

Reviewed Preprint

v1 • October 26, 2026

Not revised

For correspondence:

lisik@jhu.edu

Funding: See [page 22](#)

Reviewing editor: Iris Groen,
University of Amsterdam,
Netherlands

© 2026, McMahon et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

A Spatiotemporal Hierarchy for Social Interaction Perception in the Lateral Visual Stream

Emalie McMahon^{1,2}, Elizabeth Jiwon Im³, Michael F Bonner¹, Leyla Isik¹✉

¹Johns Hopkins University, Baltimore, United States • ²Massachusetts Institute of Technology, Cambridge, United States

• ³Stanford University, Stanford, United States

eLife Assessment

This manuscript seeks to understand the hierarchical temporal representation of social actions in the lateral visual pathway in the human brain. It introduces a potentially **valuable** methodological approach for characterizing the spatiotemporal dynamics of social-interaction representations, with the data supporting the hypothesis that the anterior portion of the superior temporal sulcus responds to communicative actions, and providing a temporal estimate of when it begins responding. However, the failure to recover several previously reported effects regarding mid-level representations in this and other visual brain regions, combined with unresolved concerns about the correspondence between EEG, fMRI, and stimulus-feature measures, limits confidence in the method's sensitivity and validity, rendering the current empirical evidence against a hierarchical feedforward architecture **incomplete**.

<https://doi.org/10.7554/eLife.112352.1.sa3>

Abstract

The lateral visual stream has been recently proposed as a third visual stream, in addition to the ventral and dorsal streams, specialized for processing dynamic social content. While prior work has suggested that the regions of this pathway form a hierarchy representing increasingly abstract information, the computations along this pathway are still largely unknown. High spatiotemporal resolution data are particularly informative for characterizing the information flow and thus neural computations across different brain regions. Using a novel regression approach, we combine data from EEG, fMRI, and behavior in response to the same videos to leverage the high temporal resolution of EEG and whole-brain spatial resolution of fMRI. We find that low-level visual features are represented in early visual cortex with a short temporal latency and are not represented in higher-level regions of the lateral stream. Further, we find that mid-level features are represented in mid-level lateral regions with a shorter latency than high-level features in more anterior regions of the lateral pathway. However, both mid- and high-level features were decodable in anterior regions of the lateral pathway with a similar latency. Together, these results provide evidence that features of social actions are processed rapidly in the lateral visual stream in a manner that is consistent with hierarchical processing, but the lateral stream does not exhibit a strict temporal sequence of representational transformations along the posterior-to-anterior axis.

1 Significance Statement

Understanding how the brain processes social actions is critical to understanding the social mind. This study leverages a novel EEG-fMRI fusion method to reveal, for the first time, the spatiotemporal hierarchy of social action representations in the lateral visual pathway from early visual cortex to the superior temporal sulcus. Notably, we can decode communicative interactions

in the anterior superior temporal sulcus (aSTS) within 200 ms shortly following decoding of visuospatial features about the social scene. This finding underscores the brain's remarkable efficiency in extracting social meaning. Together, these results advance our understanding of the neural computations underlying social perception and highlight the lateral visual stream as a specialized pathway for dynamic social perception.

2 Introduction

Perceiving the actions of others is essential in daily life. Among the most common actions we witness are social actions that involve two or more people, like talking, dancing, or gesturing (1; 2; 3). The brain primarily represents actions of others in lateral regions including the lateral occipital temporal cortex (LOTc) and the superior temporal sulcus (STS) (2; 3). In these regions, one of the key organizing features of actions is their sociality, or the extent to which an action is directed at another person (2; 3; 4). Other related work has identified regions in the STS that respond selectively to social interactions (5; 6; 7; 8). These findings, combined with research on dynamic face processing, have led to new ideas about a third visual pathway projecting laterally from primary visual cortex, through LOTc, to the STS (9). The lateral visual pathway is hypothesized to be separate from the classic ventral and dorsal visual pathways (10; 11) and to be specialized for understanding dynamic social content.

Features of social interactions are hypothesized to be hierarchically processed along the lateral visual pathway (12). This is supported by studies with simple, controlled stimuli showing that there are increasingly abstract social action representations along the lateral surface. For instance, mid-level visual cues indicative of social actions, or “social primitives”, such as whether two bodies are facing (13; 14) or moving toward one another (15) are represented in the body-selective extrastriate body area (EBA), a mid-level region in the lateral pathway. More anterior regions along the STS are selective for more abstract, high-level social action information, including the presence and valence of social interactions (5; 6; 7; 8). Further, a recent comprehensive fMRI study investigated the organization of low-to-high level social features. This work found a posterior-to-anterior gradient of increasingly abstract social feature representations in the lateral visual pathway (6). Related work has found a temporal hierarchy of action representations using electroencephalography (EEG) (1) and magnetoencephalography with source-reconstruction (16).

While the prior research provides evidence for hierarchical representations of social actions, none of these studies have been able to address how information flows through these lateral visual regions due to the use of methods that either have high temporal or high spatial resolution but not both. To investigate the question of how representations in these regions evolve over time, in the current study, we use EEG-fMRI fusion to combine the high-temporal resolution of EEG with the high-spatial resolution of fMRI. EEG-fMRI fusion is a previously developed method (17) that has been used to investigate how object (18) and social representations (19) evolve in space and time in the human brain. Here, we make several significant methodological advances to the original EEG-fMRI fusion approach. We adopt an encoding model approach that (1) links EEG and fMRI in a cross-validated manner allowing us to determine the generalizability of representations in held-out stimuli, and (2) defines a hypothesis space of the tuning profile of different regions of interest (ROIs) to stimulus features and to EEG at each time point (20; 21).

We combine our previously published fMRI data (6) with new EEG recordings in separate subjects while they watched the same video clips as the fMRI participants (Figure 1A). Using EEG decoding and EEG-fMRI fusion, our results support a hierarchy of features in computing social interactions but also suggest that the lateral stream may not be a strict feedforward hierarchy.

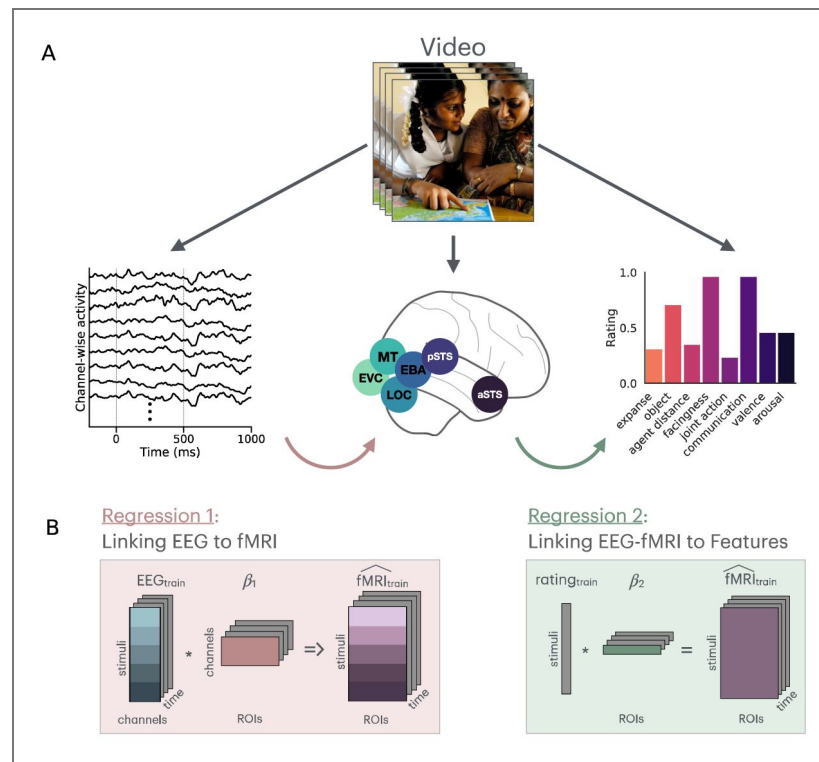


Figure 1. Data and method overview

A) One example video depicting two person actions (top). Participants viewed the videos while their neural responses were recorded with EEG (middle-left) or fMRI (middle-middle), and participants online provided behavioral ratings for eight different features of the visual and social scene. The EEG data is an example of activity in 10 channels of the 64 channels used in the regression, the brain depicts a schematic of the lateral stream ROIs that were predicted from EEG, and the behavioral ratings are example ratings for one video. **B)** Two-step regression schematic. *Regression 1* is fit between the EEG and fMRI data at each time point using 5-fold cross-validation (pink arrow in A). The color bands in the fMRI and EEG data represent the cross-validation folds. The result of this step is the predicted fMRI response for all stimuli in the training data. *Regression 2* is fit between the EEG-predicted fMRI response and the behavioral ratings of the stimuli (green arrow in A). Because of license restrictions, the image in (A) is only representative of a video in the stimulus set. The image is “Parents and kids learn together” by DFID - UK Department for International Development licensed under CC BY 3.0.

3 Results

3.1 Social Action Videos Evoke Reliable EEG Responses

A group of participants ($n = 20$) participated in an EEG experiment in which their neural activity was recorded with a 64-channel EEG, while they viewed short videos that were used in our prior study of social action perception (6). This stimulus set comprises 250 three-second videos depicting social actions, which are split into 200 videos for training and 50 videos for evaluation. Here we maintained the defined train-test split for the stimuli but reduced the videos to the central 500 ms to facilitate temporal time-locking in EEG. In the EEG experiment, videos in the training set were repeated four times, and videos in the test set were repeated sixteen times to ensure a high signal-to-noise ratio of the target data.

After minimal preprocessing, temporal resampling of the EEG data to 400 Hz, and temporal smoothing (see 5.4.1 (4)), we estimated the signal quality of EEG data by calculating the split-half reliability in the test set. Overall, the reliability is significant in the 72–438, 458–622, and 658–800 ms time ranges, revealing that there is detectable signal in EEG (Figure S1 (4)) and that the data are of high enough quality for the encoding/decoding framework adopted below.

3.2 Early-Late Temporal Dissociation in the Lateral Pathway

To understand the time course of neural processing across the lateral stream, we used an encoding approach to fuse the EEG data with our previously collected fMRI data (6). At each time point for each EEG participant, we predicted the average ROI response in fMRI data from each subject. Our fMRI dataset (6) includes voxel-wise fMRI neural responses (β values) to each video from four participants. The fMRI data includes both anatomically defined regions of interest (ROIs), such as the early visual cortex (EVC) and the motion-selective middle temporal area (MT), as well as functionally defined ROIs in the lateral stream. The lateral ROIs consist of the extrastriate body area (EBA), which processes bodies and their spatial relationships (22; 13), the lateral occipital cortex (LOC), which is selective for objects (23) and is involved in processing object-directed actions (3), and posterior and anterior social-interaction selective regions in the superior temporal sulcus (pSTS and aSTS) (5; 8; 7; 6).

Following the encoding, we averaged the prediction across EEG participants and fMRI participants. Our results show a dissociation between early activation in the most posterior region (EVC: $t_{onset} = 60$ ms) and late activation in all other lateral regions of the brain (MT: $t_{onset} = 310$ ms, LOC: $t_{onset} = 130$ ms, EBA: $t_{onset} = 112$ ms, pSTS-SI: $t_{onset} = 125$ ms, and aSTS-SI: $t_{onset} = 115$ ms, Figure 2 (4)). The onset in MT is surprisingly late, and MT has low overall prediction, pointing to low correspondence between EEG responses and fMRI responses in MT.

From these results, we do not find the classical latency profile of feedforward hierarchical computations (24; 9) as mid-level regions (EBA and LOC) are significantly predicted at the same time as anterior regions in the STS, suggesting information in mid- and high-level regions comes online at the same time.

Latency estimations are known to be influenced by the overall decodability of information (e.g., the higher decoding in EVC compared to other regions). We performed an additional analysis to increase our statistical power and minimize bias due to overall decodability. To do this, we binned the decoding performance, variance distribution, and permutation distribution within-participant in 50 ms bins from 0 to 300 ms. We then averaged each distribution across participants. These results mirror the time-resolved analysis, with EVC first decodable in the 50–100 ms interval, and all other ROIs decodable by the 100–150 ms interval, confirming this early-late distinction (Figure S2 (4)).

