A SPATIOTEMPORAL HIERARCHY FOR SOCIAL INTERACTION PERCEPTION IN THE LATERAL VISUAL STREAM

Emalie McMahon*[1, 2], Elizabeth Jiwon Im[3], Michael F. Bonner[1], and Leyla Isik[1]

[1]Johns Hopkins University
[2]Massachusetts Institute of Technology
[3]Stanford University

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 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

*Correspondence: emaliem@mit.edu

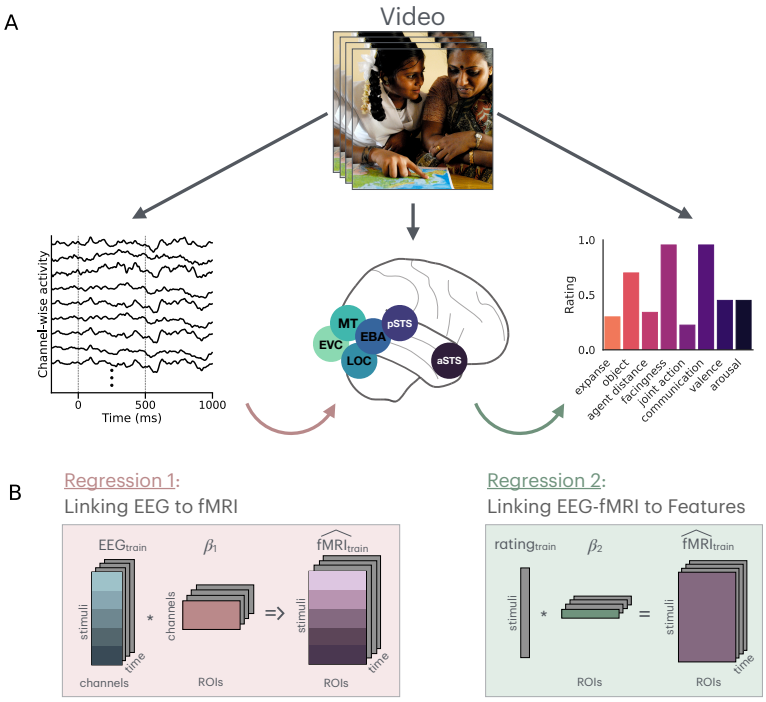

Figure 1: **Data and method overview A)** One example video depicting two person actions (top). Participants viewed the videos while their neural responses were recoded with EEG (middle-left) or fMRI (middle-middle), and participants online provided behavioral ratings for eight different features of 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).

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 social action representations using electroencephalography (EEG) (1).

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 (16) that has been used to investigate how object (17) and social representations (18) 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 (19; 20).

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.

2 RESULTS

2.1 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 4.4.1), 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 S5) and that the data are of high enough quality for the encoding/decoding framework adopted below.

2.2 SOCIAL ACTION FEATURES ARE DECODABLE FROM EEG ALONG A TEMPORAL HIERARCHY

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.

To determine the temporal latency of each of these feature representations, we used an EEG decoding procedure in which, for each EEG participant and time sample, we used the whole-brain 64-channel EEG signal to predict the uni-dimensional stimulus features.

Using this approach, we were able to decode all the annotated features in our dataset, except the extent to which the two people in the video were facing (Figure 2 and Supplemental Figure 6). In general, we tend to see that mid-level features of the scene and spatial relations among people in the scene (spatial expanse, $t_{onset} = 112$ ms, object directedness, $t_{onset} = 188$ ms, and agent distance, $t_{onset} = 128$ ms) have a shorter temporal latency than high-level social features (communication, $t_{onset} = 250$ ms, and arousal, $t_{onset} = 335$ ms), although joint action ($t_{onset} = 100$ ms) and valence ($t_{onset} = 158$ ms) both have relatively short latencies.

Latency estimations are known to be influenced by the overall decodability of a feature. Here, we see that some features have a higher decoding magnitude than others (e.g., agent distance versus communication), so 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. While this analysis reveals an earlier latency than the previous analysis for some features (e.g., communication is significantly decodable in the

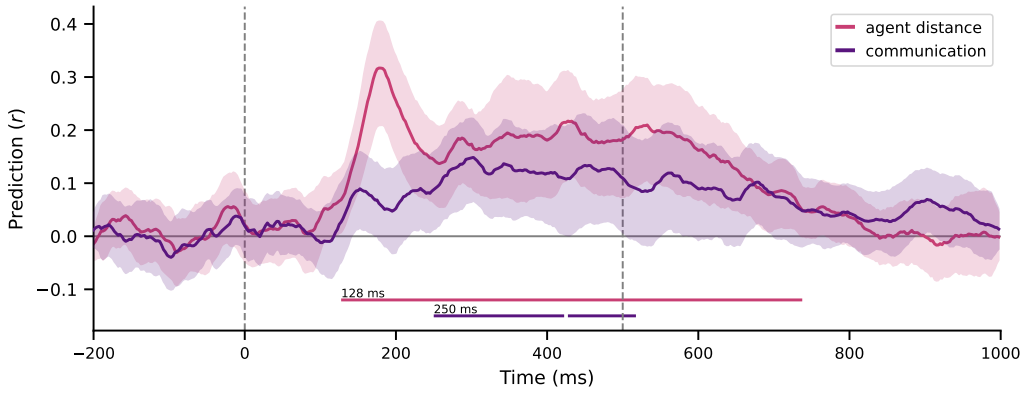

Figure 2: **Feature decoding from EEG.** Decodability of one mid-level (agent distance) and one high-level (communication) social action feature in our dataset from the EEG signal across time. Shaded regions are the 95% confidence intervals from bootstrapped variance distributions. Bold horizontal lines 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. See Figure S6 for decoding of all features.

150–200 ms time window), the overall trend of earlier decoding for lower-level features holds (Figure S7).

2.3 Early-late temporal dissociation in the lateral pathway

To understand the time course of neural processing across the lateral stream, we next 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 (21; 13), the lateral occipital cortex (LOC), which is selective for objects (22) 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 3A and Figure S8). 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 (23; 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.

As with the feature decoding, we again binned the decoding time course into 50 ms intervals from 0 to 350 ms. EVC is first decodable in the 50–100 ms interval, and all other ROIs are decodable by the 100–150 ms interval, confirming this early-late distinction (Figure S9).

