggling morphometry of human brains. *PLOS Computational Biology* **13**, e1005350 (2017). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005350. Publisher: Public Library of Science.

[74] Fonov, V. *et al.* Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage* **54**, 313–327 (2011).

[75] Evans, A. C., Janke, A. L., Collins, D. L. & Baillet, S. Brain templates and atlases. *NeuroImage* **62**, 911–922 (2012). URL https://www.sciencedirect.com/science/article/pii/S1053811912000419.

[76] Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**, 825–841 (2002).

[77] Cox, R. W. & Hyde, J. S. Software tools for analysis and visualization of fMRI data. *NMR in biomedicine* **10**, 171–178 (1997).

[78] Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).

[79] Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).

[80] Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).

[81] Satterthwaite, T. D. *et al.* An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage* **64**, 240–256 (2013).

[82] Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013).

[83] Lanczos, C. A Precision Approximation of the Gamma Function. *SIAM Journal on Numerical Analysis* **1**, 86–96 (1964).

[84] Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics* **8** (2014). URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2014.00014/full.

[85] Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945). URL https://onlinelibrary.wiley.com/doi/abs/10.2307/1932409. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1932409.

[86] Jolly, E. Pymer4: Connecting R and Python for Linear Mixed Modeling. *Journal of Open Source Software* **3**, 862 (2018). URL https://joss.theoj.org/papers/10.21105/joss.00862.

[87] Thomee, B. *et al.* YFCC100M: The New Data in Multimedia Research. *Communications of the ACM* **59**, 64–73 (2016). URL http://arxiv.org/abs/1503.01817. ArXiv:1503.01817 [cs].

[88] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A Simple Framework for Contrastive Learning of Visual Representations*, 1597–1607 (PMLR, 2020). URL https://proceedings.mlr.press/v119/chen20j.html. ISSN: 2640-3498.

[89] Mu, N., Kirillov, A., Wagner, D. & Xie, S. *SLIP: Self-supervision Meets Language-Image Pre-training*, 529–544 (Springer-Verlag, Berlin, Heidelberg, 2022). URL https://doi.org/10.1007/978-3-031-19809-0_30.

[90] Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications* **15**, 9383 (2024). URL https://www.nature.com/articles/s41467-024-53147-y. Publisher: Nature Publishing Group.

[91] Achlioptas, D. *Database-friendly random projections*, PODS '01, 274–281 (Association for Computing Machinery, New York, NY, USA, 2001). URL https://doi.org/10.1145/375551.375608.

[92] Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? (2022). URL http://biorxiv.org/lookup/doi/10.1101/2022.03.28.485868.

[93] Boersma, P. & Weenink, D. PRAAT, a system for doing phonetics by computer. *Glot international* **5**, 341–345 (2001).

[94] Vaidya, A. R., Jain, S. & Huth, A. G. Self-supervised models of audio effectively explain human cortical responses to speech (2022). URL http://arxiv.org/abs/2205.14252. ArXiv:2205.14252 [cs].

[95] Tang, J., Du, M., Vo, V., LAL, V. & Huth, A. Oh, A. *et al.* (eds) *Brain encoding models based on multimodal transformers can transfer across language and vision.* (eds Oh, A. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 36, 29654–29666 (Curran Associates, Inc., 2023). URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5ebbbac62b968254093023f1c95015d3-Paper-Conference.pdf.

[96] LeBel, A., Jain, S. & Huth, A. G. Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *The Journal of Neuroscience* **41**, 10341–10355 (2021). URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0118-21.2021.

[97] Gong, X. L. *et al.* Phonemic segmentation of narrative speech in human cerebral cortex. *Nature Communications* **14**, 4309 (2023). URL https://www.nature.com/articles/s41467-023-39872-w.

[98] LeBel, A. *et al.* A natural language fMRI dataset for voxelwise encoding models. *Scientific Data* **10**, 555 (2023). URL https://www.nature.com/articles/s41597-023-02437-z. Publisher: Nature Publishing Group.

[99] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). URL http://arxiv.org/abs/1810.04805. ArXiv:1810.04805 [cs].

[100] Radford, A. *et al.* *Language Models are Unsupervised Multitask Learners* (2019). URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

[101] Kim, J. *et al.* A study in affect: Predicting valence from fMRI data. *Neuropsychologia* **143**, 107473 (2020). URL https://www.sciencedirect.com/science/article/pii/S0028393220301445.

[102] Chapman, J. & Wang, H.-T. CCA-Zoo: A collection of Regularized, Deep Learning based, Kernel, and Probabilistic CCA methods in a scikit-learn style framework. *Journal of Open Source Software* **6**, 3823 (2021). URL https://joss.theoj.org/papers/10.21105/joss.03823.

[103] Deniz, F., Tseng, C., Wehbe, L., Dupré La Tour, T. & Gallant, J. L. Semantic Representations during Language Comprehension Are Affected by Context. *The Journal of Neuroscience* **43**, 3144–3158 (2023). URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2459-21.2023.

[104] Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin* **57**, 116–131 (1960). Place: US Publisher: American Psychological Association.

[105] Jw, P. Dividing the indivisible : using simple symmetry to partition variance explained. *Proceedings of the second international Tampere conference in statistics, 1987* 245–260 (1987). Publisher: Department of Mathematical Sciences, University of Tampere.

[106] Meschke, E. X., Castello, M. V. d. O., Tour, T. D. l. & Gallant, J. L. Model connectivity: leveraging the power of encoding models to overcome the limitations of functional connectivity (2023). URL https://www.biorxiv.org/content/10.1101/2023.07.17.549356v1. Pages: 2023.07.17.549356 Section: New Results.