We see similar trends in whole-brain voxelwise encoding analyses, with high EVC prediction by 100 ms, and onset of prediction in all other regions by 150 ms (Figure 3 (4) and Figure S3 (4)). Together, these results suggest that lateral visual regions may not be organized in a strict feedforward hierarchy as has previously been suggested (6; 9). This is in contrast to the ventral

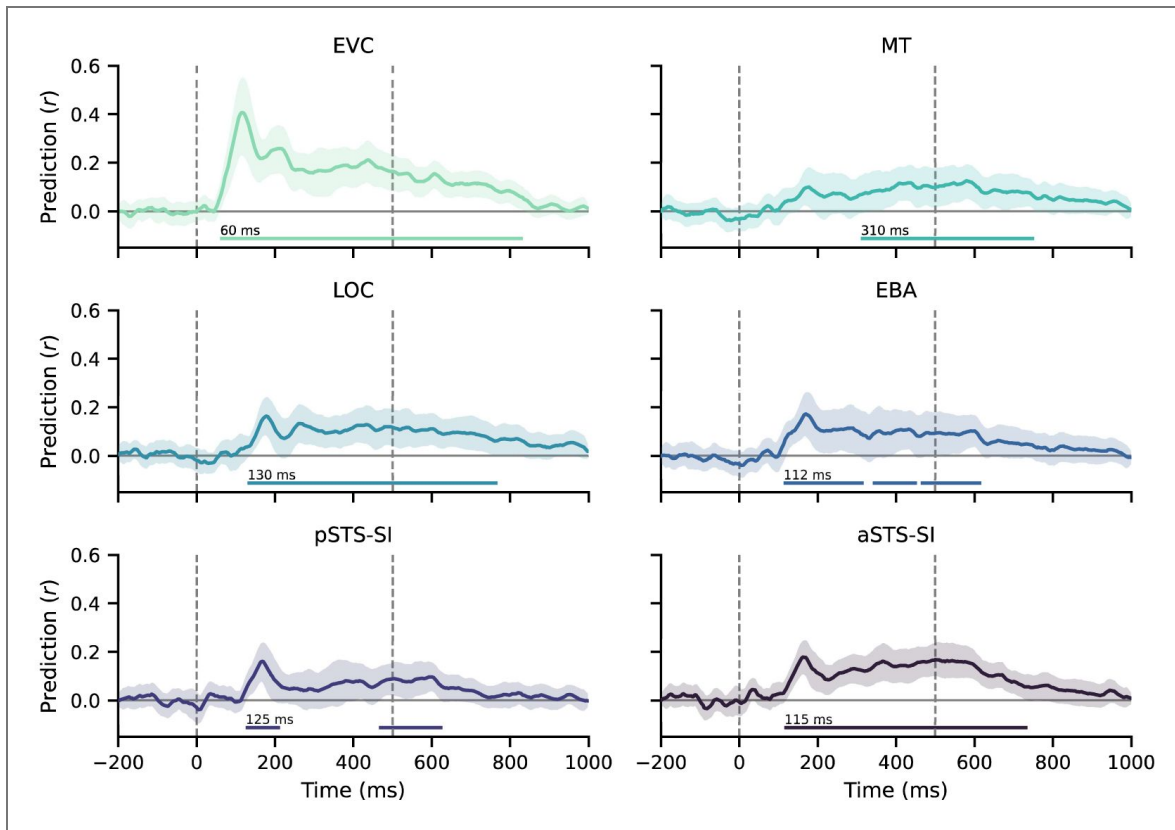


Figure 2. ROI decoding from EEG.

Decodability of responses lateral stream ROIs from the EEG signals across time. **Top:** early regions EVC and MT. **Middle:** mid-level regions LOC and EBA. **Bottom:** high-level regions pSTS-SI and aSTS-SI. Shaded regions are the 95% confidence intervals from bootstrapped variance distributions. Bold horizontal lines indicate significant decoding (permutation testing, cluster-corrected $p < 0.05$). The time leading significant lines indicates significance onset. The vertical dashed lines indicate the stimulus presentation period. For visualization purposes alone, the prediction time courses were further smoothed using a 25 ms sliding window. Significant time points are independent of this smoothing kernel.

stream, where, by 100 ms, there is also prediction in posterior ventral temporal regions, suggesting that the ventral stream may be organized in a more strict feedforward hierarchy than the lateral stream (Figure 3 and Figure S3).

3.2.1 A Spatiotemporal Hierarchy of Social Action Features Across the Lateral Stream

Our stimulus set includes human behavioral ratings that capture various visual and social scene features. These rated dimensions include mid-level features about the spatial layout of the scene and the configuration between people, including descriptions of the scene's size (*spatial expanse*), the proximity of individuals in the video (*agent distance*), the degree to which the individuals are facing each other (*facingness*), and the extent to which an action involves an object (*object directedness*). As well as high-level features about the social content, including the extent to which people were engaged in a joint or physical interaction (*joint action*), the degree of communication between people (*communication*), and affective characteristics (*valence* and *arousal*). The ratings were obtained using a Likert scale from at least ten subjects, and our prediction target is the average of the participants' ratings. In addition to human annotations, we also used AlexNet-conv2 and motion energy as models of low-level visual processing (25) and motion activations in MT (26), respectively.

We then asked how these social action features arise in the brain over space and time using a novel two-step regression method (Figure 1B). Briefly, the two-step procedure involves predicting the fMRI responses from the 64-channel EEG activity at each time point (as in Section 3.2). Then we used the stimulus features to predict the EEG-predicted fMRI activity ($\widehat{fMRI}_{train}$). This follows the same logic as the procedure for removing nuisance regressors, but takes an opposite approach to “regress in” the EEG-relevant fMRI signals. We scored the regression by predicting the EEG-predicted fMRI activity ($\widehat{fMRI}_{test}$) to the held-out test videos and then correlating the actual fMRI response ($fMRI_{test}$) in the test set to this prediction at each time point ($\widehat{fMRI}_{test}$). A more detailed description of the procedure is available in Section 5.4.5.

Using this method, we find that EVC represents low-level visual and motion features with a short latency (AlexNet-conv2: $t_{onset} = 68$ ms, motion energy: $t_{onset} = 60$ ms, Figure 4A and Figure S4). However, feature representations in MT are relatively late and have poor overall prediction (motion energy: $t_{onset} = 140$ ms and arousal: $t_{onset} = 305$ ms, Figure S5), which suggests low correspondence between EEG and fMRI signals in MT.

Mid-level lateral regions represent social primitive features later than low-level features in EVC. In particular, LOC activity was predictable by motion energy ($t_{onset} = 268$ ms), spatial expanse ($t_{onset} = 160$ ms), and agent distance ($t_{onset} = 160$ ms, Figure 4B and Figure S6), and EBA by motion energy ($t_{onset} = 302$ ms, Figure S7).

In contrast, high-level lateral regions in the STS represent both social primitive and social interaction features. pSTS-SI was predicted by spatial expanse ($t_{onset} = 148$ ms) and agent distance ($t_{onset} = 148$ ms, Figure S8) while aSTS-SI was predicted by spatial expanse ($t_{onset} = 215$ ms), agent distance ($t_{onset} = 150$ ms), and communication ($t_{onset} = 195$ ms, Figure 4C and Figure S9).

Social primitive features are represented in both LOTC (EBA and LOC) and STS (pSTS-SI and aSTS-SI) at a similar latency, mirroring the similar EEG-fMRI time course in (Section 3.2). Importantly, however, communication representations are present only in aSTS-SI, not lower-level regions, and have a later onset and peak time than social primitive features in any ROI. Together, these support the idea that social information is extracted hierarchically along the lateral stream (Figure S10).

4 Discussion

Building on existing theories (3; 9), we hypothesized that the perception of social interactions relies on hierarchical computations in the lateral visual pathway (12). Here, we examined the spatiotemporal flow of information through lateral visual regions as a unique window into the underlying neural computations. Using novel methods in EEG-fMRI fusion, we found a temporal

Figure 3. Whole-brain fMRI decoding from EEG.

The prediction of voxelwise activity in the whole brain from EEG signals across time. Shown are the lateral (top) and ventral (bottom) views of the right hemisphere in the native space of one representative fMRI subject. This is shown as a snapshot of encoding performance in 50 ms increments from 0 to 300 ms. Other fMRI participants are visible in [Figure S3](#).

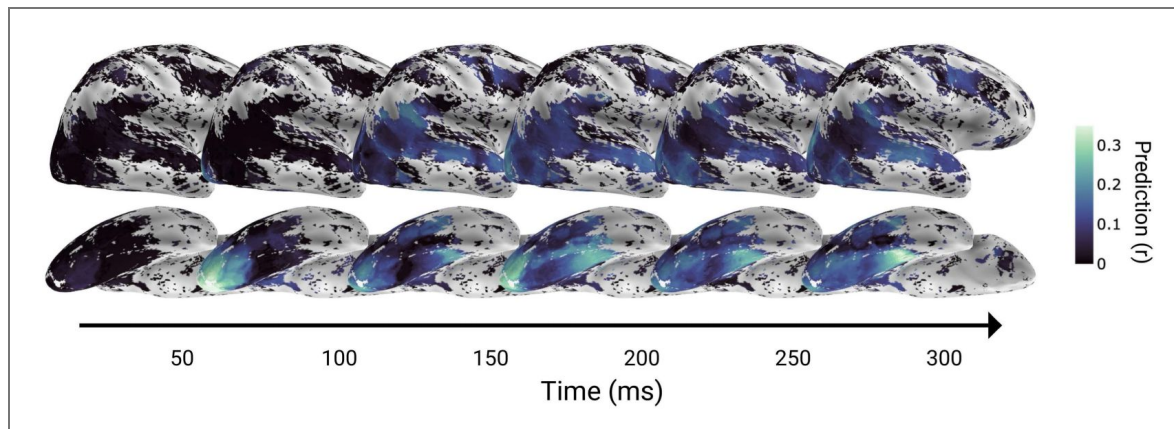
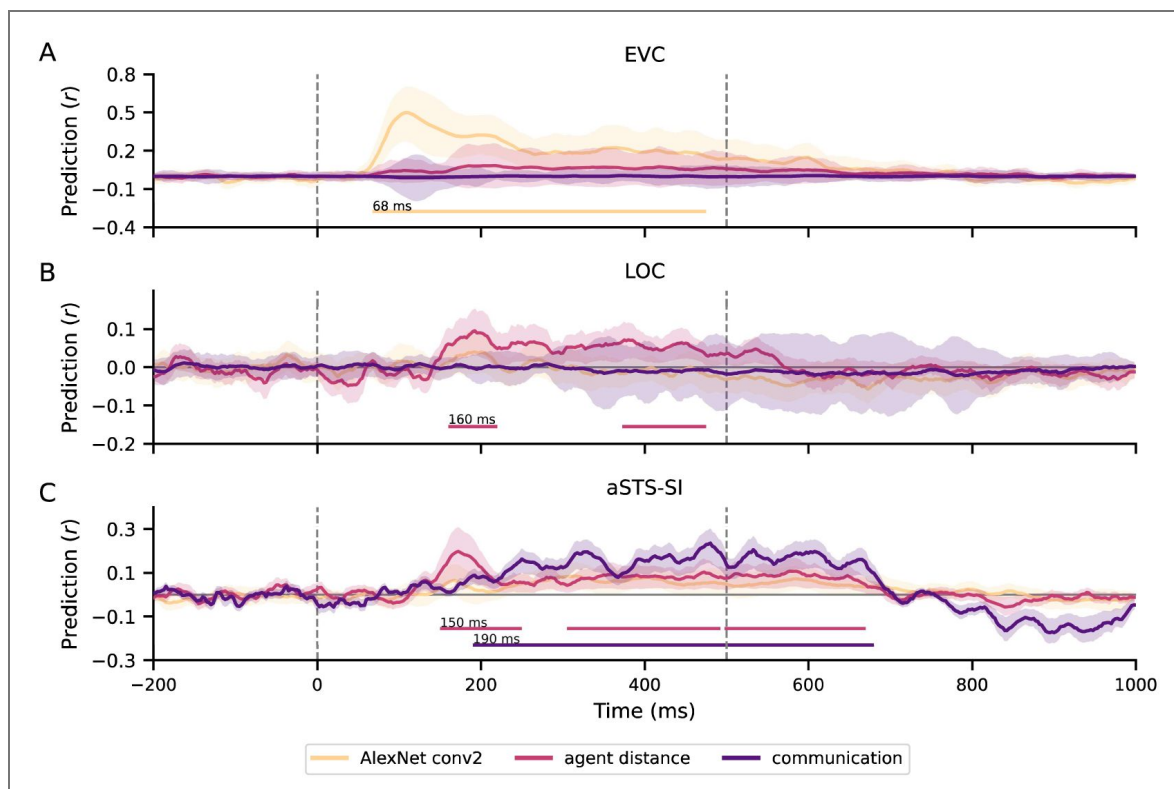


Figure 4. Joint EEG-feature fMRI encoding.

Joint EEG-feature encoding in three select ROIs in (A) EVC, (B) LOC, and (C) aSTS-SI for a low-level feature (Alexnet-conv2, yellow), mid-level feature (agent distance, red), and high-level feature (communication, purple). All other plotting conventions follow those of [Figure 2](#). See [Figures S4–S9](#) for the time course of feature encoding in each ROI.



hierarchy in feature decoding—more abstract features were decodable at longer latencies in fMRI fused responses in the lateral stream. These results provide novel support for a spatiotemporal hierarchy of social action features along lateral visual regions.