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 3B and Figure S10). Together,

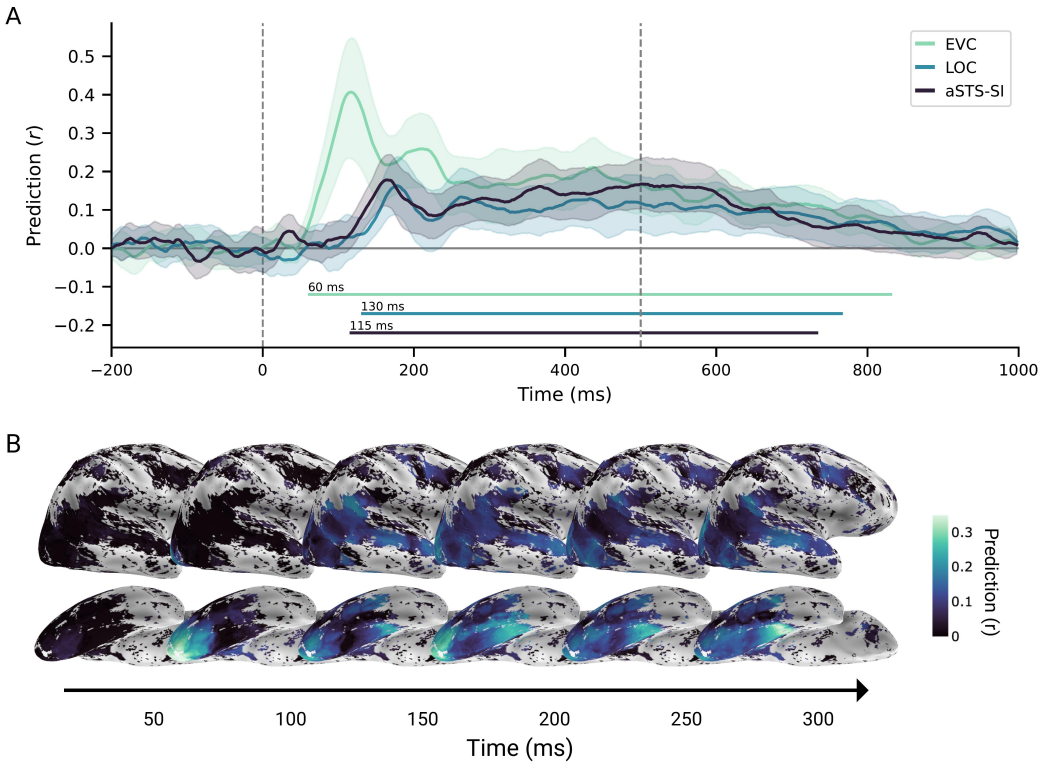

Figure 3: **ROI decoding from EEG. A)** Decodability of responses in one early (EVC, green), mid-level (LOC, blue), and high-level (aSTS-SI, dark blue) lateral stream ROI from the EEG signal across time. Plotting conventions are the same as Figure 2A. See Figure S8 for the timecourse of all ROIs. **B**) The prediction of voxelwise activity in the whole brain 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 S10.

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 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 3B and Figure S10).

2.3.1 A SPATIOTEMPORAL HIERARCHY OF SOCIAL ACTION FEATURES ACROSS THE LATERAL STREAM

Next, we turn to our central analysis that directly asks how information about social actions changes over space and time using our 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. Then we used the stimulus features to predict the EEG-predicted fMRI activity ($\widehat{fMRI}_{train}$). 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 4.4.6.

In addition to the behavioral annotations that were decoded in Section 2.2, we also used AlexNet-conv2 and motion energy as predictors of the EEG-predicted fMRI data as models of low-level visual processing (24) and motion activations in MT (25), respectively.

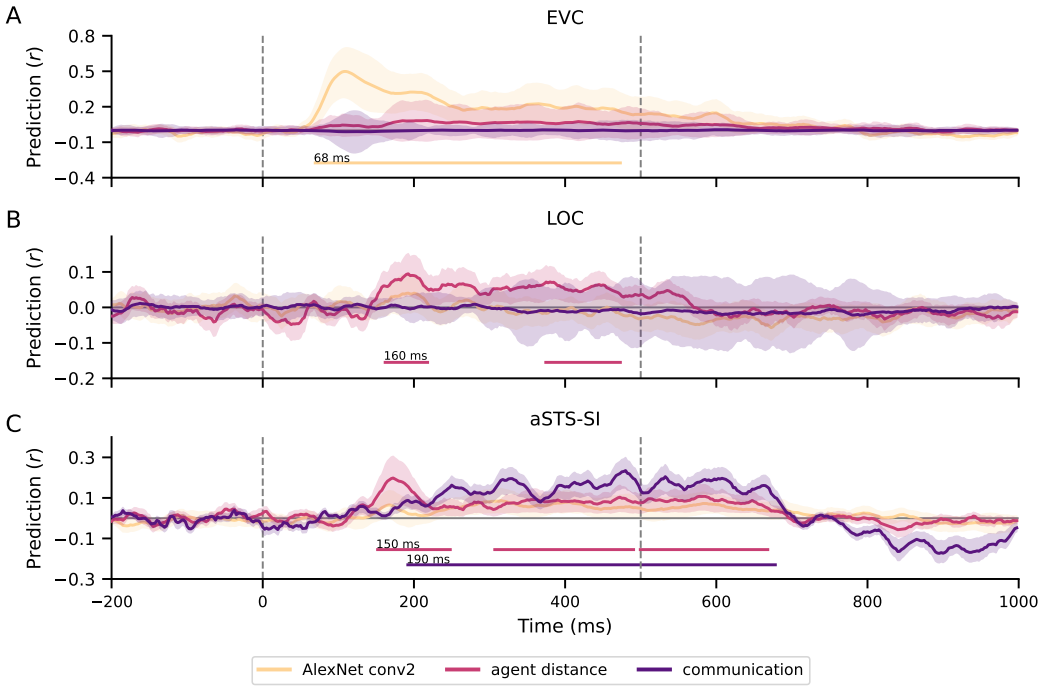

Figure 4: **Joint EEG-feature fMRI encoding.** Joint EEG-feature encoding in three select ROIs in (**A**) EVC, (**B**) LOC, and (**C**) aSTS-SI for a low-level feature (Alexnet-conv2, yellow), mid-level feature (agent distance, red), and high-level feature (communication, purple). All other plotting conventions follow those of Figure 2. See Figures S11–S16 for the time course of feature encoding in each ROI.

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 S11). 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 S12), 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 S13), and EBA by motion energy ($t_{onset} = 302$ ms, Figure S14).

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 S15) 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 S16).

The latency of social primitive representations is comparable between regions of LOTC (EBA and LOC) and STS (pSTS-SI and aSTS-SI), and earlier than the latency of high-level social features. Thus, while we again find a temporal hierarchy of features (Results 2.2), we find that mid-level features (i.e., social primitives) are broadly represented in both mid- and high-level lateral regions at the same time. Importantly, however, social interaction representations in aSTS-SI have a later onset and peak time than social primitive representations in the STS or elsewhere, supporting hierarchical feature representations across lateral stream regions (Figure S17).

3 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 hierarchy in feature decoding—more abstract features were decodable at longer latencies, both across the whole brain EEG and within the fMRI fused responses in the lateral stream. These results provide novel support for a spatiotemporal hierarchy of social action features along lateral visual regions.