5 Supplementary Information

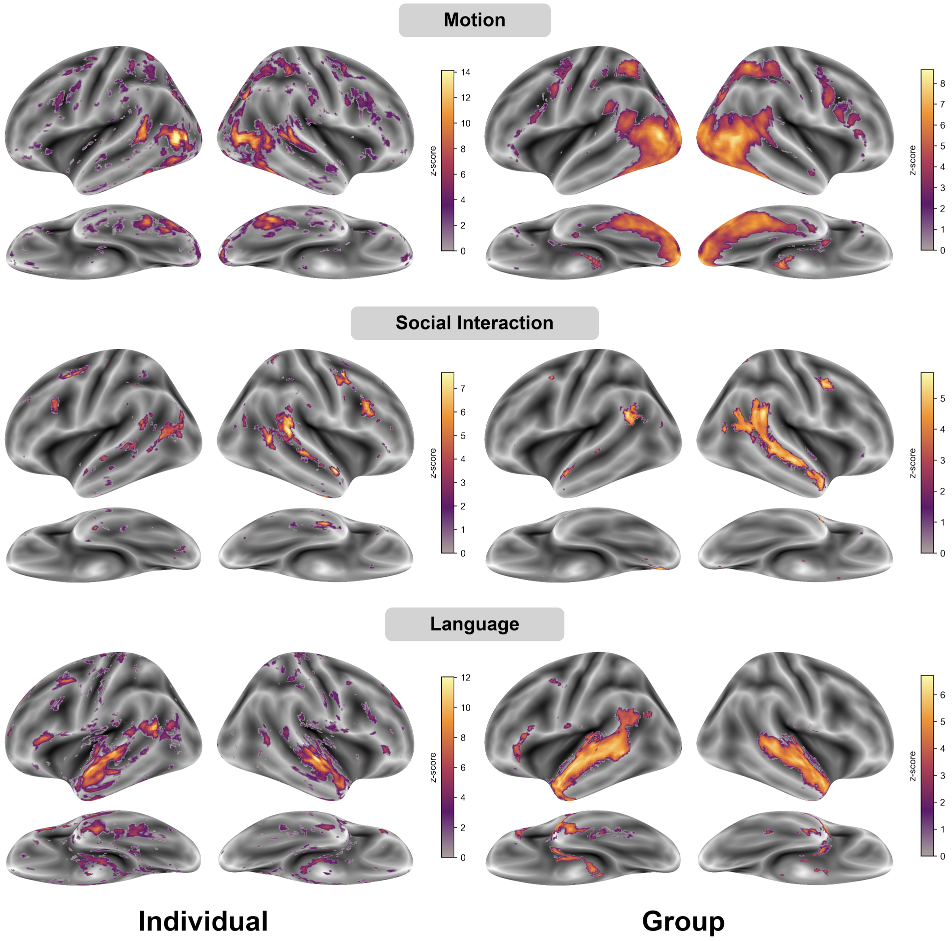


Fig. 1 Left: A representative participant and Right: group maps of contrasts used to identify motion (interacting & non-interacting point-lights), social interaction perception (interacting - non-interacting point-lights), and language (intact-degraded speech) regions. All maps are thresholded to FDR q<0.01.