4.1 Early Representations of Communicative Actions in the STS

We found that communicative interactions are decodable from EEG signals and are represented in anterior STS, supporting our hypothesis that representing communication among others is a computational goal of the lateral visual stream (6). Relatedly, we have hypothesized that representing the social interactions between others is a visual process (12). Here we show for the first time that communicative interactions can be decoded from the human STS within 200 ms of video onset. This fast representation is consistent with estimates of feedforward visual processing (24; 27), providing novel empirical support for the theory that social interactions are extracted by the human visual system. While prior work failed to find such rapid decoding in natural stimuli (28; 1), our ability to detect early social interaction signals seems to be due to the power of our novel methods in EEG-fMRI fusion and its advantages over whole brain M/EEG decoding (see below).

4.2 Broad Representations of Social Primitive Features Across the Lateral Stream

Decoding of communication follows shortly after decoding mid-level social features, such as the distance between agents, which have been referred to as “social primitives” or precursors of social interaction in prior work (12; 6). This lends further support to the theory that social interactions are extracted via these mid-level primitives.

Our results, however, find that these representations of social primitives are spatially distributed across both mid- and high-level lateral stream regions. In a recent transcranial magnetic stimulation (TMS) study, only EBA was causally implicated in the perception of social primitives, in direct contrast to nearby regions in LOTC (29). One key difference between Gandolfo et al. (29) and the current study is the use of dynamic versus static stimuli. It is possible that social primitives of people in motion are represented broadly, while social primitives in static images are only represented in the EBA.

Another possible reason that we find broad representation of social primitives in the current study is that social primitives are correlated with geometric scene structure in the stimulus set (6). As a result, it may be that scene content, rather than social primitives, is broadly represented throughout the visual cortex. Previous work did find representations of scene structure and social primitives in STS regions before controlling for shared variance among features (6). Future analyses controlling for shared variance among features may elucidate whether the lateral visual stream broadly represents social primitive features.

4.3 Two-Step Regression to Characterize Feature Representations in Space and Time

In the current study, we proposed a novel method for investigating the spatiotemporal dynamics of representations in the human brain non-invasively. We do this by performing a two-step regression in which we first predict fMRI responses from whole-brain EEG activity and then use stimulus features to predict the EEG-predicted fMRI representations. Using this method, we found that only low-level visual features (features from the second convolutional layer of AlexNet or a motion energy model) are predictive of EVC responses through time (Figure S4 4). This validates the specificity and utility of the method for investigating spatiotemporal representations in the human brain.

Somewhat counterintuitively, we find that some information is decoding better with this method than direct decoding from the EEG activity. For example, MT was poorly predicted by EEG directly and the onset latency was much later than would be expected based on electrophysiology (Figures ??) (24), but following the two-step regression, we find early motion energy decoding in MT

consistent with feedforward processing. These results suggest that the two-step regression procedure—in addition to linking neural signals in space and time—may also effectively denoise the EEG signal by considering only the information relevant for a given region of interest.

This two-step regression procedure can be applied to any other dataset or question if there are fMRI and EEG responses for the same stimuli. This neural data can be combined with stimulus features as is done here, but it could also be extended to other applications such as investigating the time course of any behavior in a region of interest or investigating how neural networks align with brain regions through time. Further, unlike invasive electrophysiology, this method enables the characterization of the temporal dynamics with dense, whole-brain spatial resolution.

While this is a powerful method to characterize the spatiotemporal dynamics of representations in the human brain non-invasively, the method is still limited by the noise of the recording methodologies, particularly EEG. We partially mitigate this limitation by having multiple stimulus repetitions, but it may be less suited to research questions in which averaging across trials is not possible. The method is also limited by the poor coverage of EEG signals in deeper cortical regions and sub-cortical structures. Future research comparing the spatiotemporal profile found via EEG-fMRI fusion with invasive recordings can shed light on the specificity of this method under different conditions.

4.4 Beyond Feedforward Hierarchies

Our findings suggest that the processing of social action features from posterior to anterior brain regions does not adhere to a strict feedforward hierarchy. Information in mid- and high-level regions comes online at similar time points and both regions have representations of mid-level, social primitive features. The broad representation of mid-level features could be explained by skip connections from EVC to the most anterior regions directly. Alternatively, information in separate but nearby cortical regions may be blurred due to the limitations of EEG-fMRI fusion described above.

Broadly, our results are in line with prior research indicating that even the ventral visual stream may not be accurately described by a simple feedforward hierarchy. The hypothesized hierarchy in the ventral stream in macaques (containing V1, V2, V4, and posterior inferotemporal cortex (IT), central IT, and anterior IT) (24) is largely based on observations of longer latencies of responses in more anterior regions and on retinotopic re-representation of the visual field with increasingly large receptive field sizes (24). However, a large body of work has highlighted that regions along the ventral visual stream contain many feedback connections and interconnections with dorsal and lateral regions (30; 31). In light of these complexities, more research is needed to characterize the anatomical and functional profile of lateral visual regions.

4.5 Conclusion

Here, we investigated the spatiotemporal organization of features in regions of the lateral visual stream. Replicating prior work, we found that communicative actions are represented in the STS (6), and we shed novel light on the computations underlying social perception. We also find, for the first time, that communicative interactions are rapidly extracted by the human brain in natural scenes. Finally, we find strong evidence that social interactions are computed via hierarchical computation of low- and mid-level social visual features in the lateral visual pathway. Future research should aim to better characterize the anatomical and functional profile of lateral regions to further refine our models of the neural basis of human social perception.

5 Materials & Methods

5.1 Reproducibility

For reproducibility, all code is available publicly on GitHub: github.com/Isik-lab/SIEEG_analysis.git .

5.2 Stimuli and fMRI

5.2.1 Stimulus Set and Feature Annotations

Videos, feature annotations, and fMRI are from our prior dataset (6). Here we shortened the videos from 3 s to the central 500 ms to facilitate time locking of the EEG signal as in other EEG-fMRI fusion work (19). The stimulus set includes behavioral annotations of features of the visual and social scene (see Section 3.2.1 for a full description of the stimulus features).

Similar to our previous study, in addition to the behavioral ratings, we also extracted the activations from the second convolutional layer of a pytorch implementation (32) of an ImageNet (33)-trained AlexNet (34) as a model of early visual processing (25) (referred to throughout as AlexNet-conv2). The activations were extracted for every frame and then averaged across frames.

Motion energy was estimated with an Adelson and Bergen model (26) implemented in pymoten (35) using the default pyramid with a temporal window of 10 frames. The motion energy was then averaged across spatiotemporal windows.

Both AlexNet-conv2 and motion energy were estimated on the full 3 s video because they were only used to predict the EEG-predicted fMRI activity and participants in the fMRI saw the 3 s videos. We reduced the dimensionality of AlexNet-conv2 and the motion energy using principal components (PC) analysis implemented in scikit-learn (36) to the number of samples of the training set. We learned the PCs in the training set and applied the learned components to the test set.

5.2.2 fMRI

In our prior experiment (6), participants ($n = 4$) viewed the 3 s videos in the fMRI scanner. The videos were divided into 200 videos in a training set that were repeated 9-10 times and a 50 video test set that was repeated 18-20 times. The reliability of the fMRI data was estimated as the splithalf correlation across repetitions for every voxel in the brain. In our prior study as well as in the current investigation, analyses are always limited to voxels that were determined to have a significant correlation ($p < 0.05$, uncorrected).

In addition to the main videos, the participants also completed a battery of functional localizer tasks to enable functional localization of ROIs in the ventral and lateral stream (see Section 3.2 for a description of the ROIs investigated in the current study).

5.3 EEG Experiment

5.3.1 Participants

Participants ($n = 21$, 5 Males, $M = 21.4$ years, $SD = 2.9$ years) gave informed consent prior to participation in accordance with the Johns Hopkins University Institutional Review Board and were either given course credit or monetary compensation for their time. One participant was excluded from subsequent analyses due to excessive movement (final sample, $n = 20$).

5.3.2 Experimental Procedure

Based on previous studies (19; 1), continuous EEG recordings with a sampling rate of 1000 Hz were made with a 64-channel Brain Products ActiCHamp system using actiCAP slim electrode caps in a Faraday chamber. Electrode impedances were kept below 25 k Ω when possible, and the Cz electrode was used as an online reference.

Participants were seated upright while viewing the videos on a back-projector screen situated approximately 60 cm away.

During the experiment, participants viewed 250 500 ms clips of social actions. The training-test split (200 and 50 videos in train and test sets, respectively) from McMahon et al. (6) was maintained. Videos in the training set were repeated four times in total, and videos in the test set were repeated sixteen times.

The experiment consisted of four sections. For a given section, the 200 training videos were randomly divided into four blocks, and the test videos were shuffled and presented in one block. One section consisted of the four training blocks and the test block repeated four times.

During each block, five “catch” videos were randomly sampled from 50 videos that depicted crowds of people and were randomly interspersed among the other videos. The participants’ task was to hit a button on the crowd trials. These trials were not analyzed in any subsequent analyses.

Between blocks, participants were told that they could take a short break to rest their eyes, but that they should remain still and hit a button when they were ready to continue. Between each of the four sections, participants were given longer breaks. They were told to rest as long as needed and announce when they were ready to continue.

Stimuli were presented with an Epson Home Cinema 3800 projector with a 60 Hz refresh rate. Videos were shown on a black background and subtended approximately 16×16 degrees of visual angle. Between trials, a white fixation cross was displayed that disappeared when the videos began. Participants were instructed to fixate between trials but were told that they could move their eyes during the videos. A photodiode was used to accurately track on-screen stimulus presentation times and account for projector lag. The paradigm was implemented in MATLAB R2021b using the Psychophysics Toolbox (37).

5.4 Analysis

5.4.1 EEG preprocessing

Minimal EEG data preprocessing was performed using MATLAB R2023b and FieldTrip (38). The EEG data were aligned to stimulus onset and cut to 1.2 s (0.2 s pre-stimulus to 1 s post-stimulus onset), baseline-corrected using the 0.2 s prior to stimulus onset, high-pass filtered at 0.1 Hz, and low-pass filtered to 60 Hz. Data were resampled to 400 Hz and temporally smoothed over five consecutive samples (12.5 ms windows). Finally, catch trials and false alarm trials were removed.

5.4.2 EEG reliability

We estimated the signal quality of EEG data by calculating the split-half reliability in the test set. For each EEG participant, channel, and time sample, even and odd presentations of videos were split in half and then averaged. The correlation was then computed between the two halves of the data across videos. The variance of the reliability was estimated using bootstrapped resampling of the videos 5000 times and recomputing the even-odd correlation. Reliability and variance distributions were averaged across channels and participants.

5.4.3 ROI response decoding

Within each EEG participant at each time point, we used the 64-channel EEG activity of each participant at each time point to predict the average response in each ROI for each of the fMRI participants using ridge regression. Given the high correlation between the signal recorded by separate EEG electrodes, we first used principal components analysis (PCA) to rotate and orthogonalize the EEG channel features without reducing the dimensionality. The ridge penalty (varied between 10^{-5} and 10^{30}) was fit using optimized leave-one-out prediction implemented in the DeepJuice python package (39) in the training set, and evaluation was performed in the test set. Performance was evaluated as the correlation between the predicted and actual ratings for each feature.

When plotting the time course of decoding, we further smoothed the data using a 25 ms sliding average window. This was done for visualization purposes alone and is not reflected in any of the statistical results. This smoothing procedure was done for all other time course plotting.

5.4.4 Voxelwise fMRI encoding with EEG

To perform whole-brain analyses, we predicted the voxelwise fMRI response following smoothing to a 12 mm full-width-half-max Gaussian kernel to increase signal-to-noise ratio. We performed the same regression procedure as in (Section 5.4.3 40) to predict the voxelwise activity for each

fMRI participant. Visualizations are done by averaging across EEG participants for individual fMRI participants.

5.4.5 Two-Step EEG-Feature fMRI Encoding

Inspired by the regression procedures used by others (40; 41), we devised a novel regression framework to map between EEG, fMRI, and feature annotations (Figure 1B). The procedure involved two regression steps. In step 1, we performed five-fold cross-validation from EEG to fMRI in the training set. We predicted the response for the held-out stimuli and did this across all folds. This results in a predicted fMRI training set based on the EEG data. In other words, we generated a matrix of the fMRI signal that was predictable by EEG. In step 2, we fit a regression with a train/test split with the video annotations (either human ratings or image-computed low-level visual features) as the predictor and the EEG-predicted fMRI response from step 1 as the target. In this way, the method maps features to the EEG-predicted portion of the fMRI signal. We scored the regression by predicting the response in the fMRI test set using the held-out test ratings or visual feature representations and correlating the predicted with the actual response. This procedure allows us to predict the fMRI response in a shared EEG-rating space.