3.1 EARLY REPRESENTATIONS OF COMMUNICATIVE ACTIONS IN THE STS

We found that communicative interactions are robustly 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 among others is a visual process (12). Here we show for the first time that whether or not two people are communicating can be decoded from the human STS within 200 ms of video onset. This fast representation is consistent with estimates of feedforward visual processing (23; 26), 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 (27; 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).

3.2 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 (28). One key difference between Gandolfo et al. (28) 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.

3.3 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 S 11). This validates the specificity and utility of the method for investigating spatiotemporal representations in the human brain.

Somewhat counterintuitively, we find that some features predict activity in some ROIs earlier using this method than directly decoding the features from the EEG activity. For example, the onset of communication representations in aSTS-SI was found to be earlier using the two-step regression than when decoding communication from the whole-brain EEG activity directly (Figures 2 and 4C, F). As another example, MT was poorly predicted by EEG directly and the onset latency was much later than would be expected based on electrophysiology (Figures S8) (23), but following the two-step regression, motion energy prediction of MT is higher and the onset was found to be consistent with feedforward processing. Together, these results highlight 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 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.

3.4 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) (23) 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 (23). 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 (29; 30). In light of these complexities, more research is needed to characterize the anatomical and functional profile of lateral visual regions.

3.5 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.

4 METHODS

4.1 REPRODUCIBILITY

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

4.2 STIMULI AND FMRI

4.2.1 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 (18). The stimulus set includes behavioral annotations of features of the visual and social scene (see Section 2.2 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 (31) of an ImageNet (32)-trained AlexNet (33) as a model of early visual processing (24) (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 (25) implemented in pymoten (34) 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 (35) 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.

4.2.2 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 split-half 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 2.3 for a description of the ROIs investigated in the current study).

4.3 EEG EXPERIMENT

4.3.1 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).

4.3.2 EXPERIMENTAL PROCEDURE

Based on previous studies (18; 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 kΩ 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 x 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 (36).

4.4 Analysis

4.4.1 EEG preprocessing

Minimal EEG data preprocessing was performed using MATLAB R2023b and FieldTrip (37). 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.

4.4.2 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.

4.4.3 Feature decoding

Within each EEG participant at each time point, we used the 64-channel EEG activity to predict visual and social features 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 (38) 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 in using a 25 ms sliding average window. This was down 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.

4.4.4 ROI RESPONSE DECODING

As for feature decoding (Section 4.4.3), we used the rotated 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.

4.4.5 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 4.4.4) to predict the voxelwise activity for each fMRI participant. Visualizations are done by averaging across EEG participants for individual fMRI participants.

4.4.6 TWO-STEP EEG-FEATURE FMRI ENCODING

Inspired by the regression procedures used by others (39; 40), 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.

4.4.7 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 $\alpha = 0.05$.

REFERENCES

[1] Diana C Dima, Tyler M Tomita, Christopher J Honey, and Leyla Isik. Social-affective features drive human representations of observed actions. *eLife*, 11:e75027, May 2022. ISSN 2050-084X. doi: 10.7554/eLife.75027. URL https://doi.org/10.7554/eLife.75027. Publisher: eLife Sciences Publications, Ltd.

[2] Leyla Tarhan and Talia Konkle. Sociality and interaction envelope organize visual action representations. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-16846-w.

[3] Moritz F. Wurm, Alfonso Caramazza, and Angelika Lingnau. Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. *Journal of Neuroscience*, 37(3):562–575, 2017. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1717-16.2016. URL https://www.jneurosci.org/content/37/3/562. Publisher: Society for Neuroscience _eprint: https://www.jneurosci.org/content/37/3/562.full.pdf.

[4] Rekha S. Varrier and Emily S. Finn. Seeing Social: A Neural Signature for Conscious Perception of Social Interactions. *Journal of Neuroscience*, 42(49):9211–9226, December 2022. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0859-22.2022. URL

https://www.jneurosci.org/content/42/49/9211. Publisher: Society for Neuroscience Section: Research Articles.

[5] Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43):E9145–E9152, October 2017. ISSN 1091-6490. doi: 10.1073/pnas.1714471114.

[6] Emalie McMahon, Michael F. Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, December 2023. ISSN 0960-9822. doi: 10.1016/j.cub.2023.10.015. URL https://www.cell.com/current-biology/abstract/S0960-9822(23)01373-8.

[7] Haemy Lee Masson and Leyla Isik. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, 245:118741, December 2021. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2021.118741.

[8] Jon Walbrin, Paul Downing, and Kami Koldewyn. Neural responses to visually observed social interactions. *Neuropsychologia*, 112:31–39, April 2018. ISSN 1873-3514. doi: 10.1016/j.neuropsychologia.2018.02.023.

[9] David Pitcher and Leslie G. Ungerleider. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, 25(2):100–110, February 2021. ISSN 1879-307X. doi: 10.1016/j.tics.2020.11.006.

[10] Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. The MIT Press, 1982.

[11] Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 0166-2236. doi: https://doi.org/10.1016/0166-2236(92)90344-8. URL https://www.sciencedirect.com/science/article/pii/0166223692903448.

[12] Emalie McMahon and Leyla Isik. Seeing social interactions. *Trends in Cognitive Sciences*, 27(12):1165–1179, December 2023. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2023.09.001. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(23)00248-6.

[13] Etienne Abassi and Liuba Papeo. The Representation of Two-Body Shapes in the Human Visual Cortex. *Journal of Neuroscience*, 40(4):852–863, 2020. doi: 10.1523/JNEUROSCI.1378-19.2019. Publisher: Society for Neuroscience.

[14] Etienne Abassi and Liuba Papeo. Behavioral and neural markers of visual configural processing in social scene perception. *NeuroImage*, 260:119506, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119506. URL https://www.sciencedirect.com/science/article/pii/S105381192200622X.

[15] Julia Landsiedel, Katie Daughters, Paul E. Downing, and Kami Koldewyn. The role of motion in the neural representation of social interactions in the posterior temporal cortex. *NeuroImage*, 262:119533, November 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119533. URL https://www.sciencedirect.com/science/article/pii/S1053811922006486.

[16] Radoslaw M. Cichy and Aude Oliva. A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron*, 107(5):772–781, September 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.001. URL https://www.sciencedirect.com/science/article/pii/S0896627320305183.

[17] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, March 2014. ISSN 1546-1726. doi: 10.1038/nn.3635. URL https://www.nature.com/articles/nn.3635. Publisher: Nature Publishing Group.

[18] Haemy Lee Masson and Leyla Isik. Rapid Processing of Observed Touch through Social Perceptual Brain Regions: An EEG-fMRI Fusion Study. *Journal of Neuroscience*, 43(45):7700–7711, November 2023. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0995-23.2023. URL https://www.jneurosci.org/content/43/45/7700. Publisher: Society for Neuroscience Section: Research Articles.