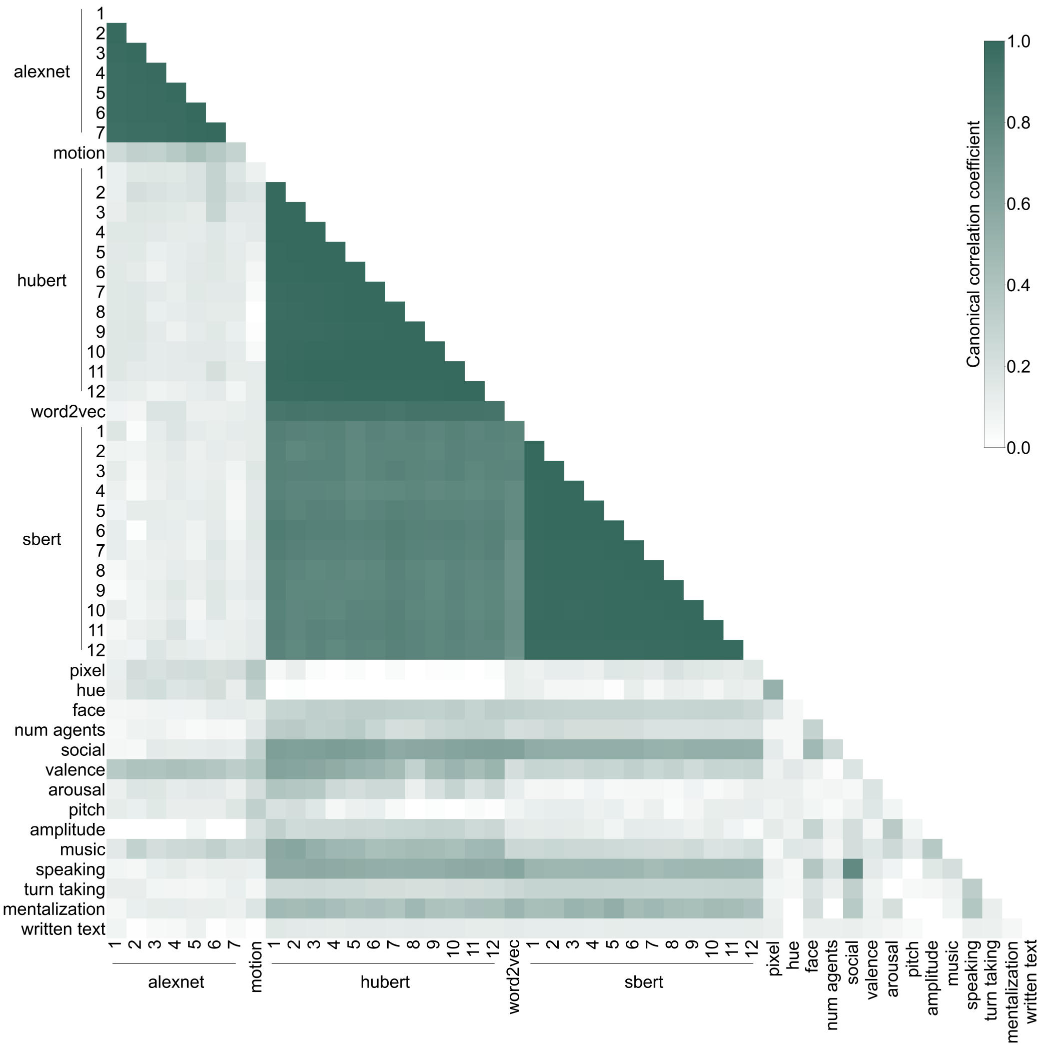


Fig. 2 Similarity between all feature spaces used in the joint encoding model. Each square shows the Canonical correlation coefficient of a pair of feature spaces, keeping all layers of AlexNet, HuBERT, and sBERT separate (see numbered layers on matrix). For each pair of feature spaces, we projected each feature space to 1 latent dimension that was maximally correlated across both feature spaces. See Main Text Methods Section 4.7.3 for more details.

Hemisphere	Region Pair	Df	Chi-Sq	p-value	Sig.
Left	MT vs STS	18	66.039	2.5e-06	***
Left	MT vs frontal	18	29.152	0.559	n.s.
Left	MT vs temporal	18	148.732	1.6e-21	***
Left	STS vs frontal	18	10.102	1	n.s.
Left	STS vs temporal	18	35.463	0.099	n.s.
Left	temporal vs frontal	18	64.195	5.1e-06	***
Right	MT vs STS	18	28.939	0.59	n.s.
Right	MT vs frontal	18	23.365	1	n.s.
Right	MT vs temporal	18	135.64	5.3e-19	***
Right	STS vs frontal	18	10.792	1	n.s.
Right	STS vs temporal	18	57.91	5.3e-05	***
Right	temporal vs frontal	18	55.656	0.00012	***

Table 1 Pairwise ANOVA comparisons of patterns of feature product measures in each region. The pairwise comparisons are done within each hemisphere and are the p-values have been adjusted for multiple comparisons using Bonferroni correction.