5.4.6 Statistical Analysis

To estimate the distribution of variance of our model performance, at every time point, we performed bootstrap resampling over the model predictions and the true response over 5000 iterations. Similarly, to estimate a statistical null distribution, we performed permutation shuffling between the predicted and true response over 5000 iterations at each time point. We averaged both the bootstrap and permutation distributions first over EEG participants and then over fMRI participants for group-level analyses. Group-level variance was estimated as the 95% confidence interval of the bootstrapped distributions, and group-level significance was the one-tailed test of true relative to the average null distribution. Significance was cluster-corrected for multiple comparisons using the maximum cluster sum across time windows and regressions performed, $\alpha = 0.05$, cluster-setting $\alpha = 0.05$.

Supplemental Figures

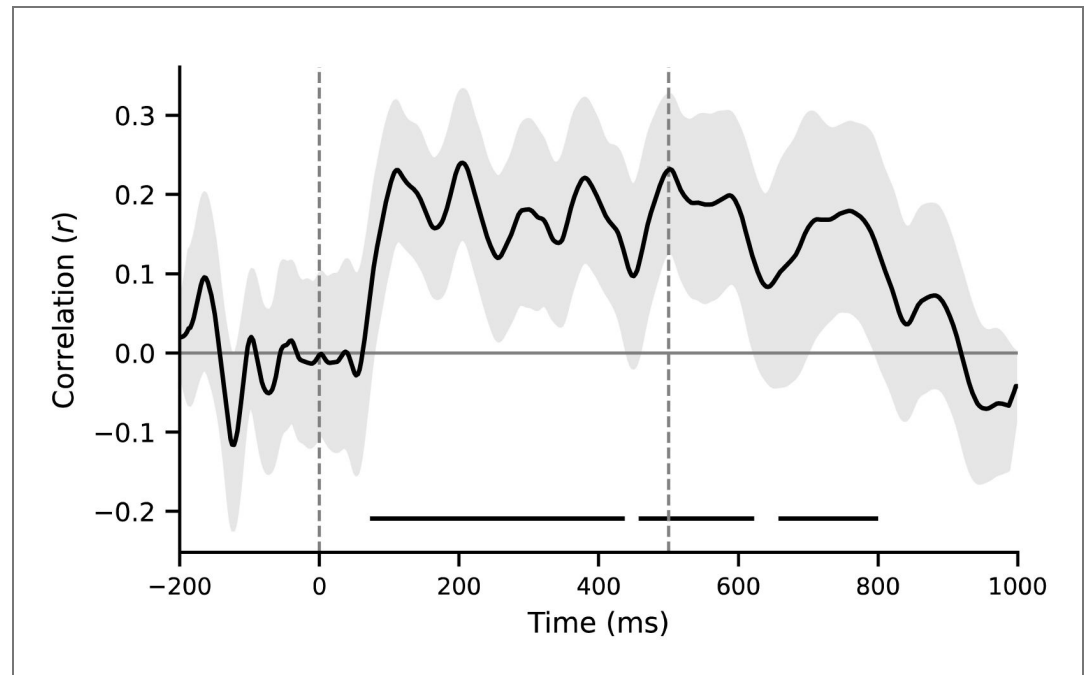


Figure S1. EEG data reliability. Data reliability in the test set averaged across participants and channels. Shaded areas are the bootstrapped 95% confidence intervals and solid horizontal lines represent time windows of significant reliability after cluster correction at a level of $p < 0.05$. Horizontal lines mark time clusters of significant reliability (permutation testing, cluster-corrected $p < 0.05$). As in [Figure 2A](#) data are smoothed in a 25 ms sliding window for visualization alone.

Figure S2. Latency of all ROI prediction from EEG decoding.

Decodability of ROI activity recorded in time bins extended from [Figure 2B](#).

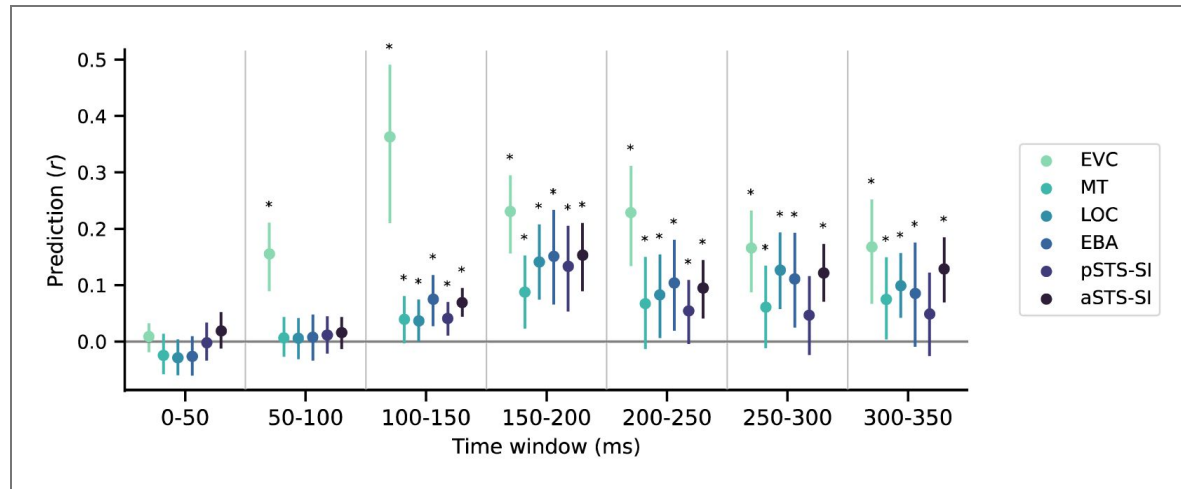
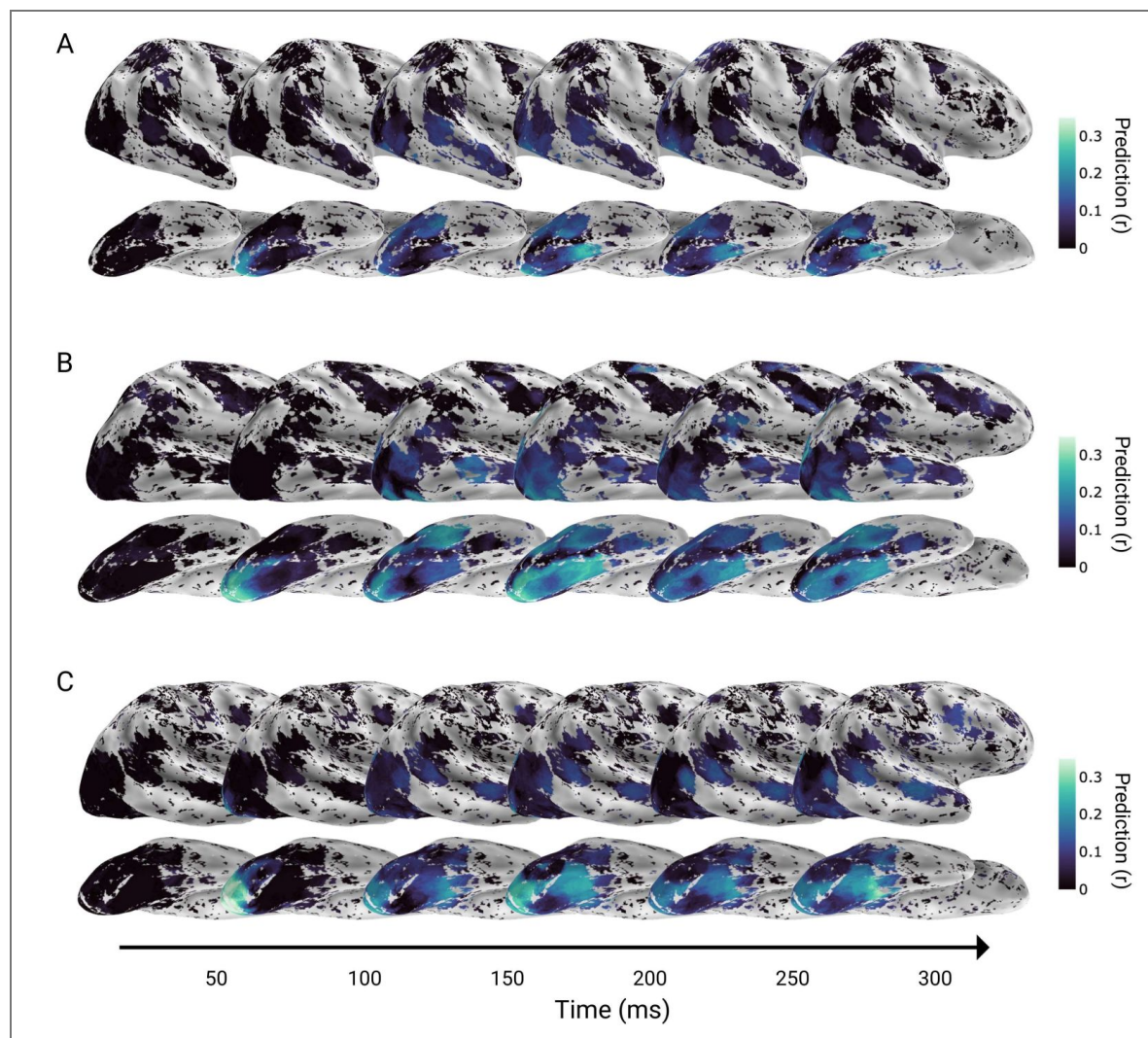


Figure S3. Whole brain prediction in other fMRI participants.

Plotting conventions are all the same as [Figure 2C](#), but for (A) sub-01, (B) sub-03, and (C) sub-04 from McMahon et al. (6).



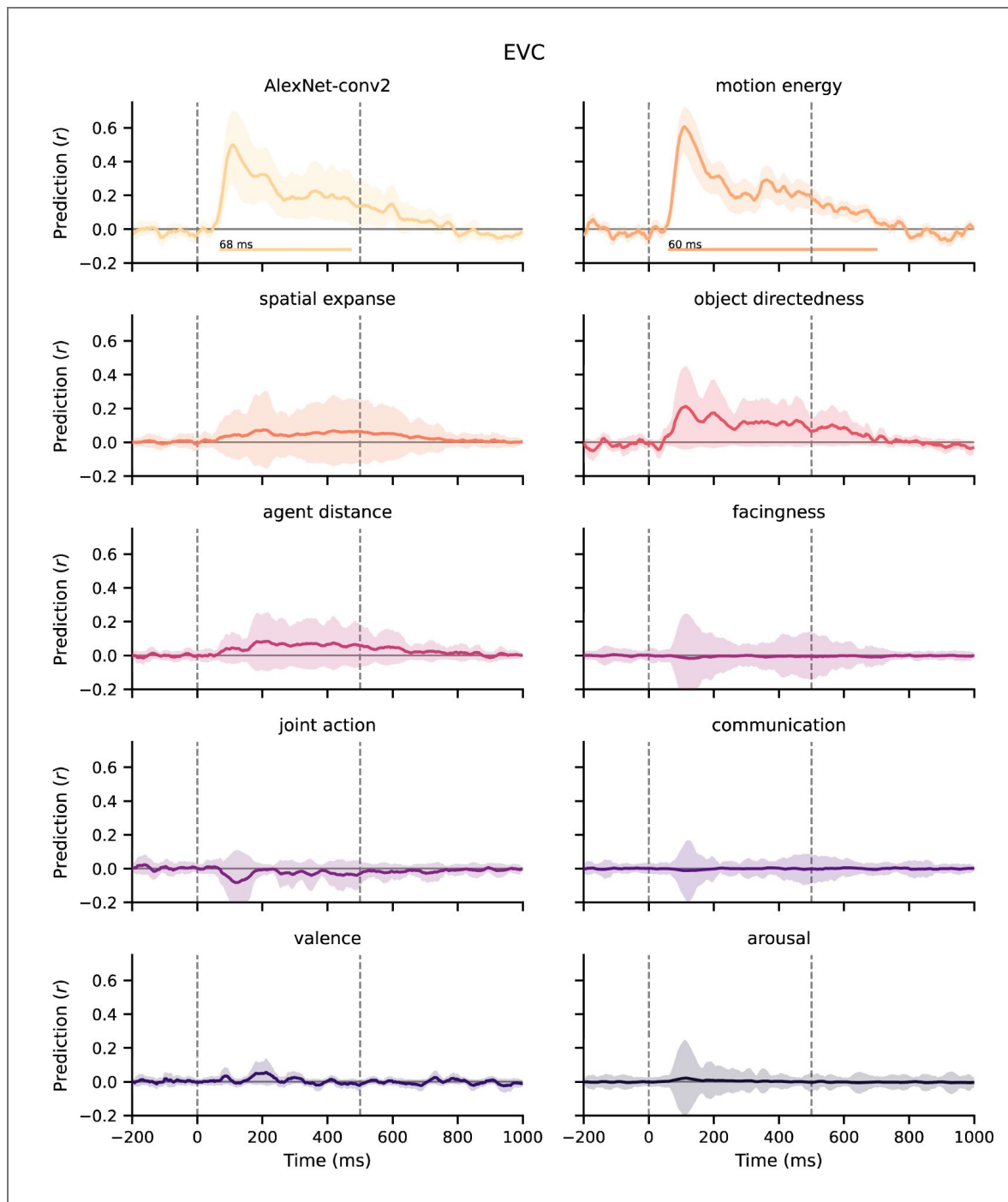


Figure S4. Time course of joint EEG-feature fMRI encoding of each feature in EVC

Extended to all features from Figure 4A. For visibility, all features are shown on separate subplots.