[19] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, April 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005508. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005508. Publisher: Public Library of Science.

[20] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: https://doi.org/10.1016/j.neuroimage.2010.07.073. URL http://www.sciencedirect.com/science/article/pii/S1053811910010657.

[21] Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A Cortical Area Selective for Visual Processing of the Human Body. *Science*, 293(5539):2470–2473, 2001. doi: 10.1126/science.1063414. Publisher: American Association for the Advancement of Science.

[22] Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10):1409–422, 2001. doi: https://doi.org/10.1016/S0042-6989(01)00073-6.

[23] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434, February 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.01.010. URL https://www.sciencedirect.com/science/article/pii/S089662731200092X.

[24] Nikolaus Kriegeskorte. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1):417–446, 2015. doi: 10.1146/annurev-vision-082114-035447.

[25] Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, February 1985. doi: 10.1364/JOSAA.2.000284. URL http://opg.optica.org/josaa/abstract.cfm?URI=josaa-2-2-284. Publisher: OSA.

[26] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, June 1996. ISSN 1476-4687. doi: 10.1038/381520a0. URL https://www.nature.com/articles/381520a0. Publisher: Nature Publishing Group.

[27] Leyla Isik, Anna Mynick, Dimitrios Pantazis, and Nancy Kanwisher. The speed of human social interaction perception. *NeuroImage*, 215:116844, 2020. doi: 10.1016/j.neuroimage.2020.116844.

[28] Marco Gandolfo, Etienne Abassi, Eva Balgova, Paul E. Downing, Liuba Papeo, and Kami Koldewyn. Converging evidence that left extrastriate body area supports visual sensitivity to social interactions. *Current Biology*, 34(2):343–351.e5, January 2024. ISSN 0960-9822. doi: 10.1016/j.cub.2023.12.009. URL https://www.sciencedirect.com/science/article/pii/S0960982223016640.

[29] Dwight J. Kravitz, Kadharbatcha S. Saleem, Chris I. Baker, Leslie G. Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1):26–49, January 2013. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2012.10.011. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(12)00247-1. Publisher: Elsevier.

[30] Dwight J. Kravitz, Kadharbatcha S. Saleem, Chris I. Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, April 2011. ISSN 1471-0048. doi: 10.1038/nrn3008. URL https://www.nature.com/articles/nrn3008. Publisher: Nature Publishing Group.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[32] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[34] Anwar Nunez-Elizalde, Fatma Deniz, Tom Dupré la Tour, Matteo Visconti di Oleggio Castello, and Jack L. Gallant. pymoten: motion energy features from video using a pyramid of spatiotemporal Gabor filters, January 2021. URL https://doi.org/10.5281/zenodo.4437446.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] D. H. Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436, 1997. ISSN 0169-1015.

[37] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011:156869, 2011. ISSN 1687-5273. doi: 10.1155/2011/156869.

[38] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53147-y. URL https://www.nature.com/articles/s41467-024-53147-y. Publisher: Nature Publishing Group.

[39] Jean-Rémi King, François Charton, David Lopez-Paz, and Maxime Oquab. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, 220:117028, October 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117028. URL https://www.sciencedirect.com/science/article/pii/S1053811920305140.

[40] Samy A. Abdel-Ghaffar, Alexander G. Huth, Mark D. Lescroart, Dustin Stansbury, Jack L. Gallant, and Sonia J. Bishop. Occipital-temporal cortical tuning to semantic and affective features of natural images predicts associated behavioral responses. *Nature Communications*, 15(1):5531, July 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49073-8. URL https://www.nature.com/articles/s41467-024-49073-8. Publisher: Nature Publishing Group.

5 SUPPLEMENTAL FIGURES

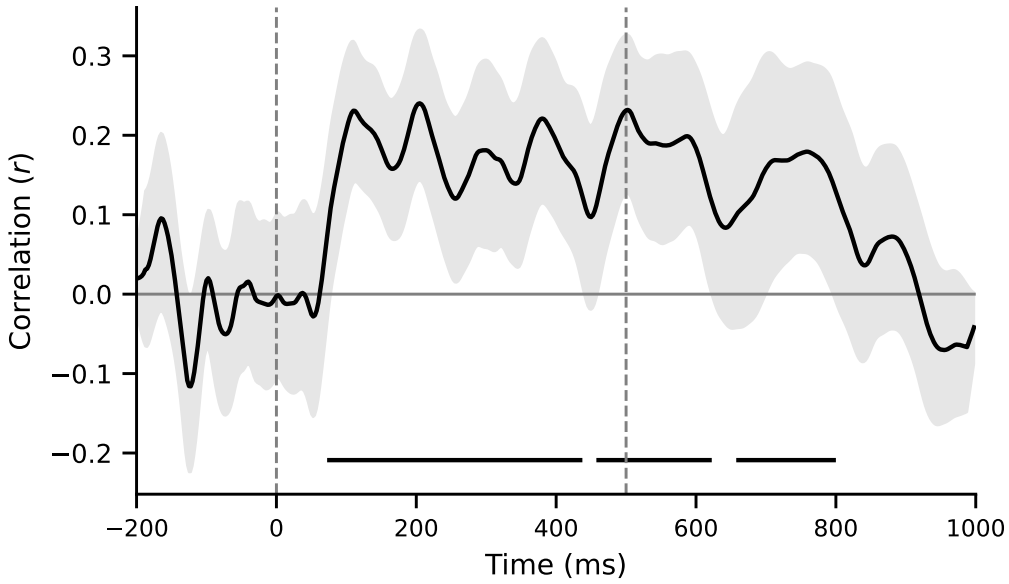

Figure 5: **EEG data reliability.** Data reliability in the test set averaged across participants and channels. Shaded areas are the bootstrapped 95% confidence intervals and solid horizontal lines represent time windows of significant reliability after cluster correction at a level of $p < 0.05$. Horizontal 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.