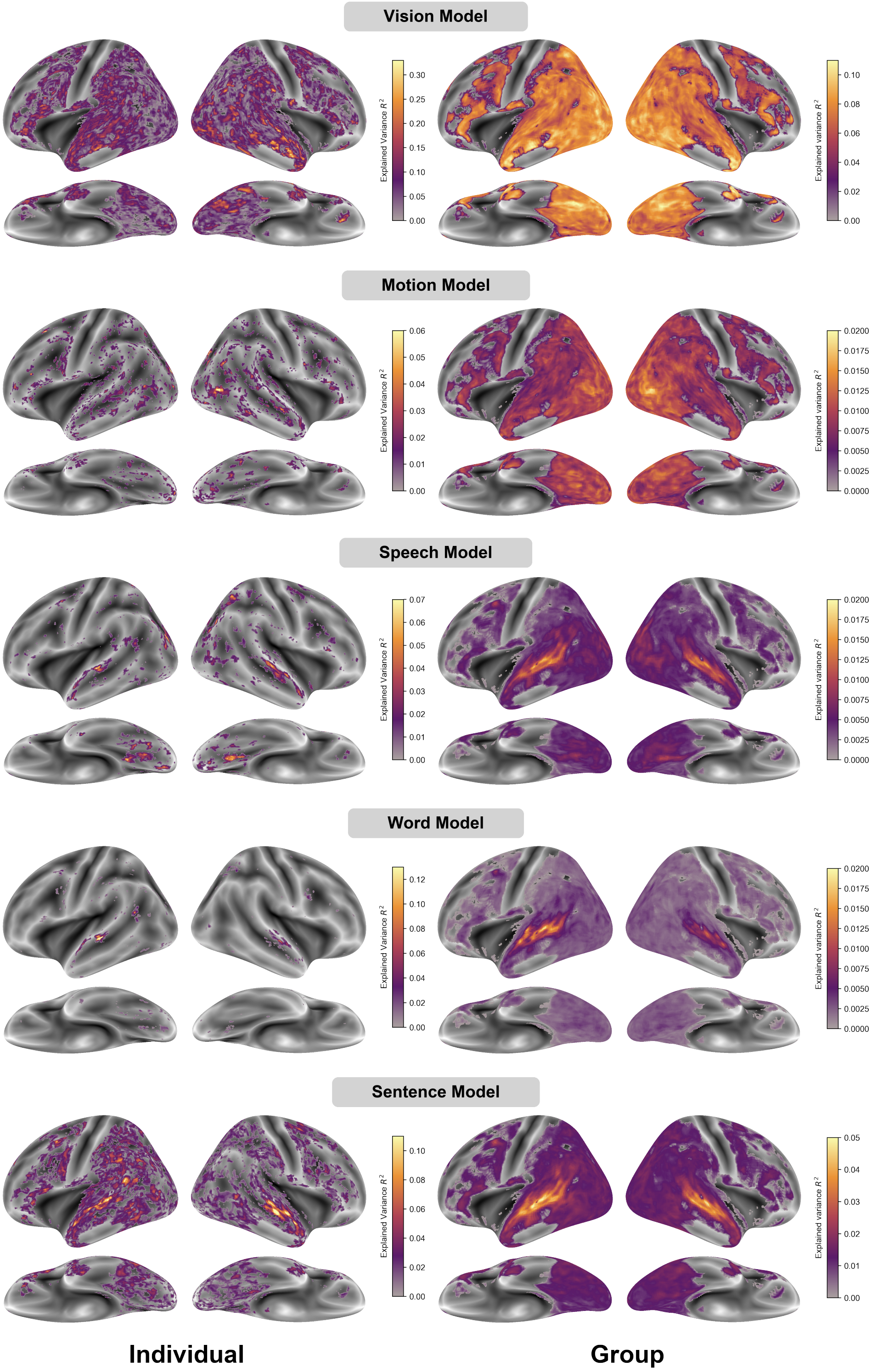
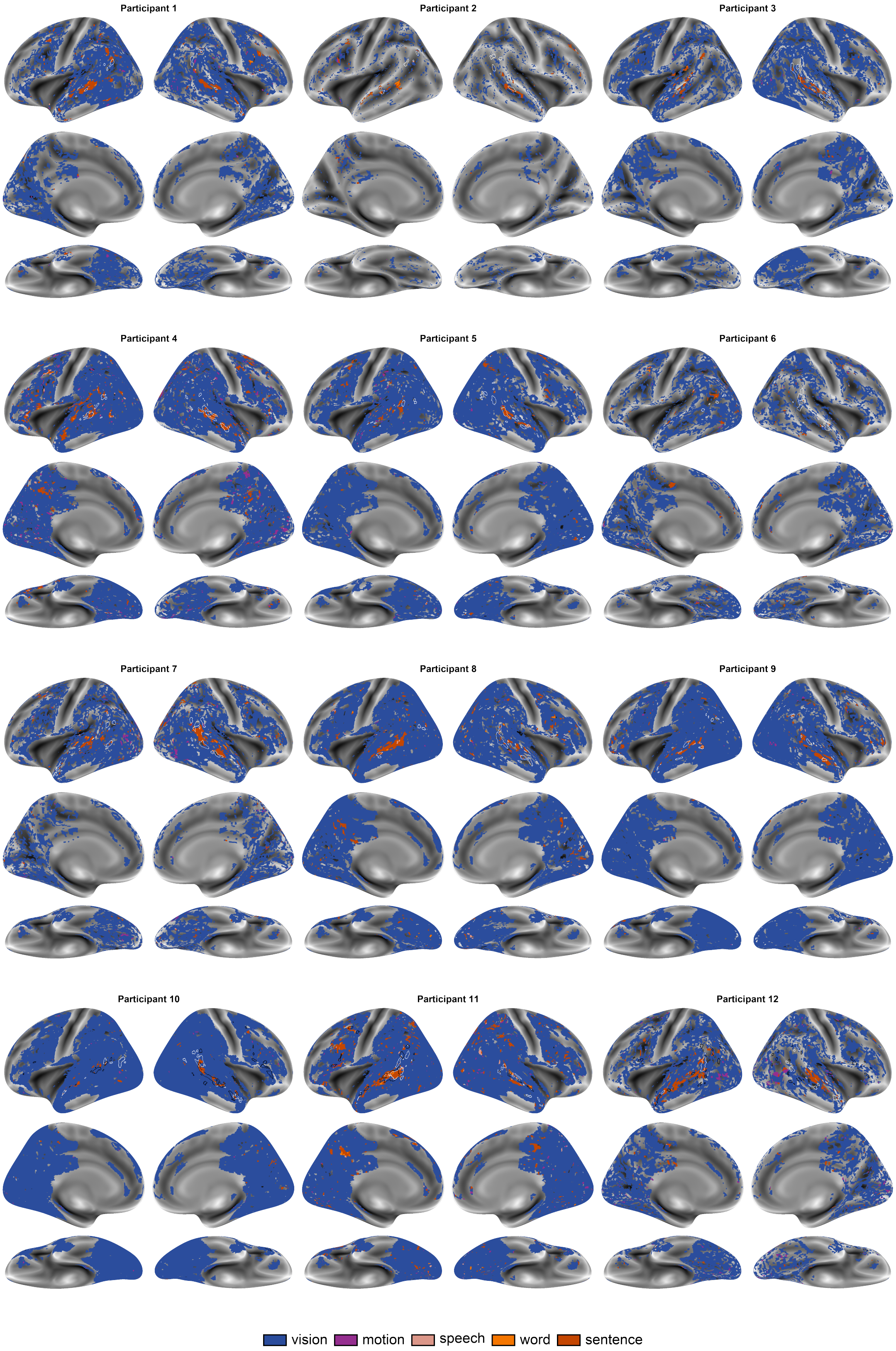