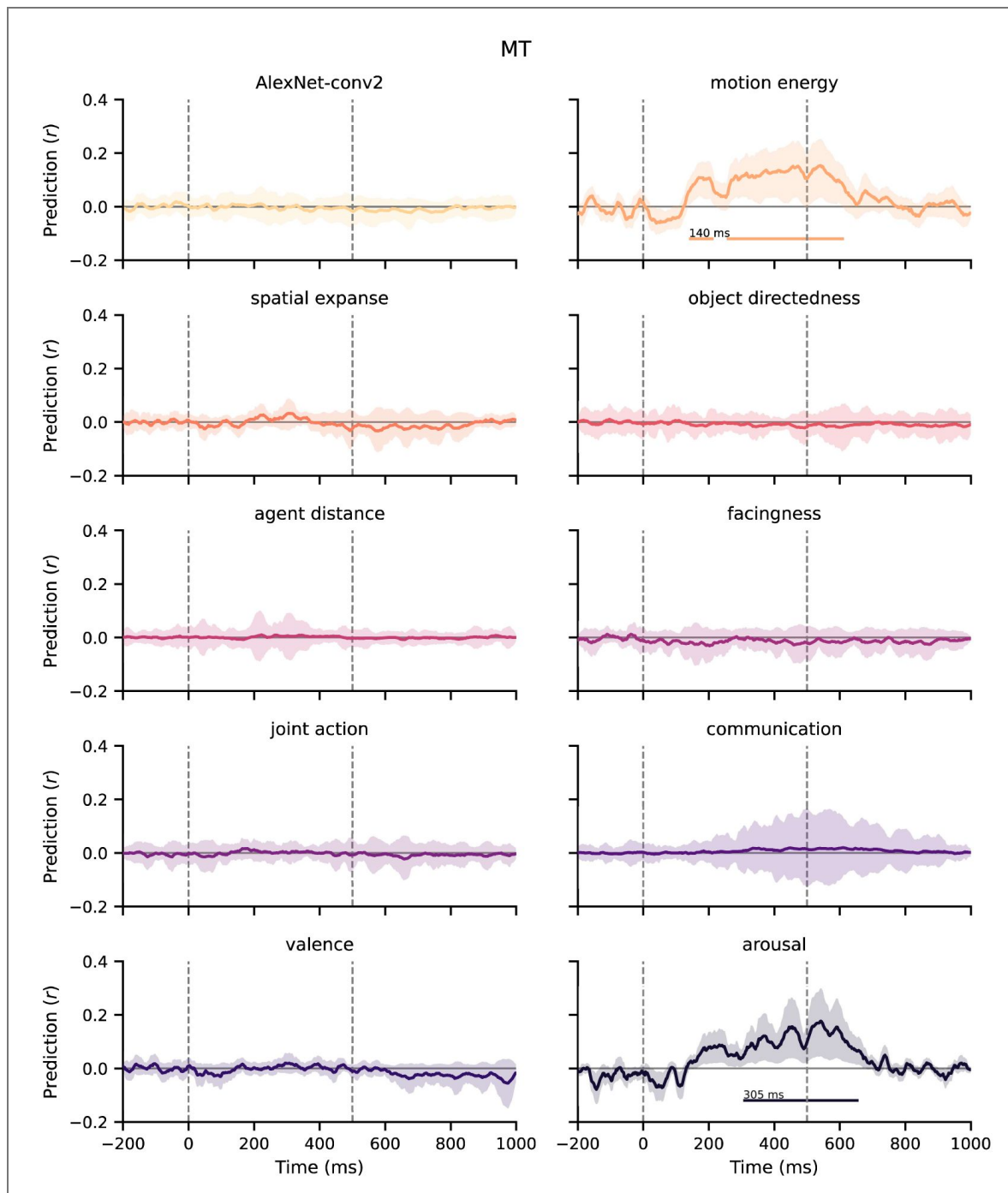


Figure S5. Time course of joint EEG-feature fMRI encoding of each feature in MT

Related to main text Figure 4 [↗](#).

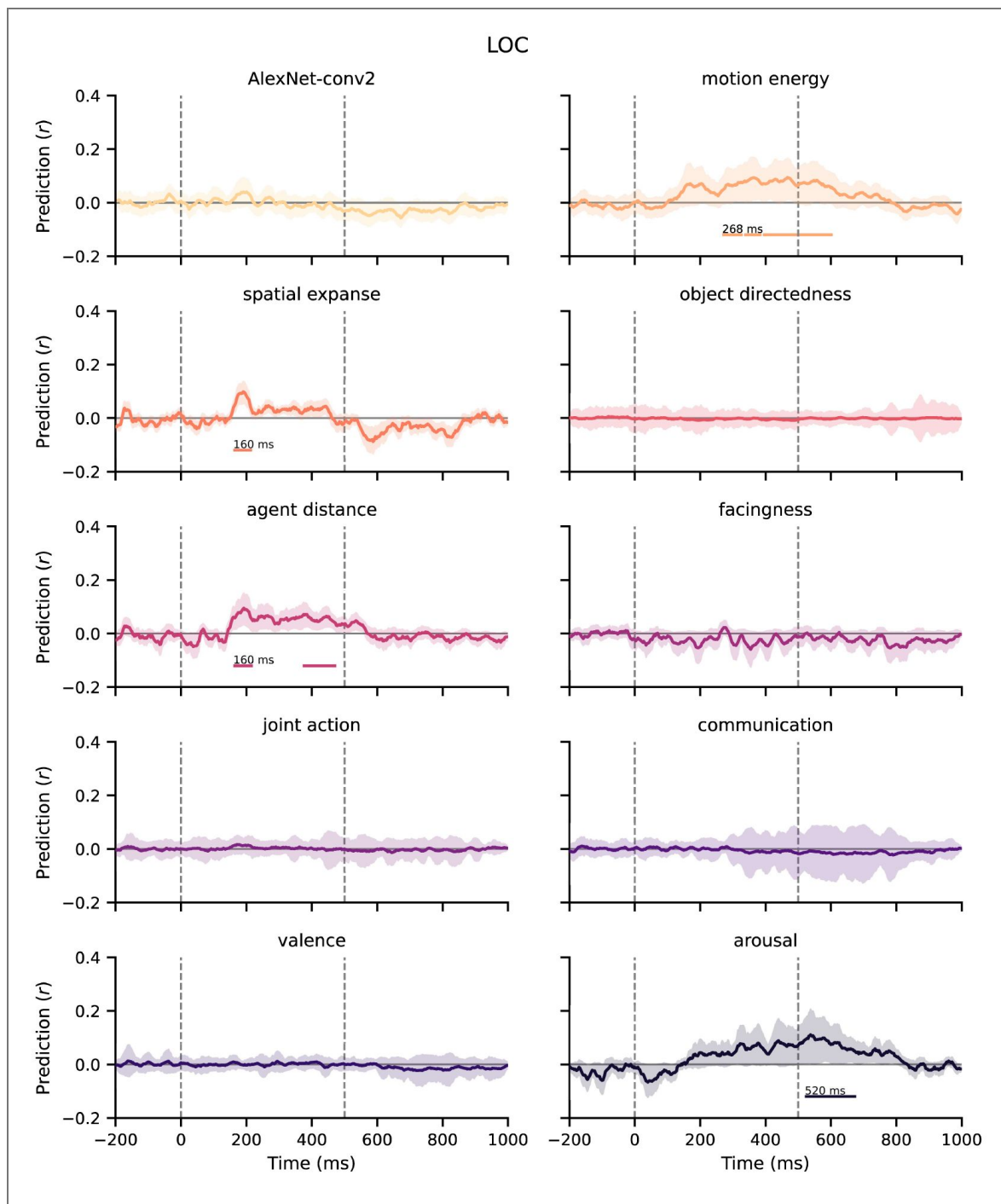


Figure S6. Time course of joint EEG-feature fMRI encoding of each feature in LOC

Extended from main text Figure 4B [for all features](#).

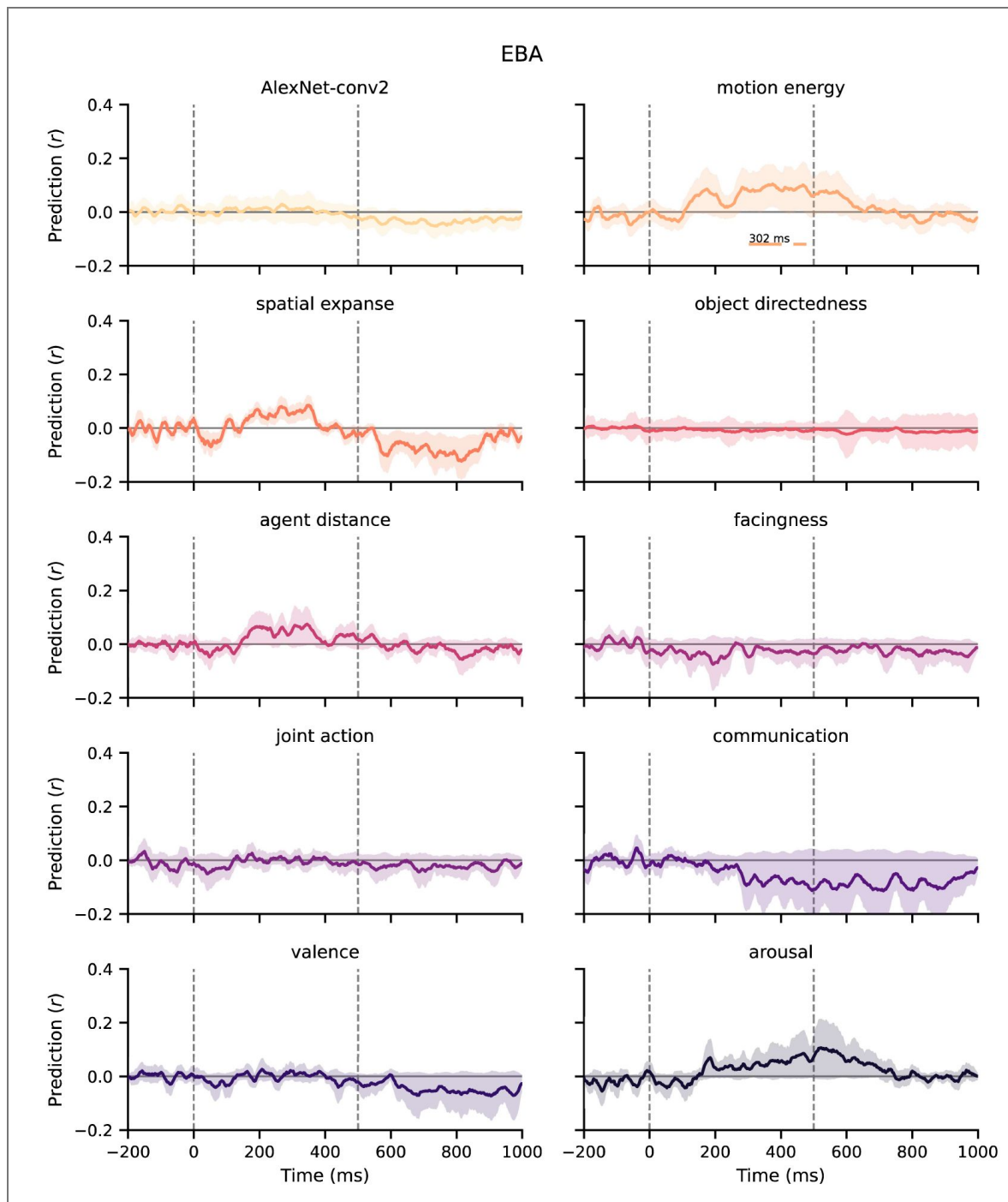


Figure S7. Time course of joint EEG-feature fMRI encoding of each feature in EBA

Related to main text Figure 4 [↗](#).