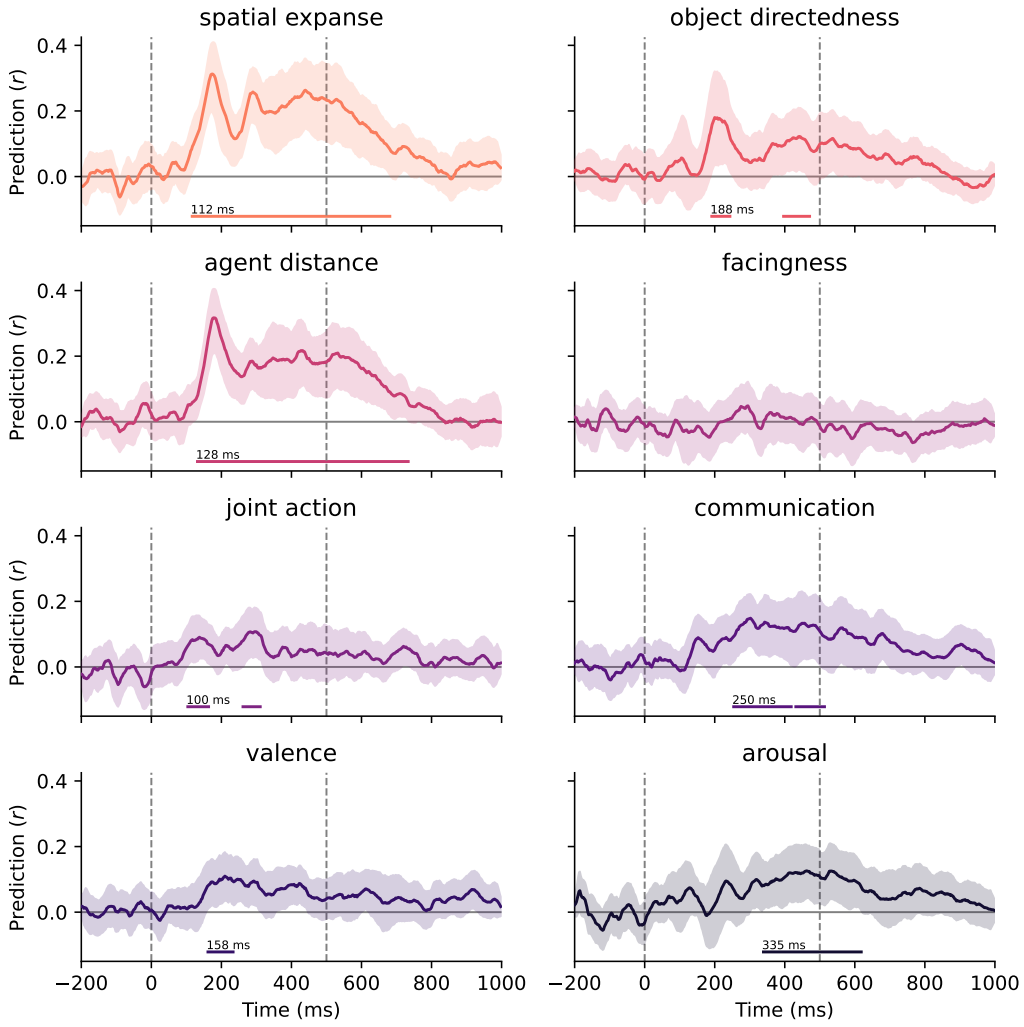

Figure 6: **Decoding of all features from EEG.** Decodability of all annotated features from EEG across time extended from the subset shown in Figure 2A. Plotting conventions are the same as Figure 2A.

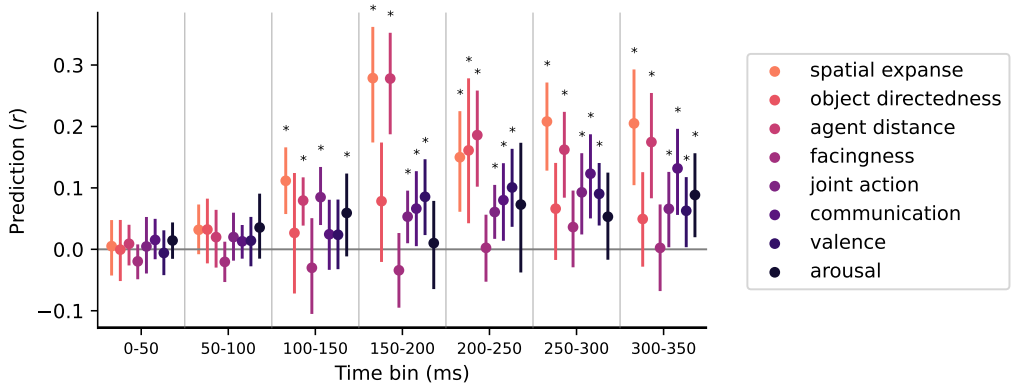

Figure 7: **Latency of all features from EEG decoding.** Decodability of annotated from the EEG in time bins as in Figure 2B extended to all features.

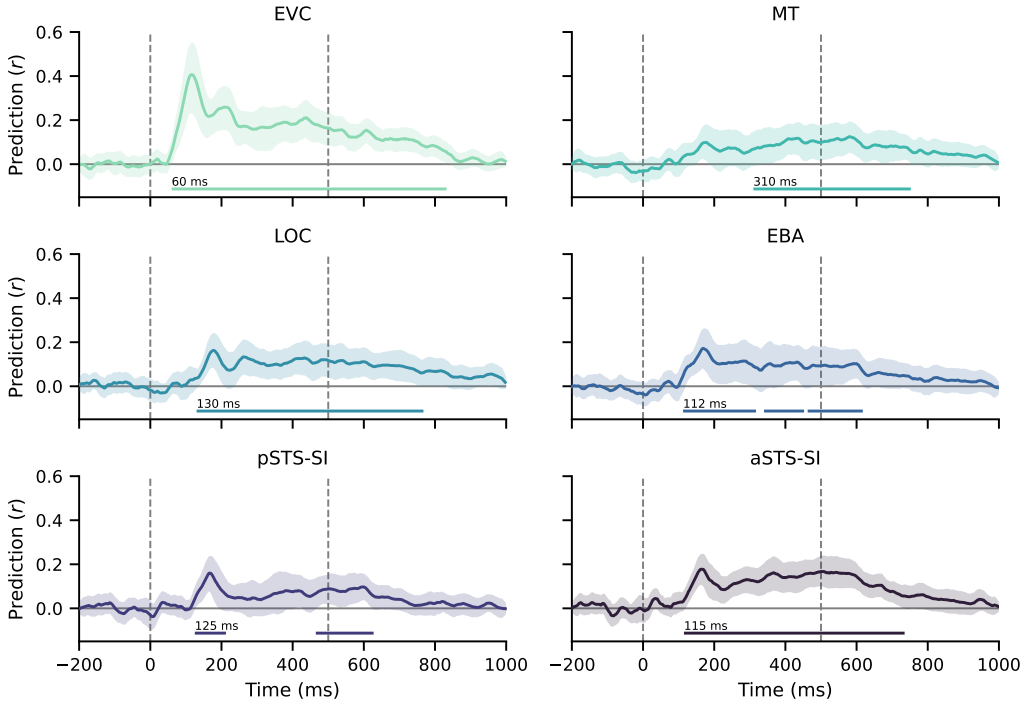

Figure 8: **Decoding of all ROIs from EEG.** Decodability of responses in all ROIs from the EEG signal across time extended from Figure 3A.