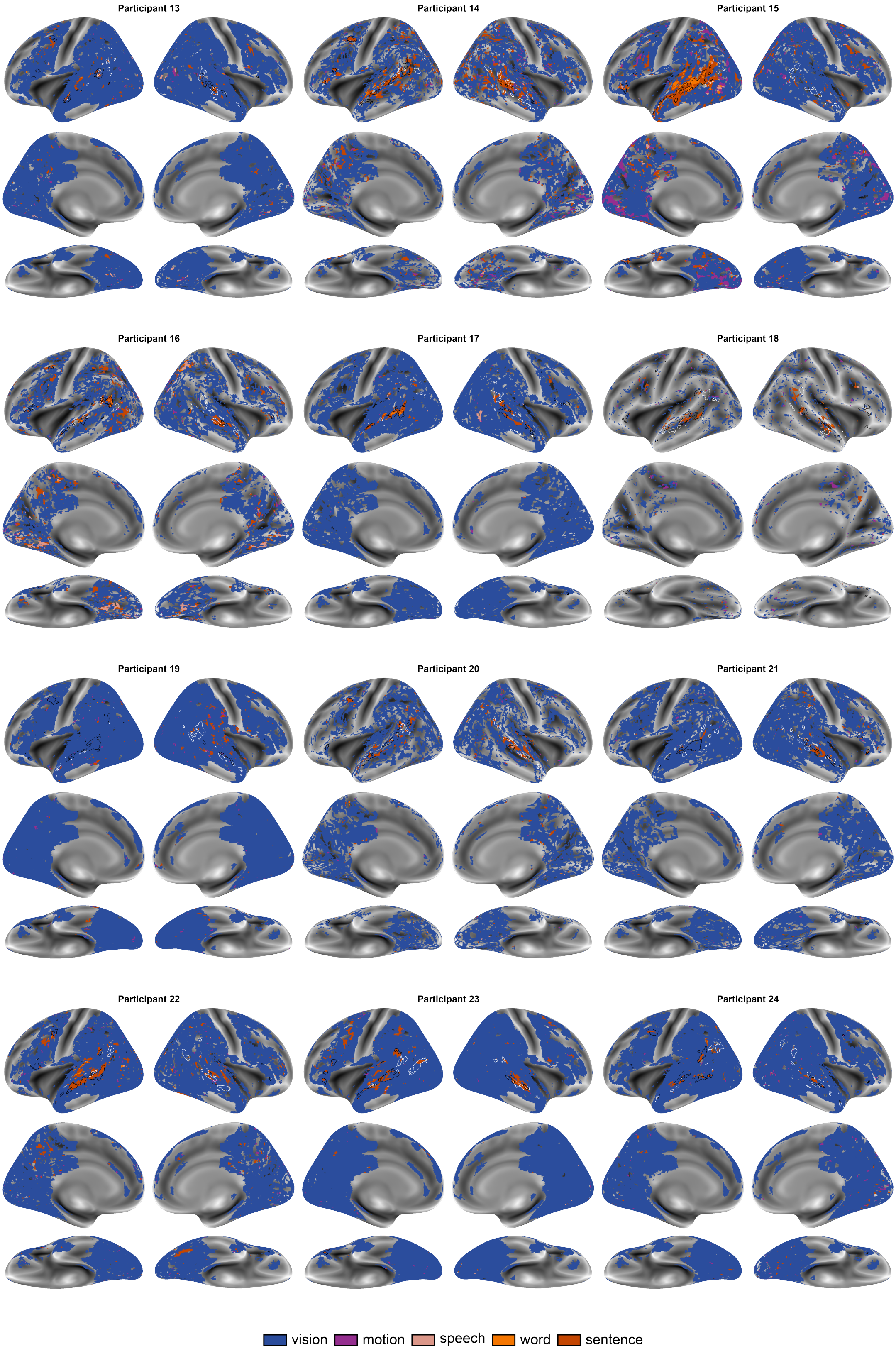