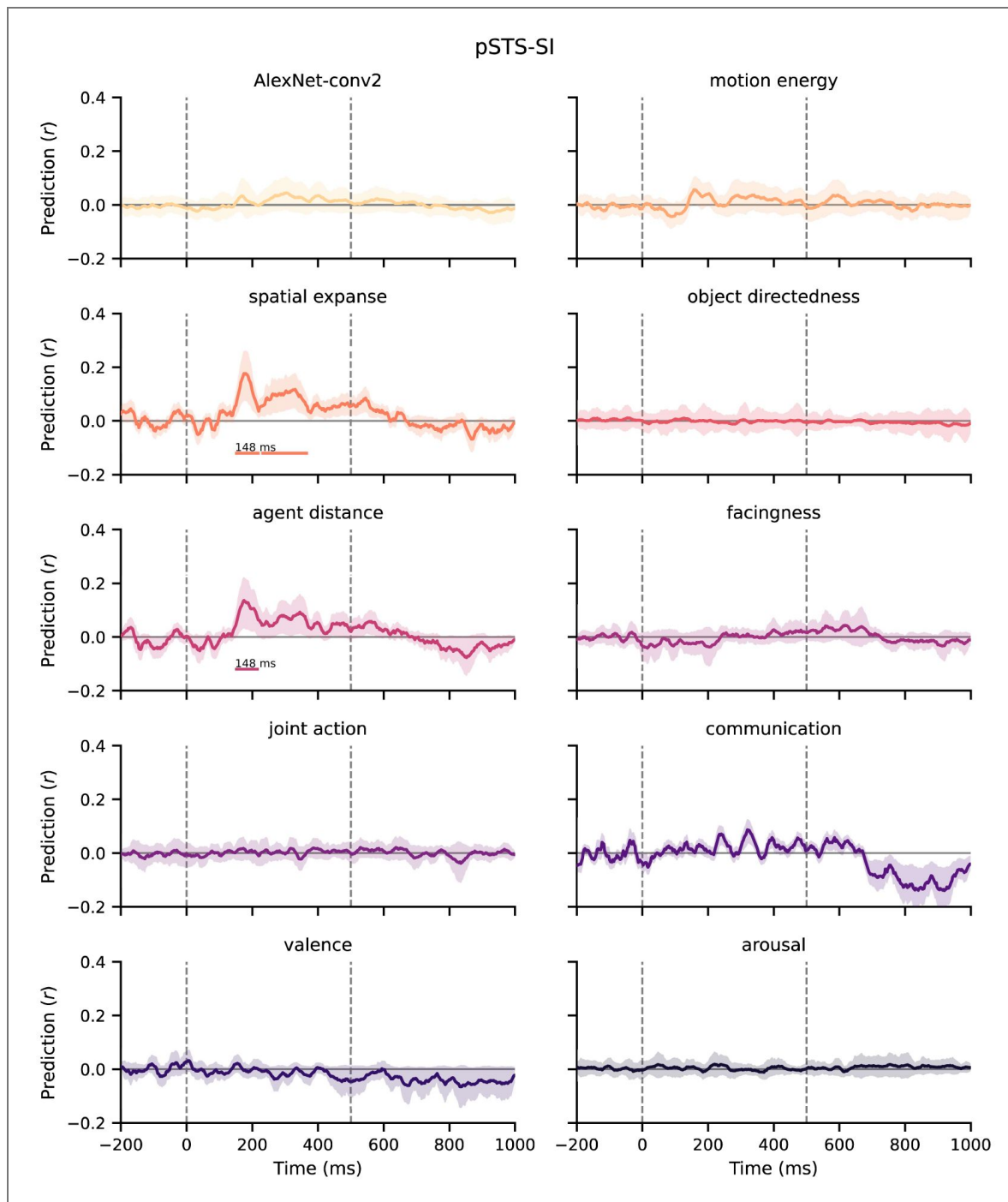


Figure S8. Time course of joint EEG-feature fMRI encoding of each feature in pSTS-SI

Related to main text Figure 4 [↗](#).

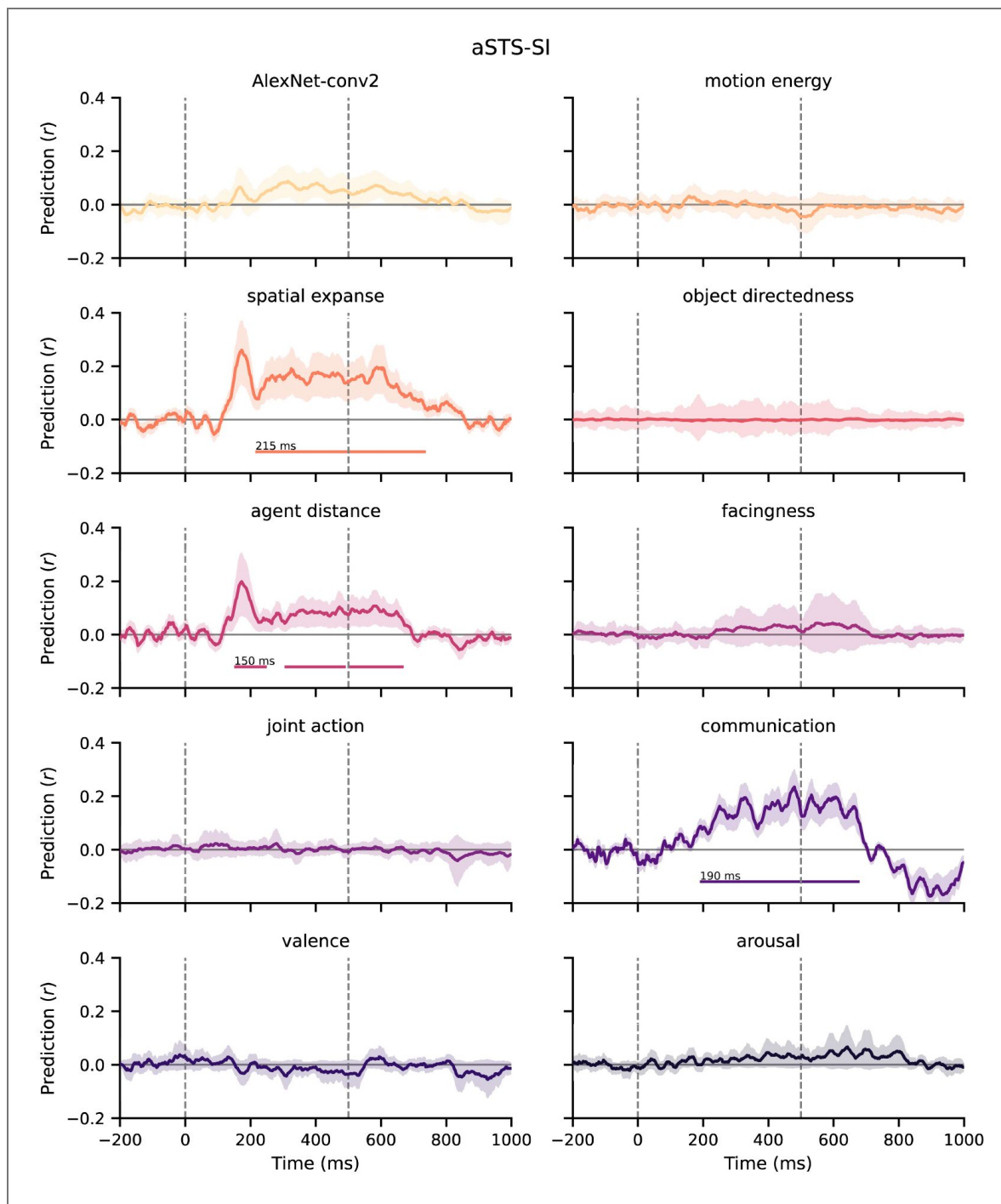


Figure S9. Time course of joint EEG-feature fMRI encoding of each feature in aSTS-SI

Extended from main text Figure 4C [for all features](#).

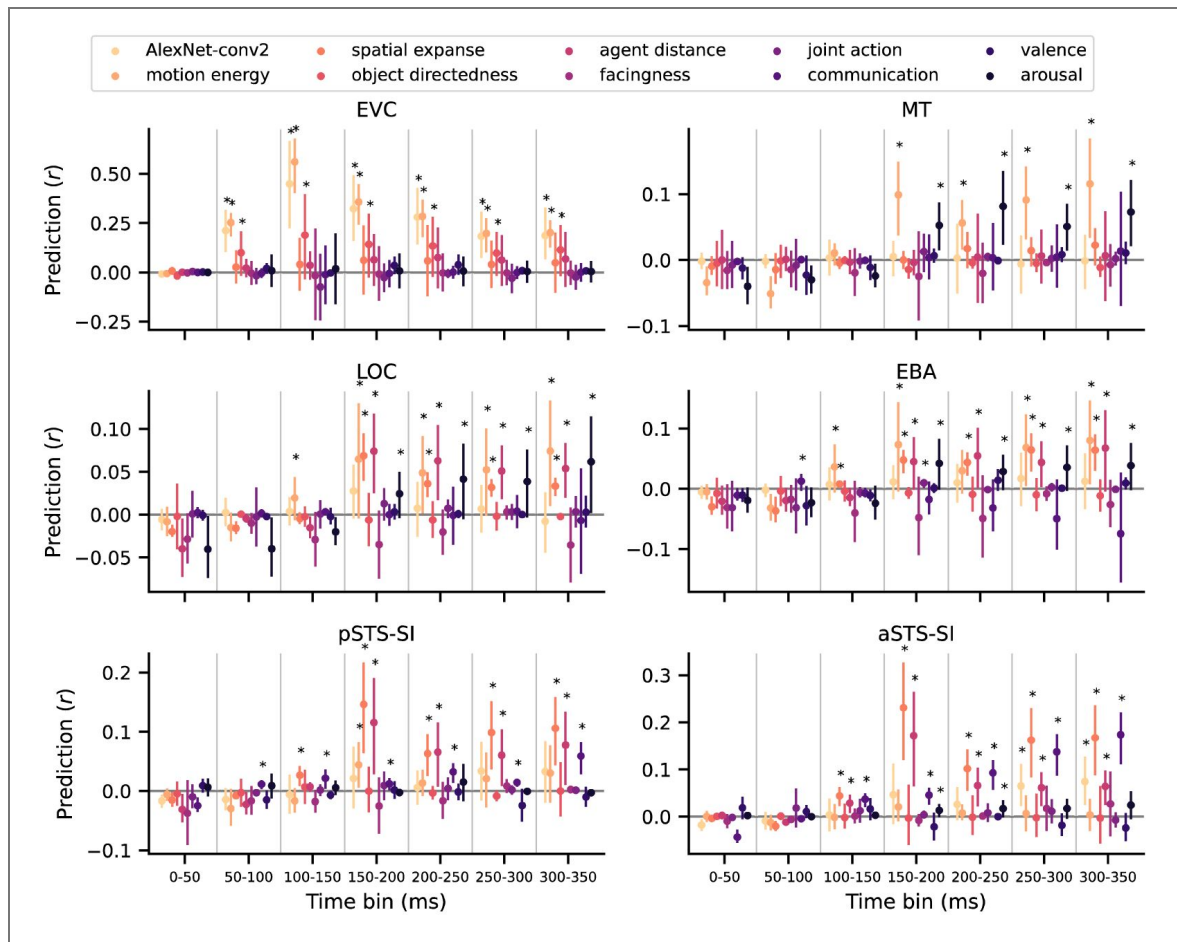


Figure S10. Latency of feature prediction in ROIs.

Extended from main text Figure 4D-F [for all features and all ROIs.](#)

Data availability

All EEG, fMRI, and behavior data have been made publicly available and are available here: <https://osf.io/2vnw4/overview>. The code for the analyses presented in this paper is openly accessible at: https://github.com/Isik-lab/SIEEG_analysis/.

Acknowledgements

This work was funded in part by NSF GRFP DGE-1746891 awarded to E.M. and NIMH R01MH132826 awarded to L.I.