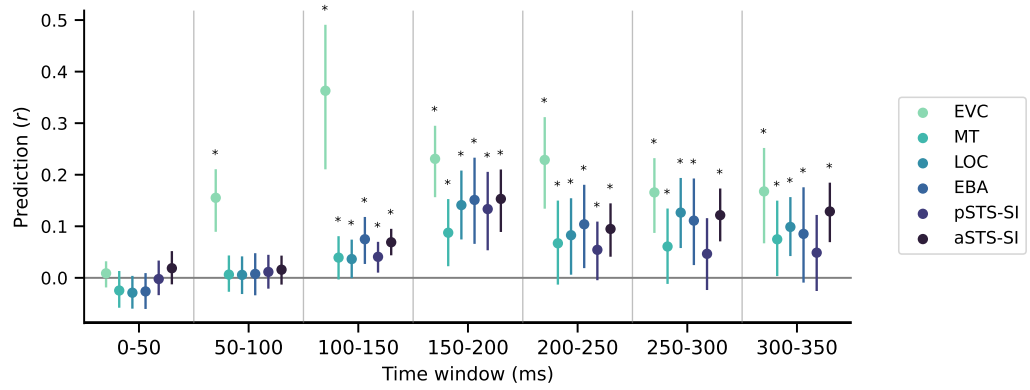

Figure 9: **Latency of all ROI prediction from EEG decoding.** Decodability of ROI activity recorded in time bins extended from Figure 3B.

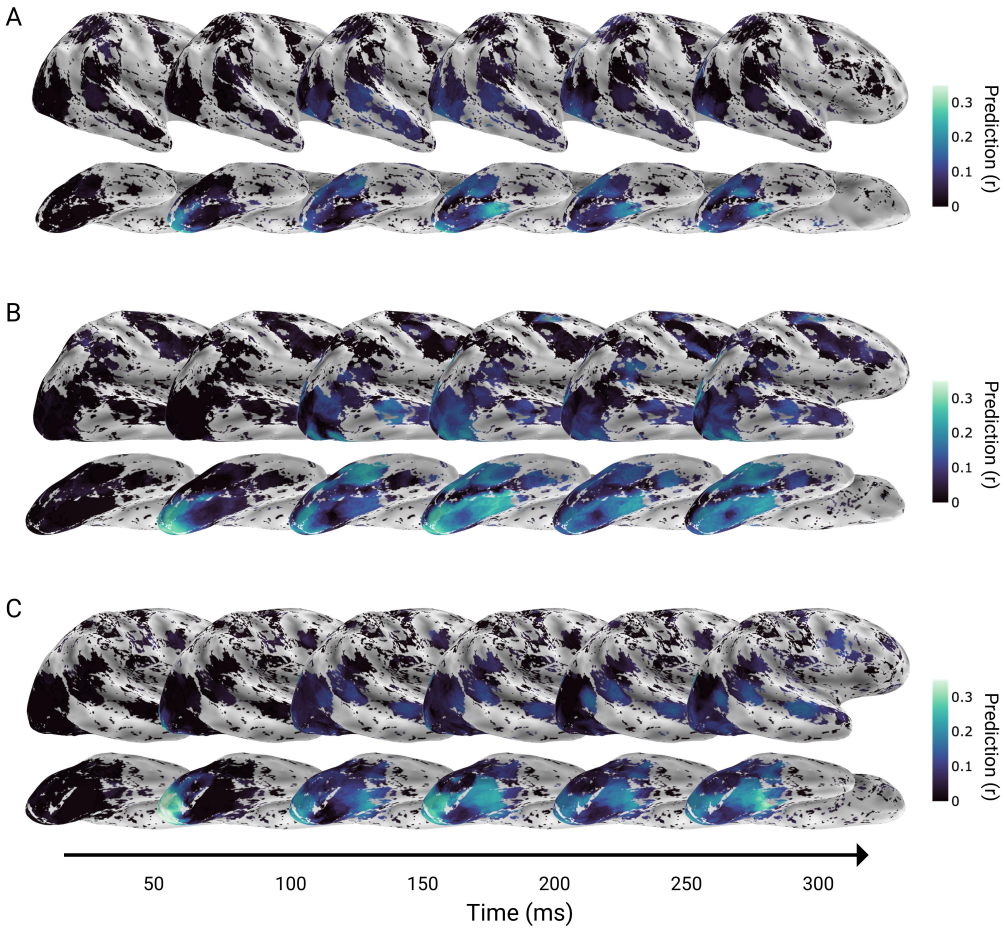

Figure 10: **Whole brain prediction in other fMRI participants.** Plotting conventions are all the same as Figure 3C, but for **(A)** sub-01, **(B)** sub-03, and **(C)** sub-04 from McMahon et al. (6).