Fig. 3 Left: Representative individual product measure maps for the vision (AlexNet), motion, speech (HuBERT), word (word2vec), and sentence (sBERT) features. The maps are thresholded to $R^2 > 0$. Right: Group averaged product measure maps for each feature. These are thresholded to FDR q<0.001 (signed permutation test).


Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
Participant 7
Participant 8
Participant 9
Participant 10
Participant 11
Participant 12
vision
motion
speech
word
sentence


Participant 13
Participant 14
Participant 15
Participant 16
Participant 17
Participant 18
Participant 19
Participant 20
Participant 21
Participant 22
Participant 23
Participant 24
vision
motion
speech
word
sentence

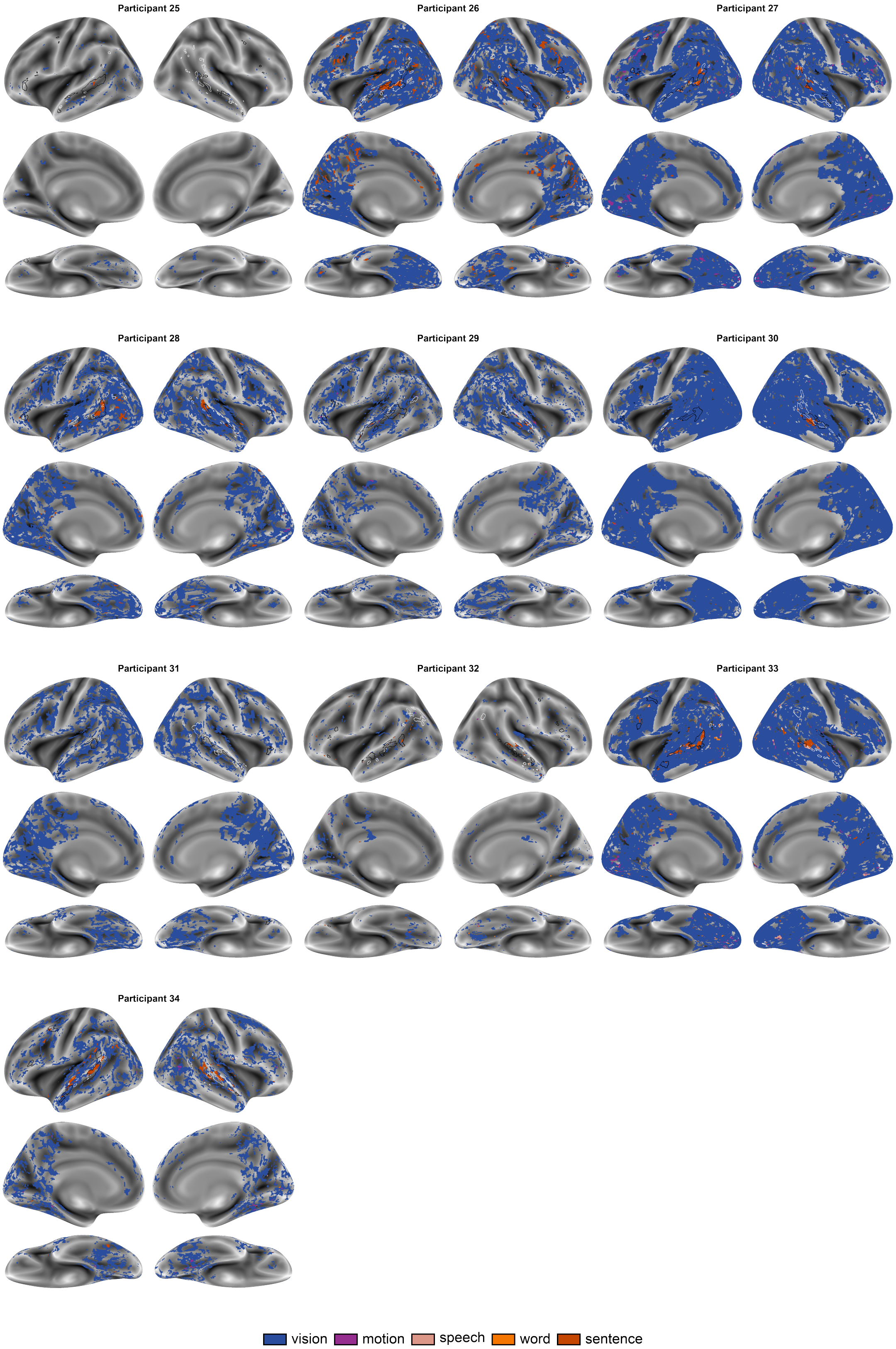


Fig. 4 Vertex-wise preference maps of feature predictivity of all participants. Individually localized social interaction (white) and language (black) regions are overlaid. All preference maps are thresholded to only include vertices with positive joint encoding model performance.

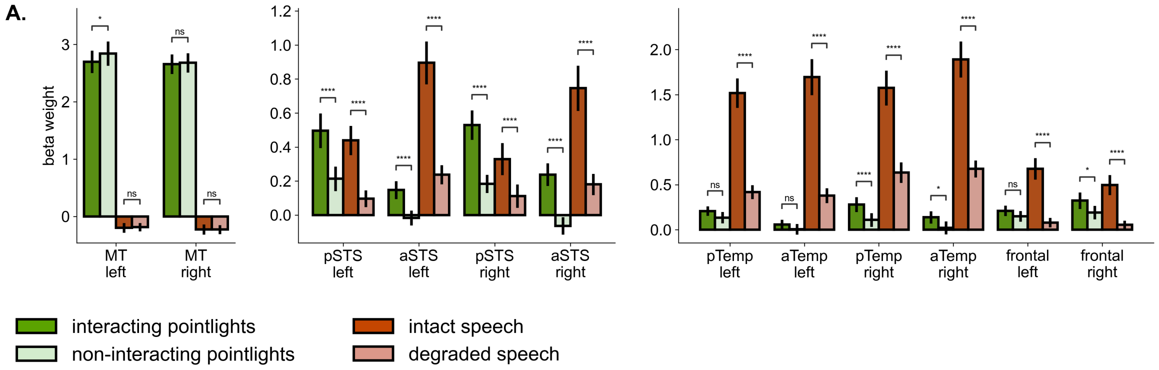


Fig. 5 Responses to the conditions of localizer experiments in motion, social interaction, and language regions. Asterisks indicate a statistical difference in response (*=p<0.05,**=p<0.01,***=p<0.001,****=p<0.0001). Although some functional overlap is evident, the social interaction selective regions are significantly more selective for social interaction than the language regions are (LME p<0.001) and language regions are significantly more selective for language than the social interaction regions are (LME p<0.001).