Additional information

Funding

Funder	Grant reference number	Author
HHS NIH National Institute of Mental Health (NIMH)	R01MH132826	Leyla Isik
National Science Foundation (NSF)	RFP DGE-1746891	Emalie McMahon

Author ORCID iDs

Leyla Isik:  <https://orcid.org/0000-0002-9255-0151>

References

- [1] **Dima Diana C**, Tomita Tyler M, Honey Christopher J, Isik Leyla (2022) Social-affective features drive human representations of observed actions. *eLife* **11**:e75027 <https://doi.org/10.7554/eLife.75027> | <https://doi.org/10.7554/eLife.75027> | PubMed
- [2] **Tarhan Leyla**, Konkle Talia (2020) Sociality and interaction envelope organize visual action representations. *Nature Communications* **11** <https://doi.org/10.1038/s41467-020-16846-w> | PubMed
- [3] **Wurm Moritz F**, Caramazza Alfonso, Lingnau Angelika (2017) Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. *Journal of Neuroscience* **37**:562-575 <https://doi.org/10.1523/JNEUROSCI.1717-16.2016> | PubMed
- [4] **Varrier Rekha S**, Finn Emily S (2022) Seeing Social: A Neural Signature for Conscious Perception of Social Interactions. *Journal of Neuroscience* **42**:9211-9226 <https://doi.org/10.1523/JNEUROSCI.0859-22.2022> | PubMed
- [5] **Isik Leyla**, Koldewyn Kami, Beeler David, Kanwisher Nancy (2017) Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America* **114**:E9145-E9152 <https://doi.org/10.1073/pnas.1714471114> | PubMed
- [6] **McMahon Emalie**, Bonner Michael F, Isik Leyla (2023) Hierarchical organization of social action features along the lateral visual pathway. *Current Biology* **33**:5035-5047.e8 <https://doi.org/10.1016/j.cub.2023.10.015> | PubMed
- [7] **Masson Haemy Lee**, Isik Leyla (2021) Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* **245**:118741 <https://doi.org/10.1016/j.neuroimage.2021.118741> | PubMed
- [8] **Walbrin Jon**, Downing Paul, Koldewyn Kami (2018) Neural responses to visually observed social interactions. *Neuropsychologia* **112**:31-39 <https://doi.org/10.1016/j.neuropsychologia.2018.02.023> | PubMed
- [9] **Pitcher David**, Ungerleider Leslie G (2021) Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences* **25**:100-110 <https://doi.org/10.1016/j.tics.2020.11.006> | PubMed

- [10] Ungerleider Leslie G, Mishkin Mortimer (1982) Two cortical visual systems. In: Ingle David J, Goodale Melvyn A, Mansfield Richard JW (Eds). *Analysis of Visual Behavior* The MIT Press. pp. 549-586
- [11] Goodale Melvyn A, Milner A David (1992) Separate visual pathways for perception and action. *Trends in Neurosciences* **15**:20-25 [https://doi.org/10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8) | PubMed
- [12] McMahon Emalie, Isik Leyla (2023) Seeing social interactions. *Trends in Cognitive Sciences* **27**:1165-1179 <https://doi.org/10.1016/j.tics.2023.09.001> | PubMed
- [13] Abassi Etienne, Papeo Liuba (2020) The Representation of Two-Body Shapes in the Human Visual Cortex. *Journal of Neuroscience* **40**:852-863 <https://doi.org/10.1523/JNEUROSCI.1378-19.2019> | PubMed
- [14] Abassi Etienne, Papeo Liuba (2022) Behavioral and neural markers of visual configural processing in social scene perception. *NeuroImage* **260**:119506 <https://doi.org/10.1016/j.neuroimage.2022.119506> | PubMed
- [15] Landsiedel Julia, Daughters Katie, Downing Paul E, Koldewyn Kami (2022) The role of motion in the neural representation of social interactions in the posterior temporal cortex. *NeuroImage* **262**:119533 <https://doi.org/10.1016/j.neuroimage.2022.119533> | PubMed
- [16] de Vries Ingmar EJ, Wurm Moritz F (2023) Predictive neural representations of naturalistic dynamic input. *Nature Communications* **14**:3858 <https://doi.org/10.1038/s41467-023-39355-y> | PubMed
- [17] Cichy Radoslaw M, Oliva Aude (2020) A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron* **107**:772-781 <https://doi.org/10.1016/j.neuron.2020.07.001> | PubMed
- [18] Cichy Radoslaw Martin, Pantazis Dimitrios, Oliva Aude (2014) Resolving human object recognition in space and time. *Nature Neuroscience* **17**:455-462 <https://doi.org/10.1038/nn.3635> | PubMed
- [19] Masson Haemy Lee, Isik Leyla (2023) (2023) Rapid Processing of Observed Touch through Social Perceptual Brain Regions: An EEG-fMRI Fusion Study. *Journal of Neuroscience* **43**:7700-7711 <https://doi.org/10.1523/jneurosci.0995-23.2023> | PubMed
- [20] Diedrichsen Jörn, Kriegeskorte Nikolaus (2017) Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology* **13**:e1005508 <https://doi.org/10.1371/journal.pcbi.1005508> | PubMed
- [21] Naselaris Thomas, Kay Kendrick N, Nishimoto Shinji, Gallant Jack L (2011) Encoding and decoding in fMRI. *NeuroImage* **56**:400-410 <https://doi.org/10.1016/j.neuroimage.2010.07.073> | PubMed
- [22] Downing Paul E, Jiang Yuhong, Shuman Miles, Kanwisher Nancy (2001) A Cortical Area Selective for Visual Processing of the Human Body. *Science* **293**:2470-2473 <https://doi.org/10.1126/science.1063414> | PubMed
- [23] Grill-Spector Kalanit, Kourtzi Zoe, Kanwisher Nancy (2001) The lateral occipital complex and its role in object recognition. *Vision Research* **41**:1409-422 [https://doi.org/10.1016/S0042-6989\(01\)00073-6](https://doi.org/10.1016/S0042-6989(01)00073-6) | PubMed
- [24] DiCarlo James J, Zoccolan Davide, Rust Nicole C (2012) How Does the Brain Solve Visual Object Recognition?. *Neuron* **73**:415-434 <https://doi.org/10.1016/j.neuron.2012.01.010> | PubMed
- [25] Kriegeskorte Nikolaus (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science* **1**:417-446 <https://doi.org/10.1146/annurev-vision-082114-035447> | PubMed
- [26] Adelson Edward H, Bergen James R (1985) Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**:284-299 <https://doi.org/10.1364/JOSAA.2.000284> | PubMed
- [27] Thorpe Simon, Fize Denis, Marlot Catherine (1996) Speed of processing in the human visual system. *Nature* **381**:520-522 <https://doi.org/10.1038/381520a0> | PubMed
- [28] Isik Leyla, Mynick Anna, Pantazis Dimitrios, Kanwisher Nancy (2020) The speed of human social interaction perception. *NeuroImage* **215**:116844 <https://doi.org/10.1016/j.neuroimage.2020.116844> | PubMed

- [29] **Gandolfo Marco**, Abassi Etienne, Balgova Eva, Downing Paul E, Papeo Liuba, Koldewyn Kami (2024) Converging evidence that left extrastriate body area supports visual sensitivity to social interactions. *Current Biology* **34**:343-351.e5 <https://doi.org/10.1016/j.cub.2023.12.009> | [PubMed](#)
- [30] **Kravitz Dwight J**, Saleem Kadharbatcha S, Baker Chris I, Ungerleider Leslie G, Mishkin Mortimer (2013) The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences* **17**:26-49 <https://doi.org/10.1016/j.tics.2012.10.011> | [PubMed](#)
- [31] **Kravitz Dwight J**, Saleem Kadharbatcha S, Baker Chris I, Mishkin Mortimer (2011) A new neural framework for visuospatial processing. *Nature Reviews Neuroscience* **12**:217-230 <https://doi.org/10.1038/nrn3008> | [PubMed](#)
- [32] **Paszke Adam**, Gross Sam, Massa Francisco, Lerer Adam, Bradbury James, Chanan Gregory, Killeen Trevor, Lin Zeming, Gimelshein Natalia, Antiga Luca, *et al.* (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems*. **32** pp. 8024-8035
- [33] **Deng J**, Dong W, Socher R, Li L, Li Kai, Fei-Fei Li (2009) ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248-255 <https://doi.org/10.1109/CVPR.2009.5206848>
- [34] **Krizhevsky Alex**, Sutskever Ilya, Hinton Geoffrey E (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. **25**
- [35] **Nunez-Elizalde Anwar**, Deniz Fatma, Dupré la Tour Tom, Castello Matteo Visconti di Oleggio, Gallant Jack L (2021) pymoten: motion energy features from video using a pyramid of spatiotemporal Gabor filters. Zenodo. <https://doi.org/10.5281/zenodo.4437446>
- [36] **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825-2830
- [37] **Brainard DH** (1997) The Psychophysics Toolbox. *Spatial Vision* **10**:433-436 <https://doi.org/10.1163/156856897x00357> | [PubMed](#)
- [38] **Oostenveld Robert**, Fries Pascal, Maris Eric, Schoffelen Jan-Mathijs (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience* **2011**:156869 <https://doi.org/10.1155/2011/156869> | [PubMed](#)
- [39] **Conwell Colin**, Prince Jacob S, Kay Kendrick N, Alvarez George A, Konkle Talia (2024) A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications* **15**:9383 <https://doi.org/10.1038/s41467-024-53147-y> | [PubMed](#)
- [40] **King Jean-Rémi**, Charton François, Lopez-Paz David, Oquab Maxime (2020) Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage* **220**:117028 <https://doi.org/10.1016/j.neuroimage.2020.117028> | [PubMed](#)
- [41] **Abdel-Ghaffar Samy A**, Huth Alexander G, Lescroart Mark D, Stansbury Dustin, Gallant Jack L, Bishop Sonia J (2024) Occipital-temporal cortical tuning to semantic and affective features of natural images predicts associated behavioral responses. *Nature Communications* **15**:5531 <https://doi.org/10.1038/s41467-024-49073-8> | [PubMed](#)

Peer reviews

Reviewer #1 (Public review):

Summary:

The manuscript presents a study combining human functional MRI (fMRI), electroencephalography (EEG), and human judgement data within a two-step regression framework to determine when several regions of interest (ROIs) along the so-called third visual pathway for social perception are activated, and which aspects of visual representations of social interactions they encode. The goal is to determine whether social

interactions are encoded through a hierarchy of representation along the third pathway, and whether this hierarchy is implemented in a feedforward manner.

The first analysis linking EEG to fMRI data identifies early activation (within 60 ms) in early visual cortex (EVC) and later (onset between 112 and 310 ms) activations in selected, more anterior ROIs along the pathway, i.e., motion-selective MT, extrastriate body area (EBA), object-selective lateral occipital cortex (LOC), and posterior and anterior social-interaction superior temporal sulcus regions (pSTS and aSTS). The second analysis linking EEG-fMRI data to stimulus features derived either computationally from the images or from human judgments shows that EVC activity arises early (within 68 ms) and is predicted by low-level visual features and motion energy. Activity in all other ROIs emerges after 140 ms (or later) and is predicted by motion energy (in MT, EBA and LOC), arousal (in MT), spatial expanse (in LOC, pSTS and aSTS), agent distance (in LOC, pSTS and aSTS) and communication (in aSTS). No effects of facingness, object-directedness, joint action, or valence are found in any ROI.

Two observations are highlighted: features important for recognizing social interactions (i.e., agent distance) are encoded with a similar latency in the mid-level LOC and in STS; the communication content is selectively encoded in anterior STS.

It is concluded that social interactions are processed in a hierarchical way, with more abstract information (i.e., communication) encoded in more anterior regions. Moreover, because this higher-level information is encoded at similar or even shorter latencies than mid-level features in mid- and high-level ROIs, these findings are taken to challenge a strictly feedforward hierarchy within the third visual pathway.

Strengths:

This is a timely manuscript concerning a newly discovered and now extensively investigated brain network, purportedly dedicated to social perception. In the manuscript, the selectivity of this network for the perception of social interactions is assumed rather than demonstrated. Instead, the study addresses two related questions that are critical for understanding how the network (or pathway) is organized and functions: (i) whether representations are hierarchically organized, progressing from more specific information (lower-level visual features) to increasingly abstract information (interaction category, e.g., communication), and (ii) whether processing follows a feedforward mechanism with prior and rapid encoding of lower-level features of social interactions, followed by encoding of mid-level and higher-level features.

Addressing these questions is important because it goes beyond the knowledge of where social information is represented, toward understanding what kinds of visual input and transformations support the recognition of social interactions, and how the processing stages are organized (e.g., in a feedforward manner or with recurrent interactions). Establishing both aspects would define the particular computational architecture supporting the human ability to understand visual social interactions.

The manuscript is also methodologically valuable, as it introduces a variation of more conventional fMRI-EEG fusion analyses. This method establishes fMRI and EEG correspondences, thus revealing spatiotemporal dynamics of brain activity (i.e., where activity occurs in the brain, at which time point), and further links these dynamics to stimulus features (i.e., which aspects of the stimuli predict activity in a given brain area at a given time). The description and discussion of the method in the manuscript are clear and transparent. This facilitates replication and allows appreciation of the rationale of the approach and its potential for future research.

Weaknesses:

The method introduced in the manuscript appears to be a valuable addition to the range of non-invasive approaches to study the spatiotemporal dynamics of representations in the human brain. The results presented in the manuscript, however, do not make a strong case for the sensitivity of the method.

The study fails to replicate several effects that have demonstrated reliability across previous studies. In particular, motion energy in the stimuli only poorly predicts - relatively late - activity in motion-sensitive MT. Moreover, unlike several previous studies, including work by McMahon, Isik, and colleagues, the study finds no effects of facingness (or any other mid-level features) in EBA, object-directedness in LOC, joint actions in pSTS, and valence in pSTS. The manuscript does not discuss discrepancies, raising concerns about the sensitivity and validity of the method. In particular, the absence of previously reported effects is difficult to interpret without considering whether the method is sufficiently sensitive to detect the relevant stimulus dimensions.