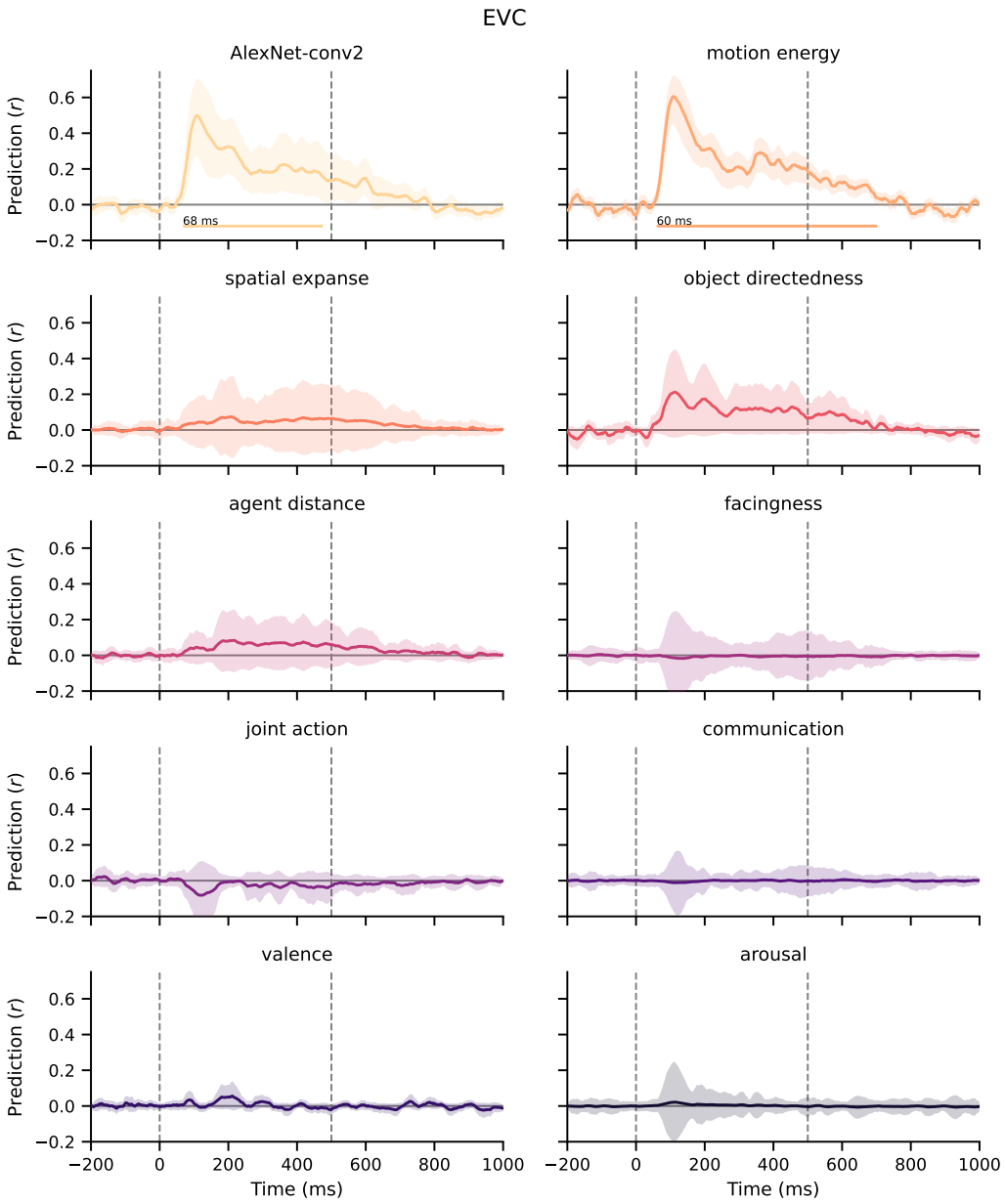

Figure 11: **Time course of joint EEG-feature fMRI encoding of each feature in EVC** Extended to all features from Figure 4A. For visibility, all features are shown on separate subplots.

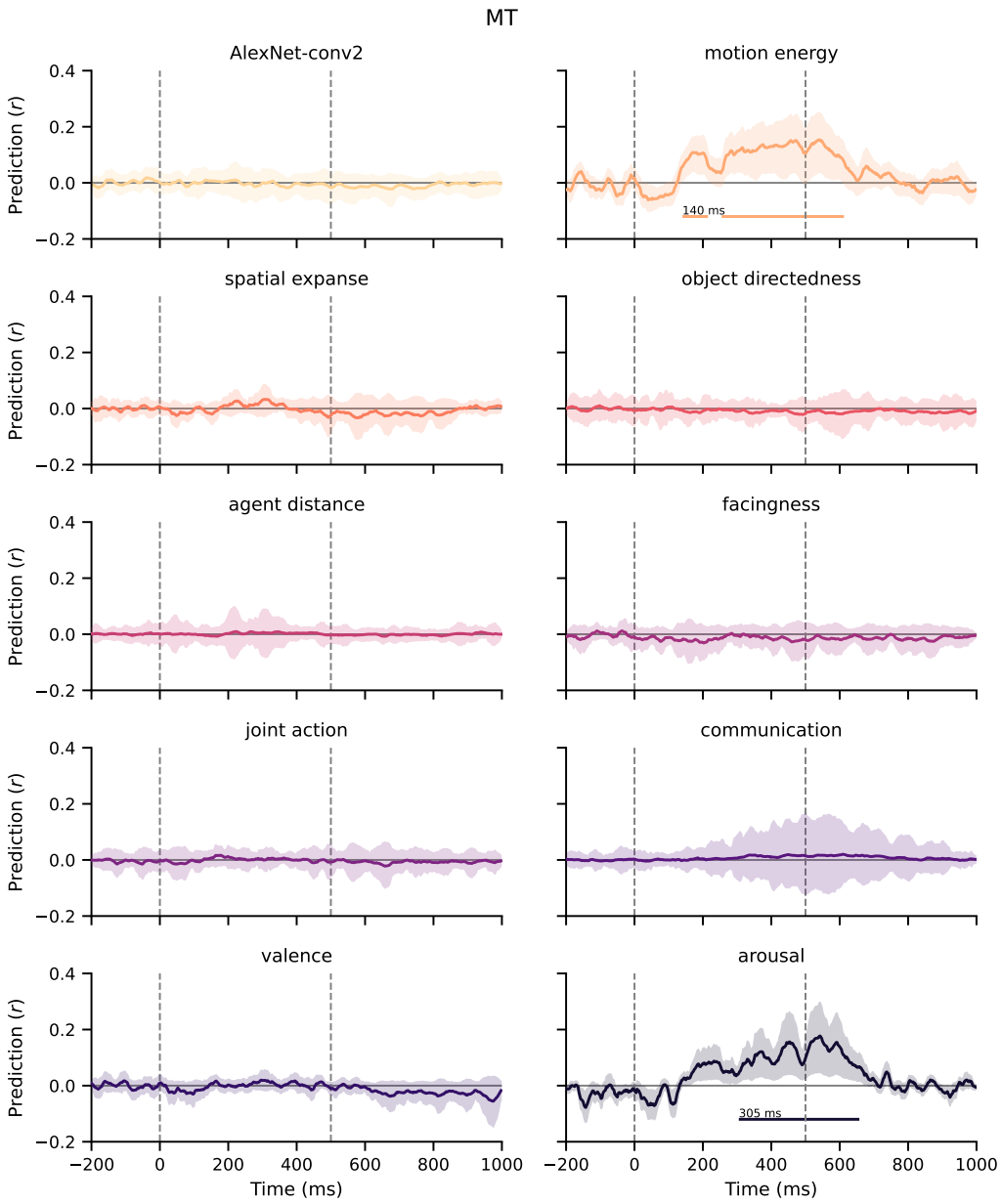

Figure 12: **Time course of joint EEG-feature fMRI encoding of each feature in MT** Related to main text Figure 4.