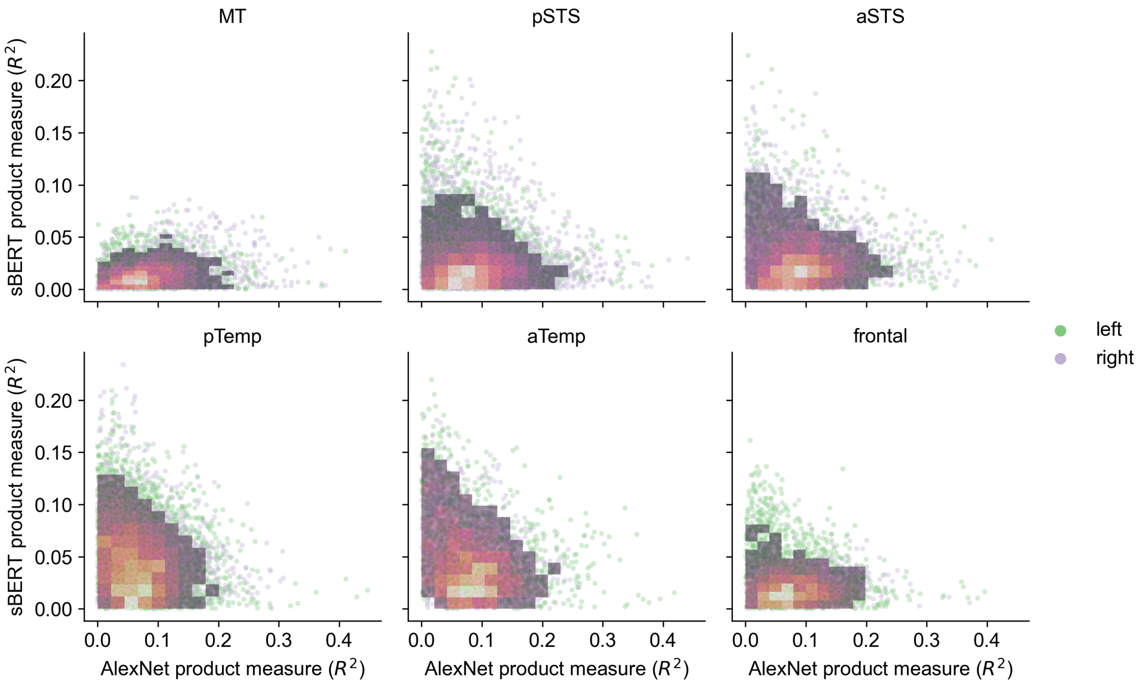


Fig. 6 Voxel-wise vision (AlexNet) and sentence (sBERT) product measures in localized motion (MT), social interaction (posterior and anterior STS), and language (posterior and anterior temporal and frontal). Each point represents the AlexNet and sBERT product measure of one voxel from one participant in a particular region (voxels from left hemisphere are in green, voxels from right hemisphere are in purple). The overlaid bivariate histogram (20 bins, proportion threshold of 0.1) shows the areas with the most voxels.

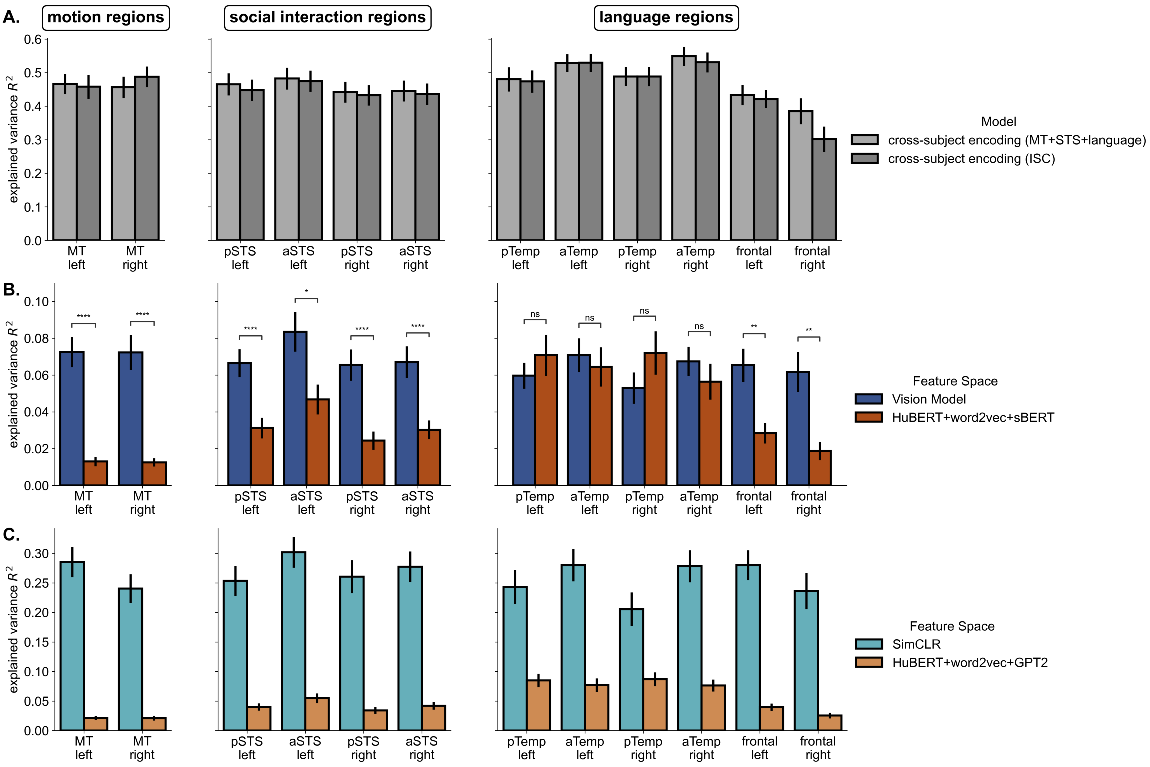


Fig. 7 A. Cross-subject encoding model performance. Light gray bars show performance of a model using participant's neural responses within the ISC mask to predict each other and gray bars show performance of a model using participant neural responses within a combined MT, STS, and language parcel to predict each other. B. Vision and language predictivity in motion, social interaction and language regions in the joint encoding model. The vision predictivity is the summation of AlexNet and motion product measures and the language predictivity is the summation of HuBert, word2vec, and sBERT product measures. C. Vision and language predictivity in motion, social interaction and language regions in an encoding model fit using more advanced DNNs. The vision predictivity is the SimCLR product measure (all layers added together) and the language predictivity is the summation of HuBert, word2vec, and GPT-2 product measures.