Another related and unexplored issue concerns the potentially limited correspondence between EEG and fMRI signals, and between stimuli used to derive the stimulus features for the analyses and stimuli used in the actual EEG experiment. EEG and fMRI data were collected in different samples, and different individuals may assign different social value to the same events, potentially affecting neural processing in regions such as the pSTS (see e.g., work by Varrier, Finn et al.). Furthermore, the stimuli used in the EEG experiment were derived from, but differed from those used in the fMRI experiment. Although this is not entirely clear from the manuscript, it appears that stimulus features were measured for the original stimuli used in the fMRI experiment and assumed to be valid also for stimuli in the EEG experiment. The manuscript does not acknowledge these issues, or other potential sources of noise arising from them, either on their own right or in relation to the above null effects. This makes it difficult to determine whether the absence of several expected effects reflects genuine characteristics of neural representations or limitations of the measurement and analysis pipeline.

Beyond the methodological concerns, the results themselves are less straightforward than the manuscript suggests. Evidence for hierarchical representation emerges from the effect of low-level visual features in EVC and the encoding of communication in aSTS. Activity in all ROIs 'in the middle', from MT, EBA and LOC to pSTS, is predicted by a mixture of low- and mid-level stimulus features, whose effects in some cases come online after the encoding of the interaction category in aSTS (at 195 ms). This raises important questions concerning how a higher-level representation of an interaction category can emerge before the encoding of the lower- or mid-level information that is presumably diagnostic of that category. The manuscript would benefit from a more thorough discussion of this issue, particularly because the temporal ordering of these effects is central to the authors' claims about hierarchical and feedforward processing.

<https://doi.org/10.7554/eLife.112352.1.sa2>

Reviewer #2 (Public review):

Summary:

The authors seek to understand the hierarchical temporal representation of social communicative actions in the posterior superior temporal sulcus and other regions of the lateral pathway. They apply an analysis framework in which they predict fMRI responses from EEG responses to the same stimuli. (The EEG stimuli were shorter versions of the same clips shown as fMRI stimuli.) They examine which temporal segments of EEG data are most predictive of responses in which part of the brain. They find that early (< 100 ms post-stimulus) EEG activity is predictive of responses in early visual cortex, and that later (~ 150

ms post-stimulus) EEG activity is predictive of responses in the lateral visual stream. They then model the fMRI responses predicted by the EEG signals - which they term the "EEG-relevant" portion of the fMRI signal - as a function of various social and visual factors. They conclude that social interactions elicit responses in aSTS-SI later than low-level visual features elicit responses in early visual cortex, and they provide some evidence that mid-level social primitives elicit responses in mid-level areas somewhere in between.

Strengths:

I find the question and the framework interesting. The extraction of "EEG-relevant" fMRI signal as a target for visual and social feature-based encoding models has potential, though the approach in this instance suffers from low signal-to-noise. The evidence that the anterior portion of the superior temporal sulcus responds to communicative actions is clear and supports prior work in this regard.

Weaknesses:

I think there are several important points that the authors need to address to strengthen their case, and I think that many points could use more explicit description, explanation, and discussion.

Major concerns:

First, the basis of the EEG predictions of fMRI activity is not clear. The paradigm relies on predicting fMRI responses at each voxel from a weighted combination of EEG responses measured anywhere over the scalp. However, the authors never show any kind of weight map to visualize which electrodes on the scalp were weighted to predict responses in each fMRI voxel or ROI. I see that the use of PCA to orthogonalize the EEG responses creates problems with localizing the predictive signal in EEG to particular electrodes. Is the PCA step necessary? In any case, some form of analysis to estimate where in the brain the predictive EEG signal is coming from seems important. It's fairly clear that the best signal quality in the EEG recordings is in the early visual cortex. Thus, it could be the case that electrodes over EVC provide the main basis for predictions across the entire brain, with different temporal phases of EVC activity associated with activity in different regions in the fMRI data. It could also be the case (tho less likely in my opinion) that activity in regions away from EVC on the scalp provides the basis for predicting more anterior brain regions in the fMRI data. These two possibilities each suggest a substantially different interpretation of the present results: in the former case, there may be some sort of feedback loop between early visual areas and lateral areas; in the latter case, the hierarchical interpretation the authors have laid out seems more plausible. It seems important to draw a distinction between these possibilities.

In any case, the authors should say more in the discussion about the sources of EEG signals and fMRI signals and the relationship between the two.

On a similar note: the authors show that their EEG data was reliable based on a split-half analysis of the data. However, the authors only show average split-half correlation across electrodes and subjects. It stands to reason that some electrodes were much more reliable than others. The authors should show a map of electrode reliability across the scalp. One relevant question is: are there reliable electrode locations throughout the visual cortex? Or are the EEG signals only, or by far most reliable, in EVC? (How reliable is the EEG signal in regions that are at least somewhat near to the location of the lateral regions that is the principal object of investigation here?)

The use of predicted fMRI signals (predicted by EEG) as a target for image & social feature encoding models is only valid to the extent that the EEG signals accurately predict the fMRI signals. The prediction accuracy going from EEG to fMRI is not very high - a correlation of less than 0.2 in most regions besides EVC and the anterior extent of the activity along the ventral

temporal lobe. In the best-predicted regions, the activity only reaches a correlation of 0.3. The majority of the signal ($> 90\%$ of the variance) is unmodeled in this analysis. Thus, I would hesitate to characterize the evidence for activation by particular stimulus features at particular times in this analysis as "strong". The authors themselves are somewhat ambivalent on this point, as well. While they write in the conclusion: "Finally, we find strong evidence that social interactions are computed via hierarchical computation of low- and mid-level social visual features in the lateral visual pathway." However, This statement conflicts somewhat with other statements the authors make in the discussion section: "Together, these results suggest that lateral visual regions may not be organized in a strict feedforward hierarchy as has previously been suggested" Later in the discussion there is more ambivalence about the hierarchical nature of the representation: "Our findings suggest that the processing of social action features from posterior to anterior brain regions does not adhere to a strict feedforward hierarchy." The authors should present a more consistent interpretation of their findings.

Finally, the low performance of most of the mid-level models in mid-level regions is a concern; it points to poor signal overall. The common threads in the models that do predict responses above chance seem to be space (spatial expanse and agent distance) and salience (to the extent that arousal may constitute salience). This raises a question: Were the social model features correlated with either of the low-level model features?

I think that the overriding issue here is poor signal quality in the EEG data. The estimate of repeatability from the split-half analysis is not high, even on average for the whole brain; it is likely to be very poor for areas outside EVC. I think this matters. An important null hypothesis to consider is that there may be a few fairly low-level, or at least simple and straightforward, elements of the response that are reflected in the noisy EEG signal. Chief among these is likely to be that the most salient or high-contrast retinotopic regions in the image, which both AlexNet and motion energy models would likely pick up on, are likely to elicit responses in EVC. This is likely why EVC is by far the best-predicted region, both based on the raw EEG data and in the two-step regression analysis. There could be elements of spatial priority at all levels of the visual system. This could be what the authors are observing: there is some relatively small (in terms of variance explained) component of the signal that is common to many regions, and which emerges somewhat later in later regions due to conduction delays.

I think an important exception to this signal problem is the finding of communicative actions eliciting responses in aSTS-SI. That seems very robust and clear.

Medium comments:

There should be a section in the methods on how the ROIs were defined. (There is a brief description in the results section, but this is not sufficient). The ROI outlines should also be displayed somewhere in the paper (not just circles showing the general region as in Figure 1). Figure 3 would be a nice place for them, as would Figure S3.

It's also worth noting in the methods and discussion that the use of template ROIs for hMT in particular is potentially sub-optimal. hMT lies in anatomically variable locations across subjects. One reason for the seemingly weak signal in hMT may be that the voxels being averaged do not fall within the true hMT for each subject.

In the results (section 3.2.1), the authors write, "We scored the regression by predicting the EEG-predicted fMRI activity (estimated fMRItest) to the held-out test videos and then correlating the actual fMRI response (fMRItest) in the test set to this prediction at each time point (estimated fMRItest)." Unless I am missing something, I think the authors can safely say that they predicted responses to the withheld stimuli without specifying the intended prediction target (whether it is the fMRI data or the estimated fMRI data), because the prediction is simply based on weights, labeled or estimated image features, and stimuli. So,

the prediction might as well be a direct prediction of the fMRI test data. It seems to me that mentioning the estimated fMRI test in this sentence only muddies the waters.

In the two-stage regression analysis, were all the visual and social features fit to the EEG-predicted fMRI data in a single regression? Was this regression regularized? How many parameters total were fit? (This is clear for the AlexNet and motion energy models, but not clear for the other features - were they all single-feature models?)

<https://doi.org/10.7554/eLife.112352.1.sa1>

Author response:

We thank the reviewers and editors for their constructive feedback and are in the process of preparing a detailed point-by-point response. In the meantime, we would like to underscore our main conclusions and highlight the consistency with this study and the prior fMRI literature.

First, while the reviewers are correct to highlight the difficulty in interpreting null effects, the claims of our paper are primarily concerned with the positive findings. To our knowledge, this is the first evidence that social interaction perception is extracted in a fast hierarchy, which progresses from early visual features to mid-level social primitives to high-level information about communicative actions in the STS. This is an exciting and novel finding that sheds new light on the neural computations that support this core human ability. Importantly, the main null finding is the lack of difference in the onset latency between mid-level and anterior ROIs, rather than an inability to decode information in these regions (see below) and does not undercut this central claim.

Second, we would like to highlight that, overwhelmingly, our results are consistent with the prior fMRI literature. For example:

- Using our two-step regression procedure, we recover representations of motion energy in MT validating the use of this method
- Further, we find representations of mid-level features in mid-level regions ~150 ms after stimulus onset, primarily the distance between people.

This consistency, and the decoding of high-level features like communication exclusively in STS regions (as highlighted by reviewer 1), provides support for our method and shows that the EEG-fMRI correspondence extends beyond early visual regions.

The reviewers also raise questions about potential inconsistencies with our prior fMRI study (McMahon et al., 2023). We would like to clarify those findings here, because the current manuscript relies on the same fMRI data, and thus we would not expect to see effects here that are not reported in that fMRI data. In our original fMRI study, we did not find an effect of facingness, joint action, or valence in any ROI. Further, in that study, the effect for communication was most robust in aSTS, rather than pSTS. The primary effect from our original fMRI study that we do not find with EEG-fMRI fusion is information about object-directed actions in LOC. We will address these in a more detailed point-by-point response and revision soon.

Beyond the practical limitations of what we could find in the current fMRI dataset, we would also like to discuss the theoretical reasons for differences between our findings and the broader literature. First, our lack of finding an effect of valence likely reflects our stimuli, not neural representations. The Moments in Time computer vision dataset that our videos are derived from does not sample strongly negative videos. This restricted valence range is also evident in the behavioral ratings that participants provided.

Further, there are many differences between how we have looked for an effect of facing direction compared to prior work in this area. First, we do not use a binary facing/not facing but a graded scale from facing to not facing, and, to our knowledge, ours is the only study that has investigated a facing effect in naturalistic videos. It is possible that due to the graded scale or the natural stimuli that the facing effect is too subtle for us to detect. Another possibility is that while facingness is a social cue that is encoded in the brain, it is not a dominant cue in naturalistic videos where other cues are present (e.g., the distance between people, physical contact, and contingent motion).

Finally, on joint action, a type of physical social interaction between people, there is now convergent evidence from multiple studies that this may not be the primary type of social interaction represented in the STS. McMahon et al. (2023) found that communicative, not joint interactions, predicted STS responses. This was replicated and extended in (Zhao et al., 2025), where we find that joint action representations are more represented in lateral occipital regions, while communicative interactions are represented in the STS. From ongoing work in controlled stimuli (McMahon & Kanwisher, 2026 VSS), we have found that regions in the STS respond more to communicative interactions than joint actions. This present study provides new insight into the computations that support this, suggesting that information about communicative interactions is extracted along a spatiotemporal hierarchy, consistent with the fast timescale of visual processing, along the newly proposed lateral stream.

References:

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).

Zhao, Y., McMahon, E. & Isik, L. Separate Neural Representations for Physical and Communicative Social Interactions: Evidence from Data-driven Voxel Decomposition. Preprint at https://doi.org/10.31234/osf.io/fmnq9_v1 (2025).

McMahon, E. & Nancy Kanwisher. The STS Responds Preferentially to Communicative Signals. Vision Science Society; May 15–19, 2026; St. Petersburg, FL.

<https://doi.org/10.7554/eLife.112352.1.sa0>