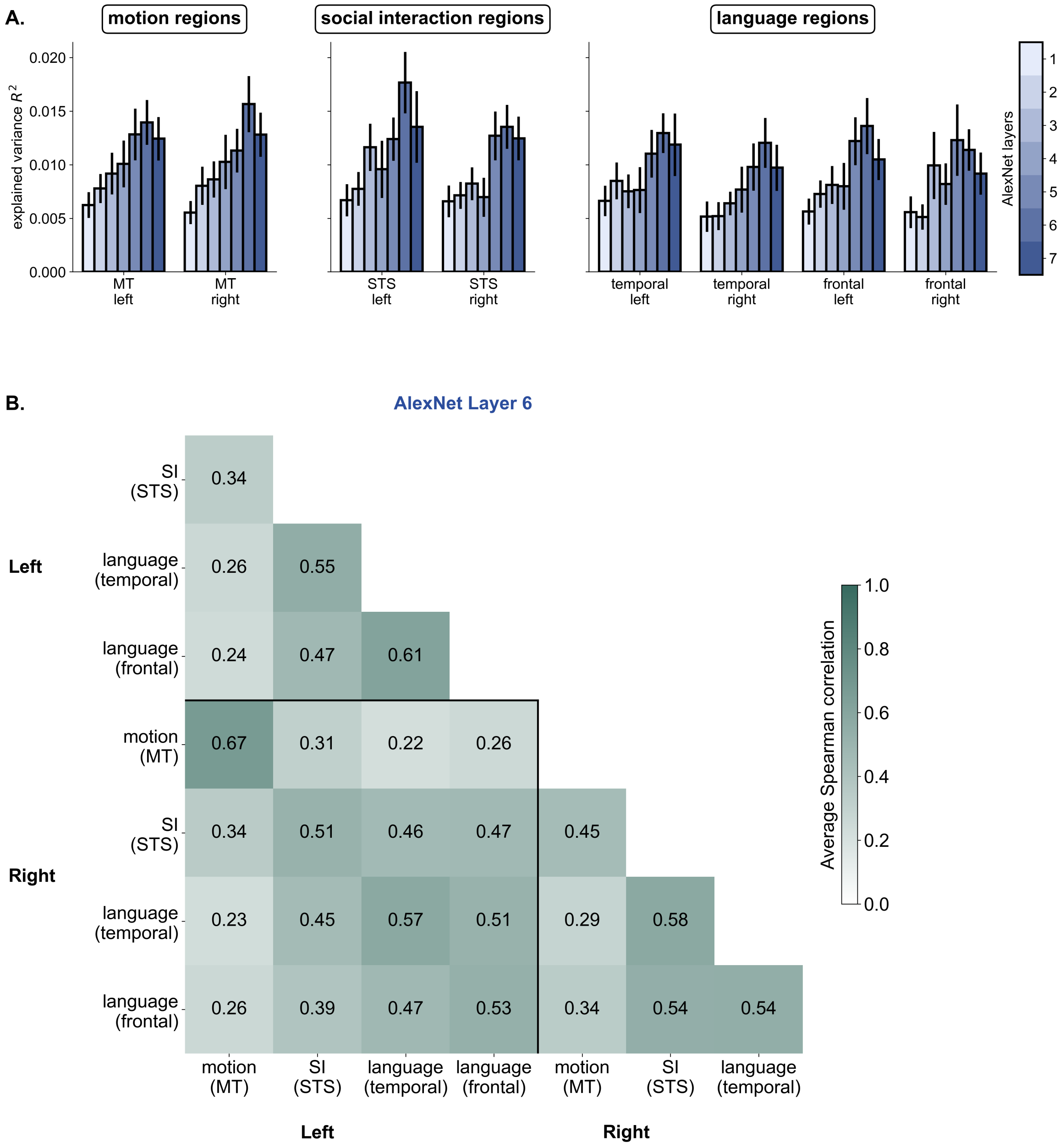


Fig. 8 A. Average product measure of each layer of AlexNet from the joint model. Posterior and anterior regions are averaged per participant before averaging across the group. B. Pairwise correlations of beta weights from AlexNet layer 6, the highest performing AlexNet layer in the joint encoding model. Posterior and anterior regions are averaged per participant before averaging across the group.

Table 2 Pairwise comparisons of region AlexNet layer-wise product measure. The pairwise comparisons are done within each hemisphere and are the p-values have been adjusted for multiple comparisons using Bonferroni correction.

Hemisphere	Region Pair 1	Region Pair 2	Estimate	p-value	Sig.
Left	MT-SI	MT-frontLang	0.114	0.189	n.s.
Left	MT-SI	MT-tempLang	0.098	0.124	n.s.
Left	MT-SI	SI-frontLang	-0.114	0.032	*
Left	MT-SI	SI-tempLang	-0.193	3.9e-08	***
Left	MT-SI	frontLang-tempLang	-0.257	2e-10	***
Left	MT-frontLang	MT-tempLang	-0.016	1	n.s.
Left	MT-frontLang	SI-frontLang	-0.228	2.2e-05	***
Left	MT-frontLang	SI-tempLang	-0.307	5.1e-11	***
Left	MT-frontLang	frontLang-tempLang	-0.371	3.4e-13	***
Left	MT-tempLang	SI-frontLang	-0.211	7.8e-07	***
Left	MT-tempLang	SI-tempLang	-0.29	1e-15	***
Left	MT-tempLang	frontLang-tempLang	-0.355	2.2e-17	***
Left	SI-frontLang	SI-tempLang	-0.079	0.247	n.s.
Left	SI-frontLang	frontLang-tempLang	-0.144	0.003	**
Left	SI-tempLang	frontLang-tempLang	-0.065	0.731	n.s.
Right	MT-SI	MT-frontLang	0.12	0.169	n.s.
Right	MT-SI	MT-tempLang	0.165	9.3e-05	***
Right	MT-SI	SI-frontLang	-0.076	0.647	n.s.
Right	MT-SI	SI-tempLang	-0.125	0.00083	***
Right	MT-SI	frontLang-tempLang	-0.073	0.834	n.s.
Right	MT-frontLang	MT-tempLang	0.045	1	n.s.
Right	MT-frontLang	SI-frontLang	-0.197	0.001	**
Right	MT-frontLang	SI-tempLang	-0.245	1.5e-06	***
Right	MT-frontLang	frontLang-tempLang	-0.194	0.002	**
Right	MT-tempLang	SI-frontLang	-0.241	4.5e-08	***
Right	MT-tempLang	SI-tempLang	-0.29	1.2e-15	***
Right	MT-tempLang	frontLang-tempLang	-0.238	1e-07	***
Right	SI-frontLang	SI-tempLang	-0.048	1	n.s.
Right	SI-frontLang	frontLang-tempLang	0.003	1	n.s.
Right	SI-tempLang	frontLang-tempLang	0.051	1	n.s.