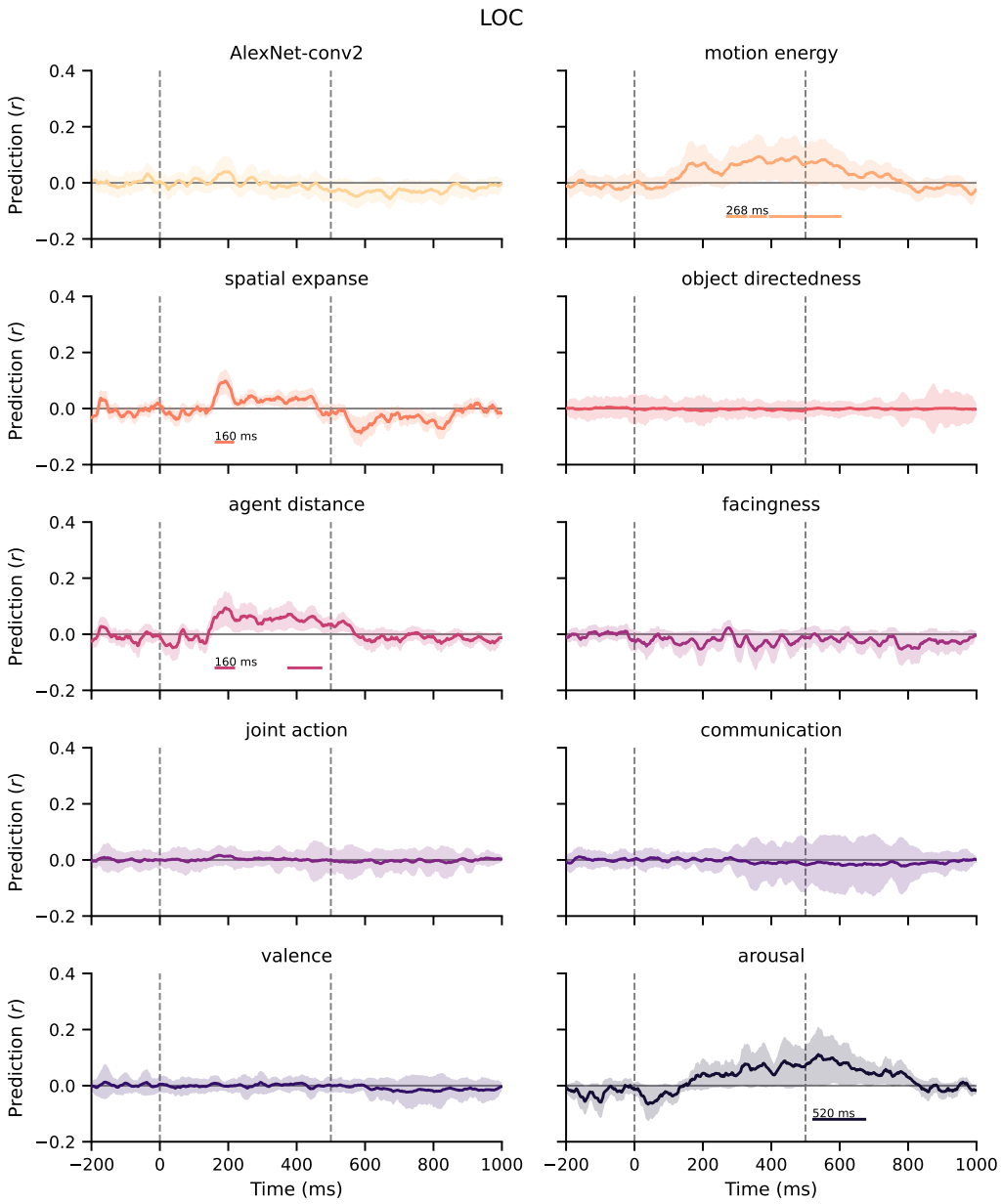

Figure 13: **Time course of joint EEG-feature fMRI encoding of each feature in LOC** Extended from main text Figure 4B for al features.

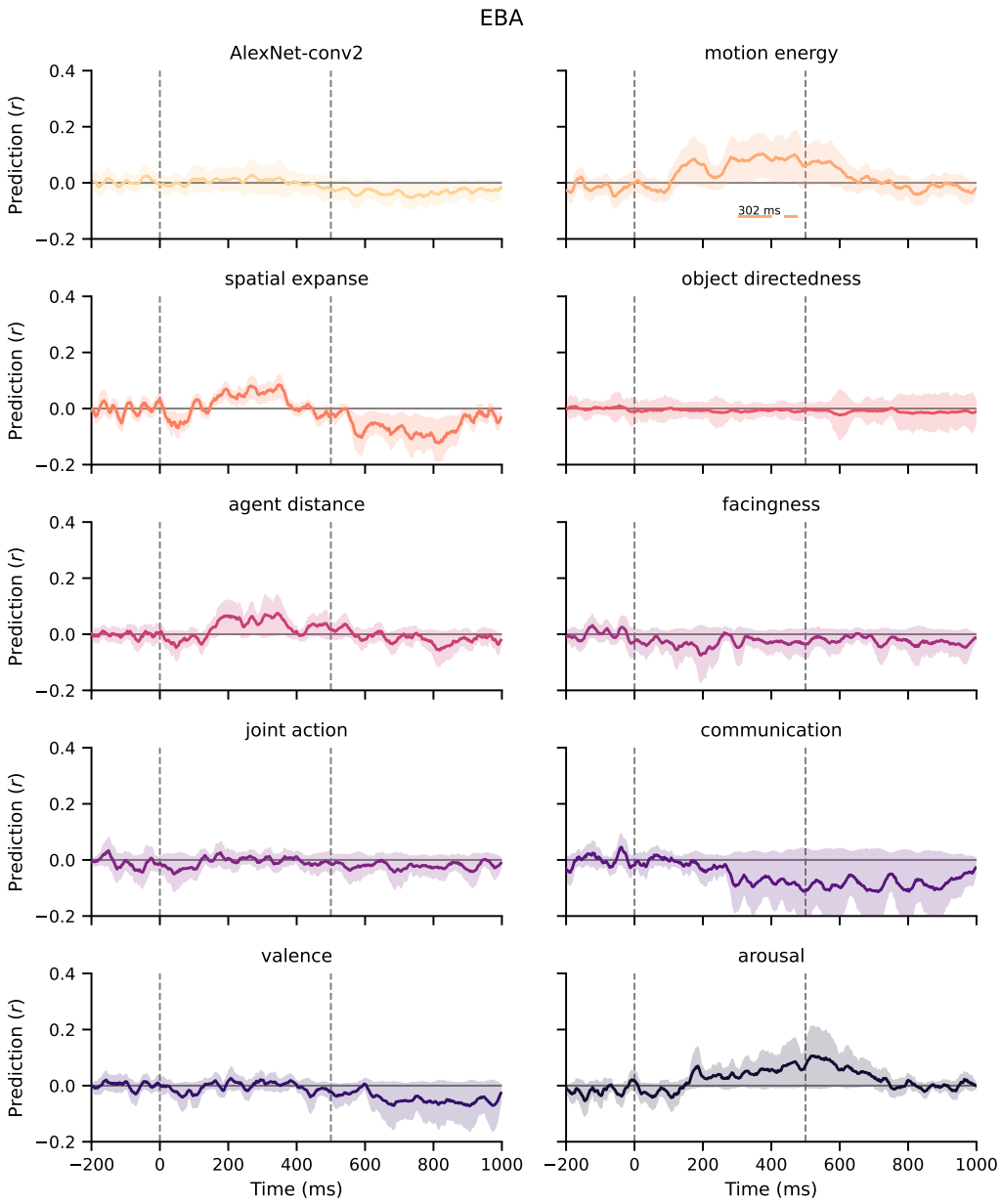

Figure 14: **Time course of joint EEG-feature fMRI encoding of each feature in EBA** Related to main text Figure 4.

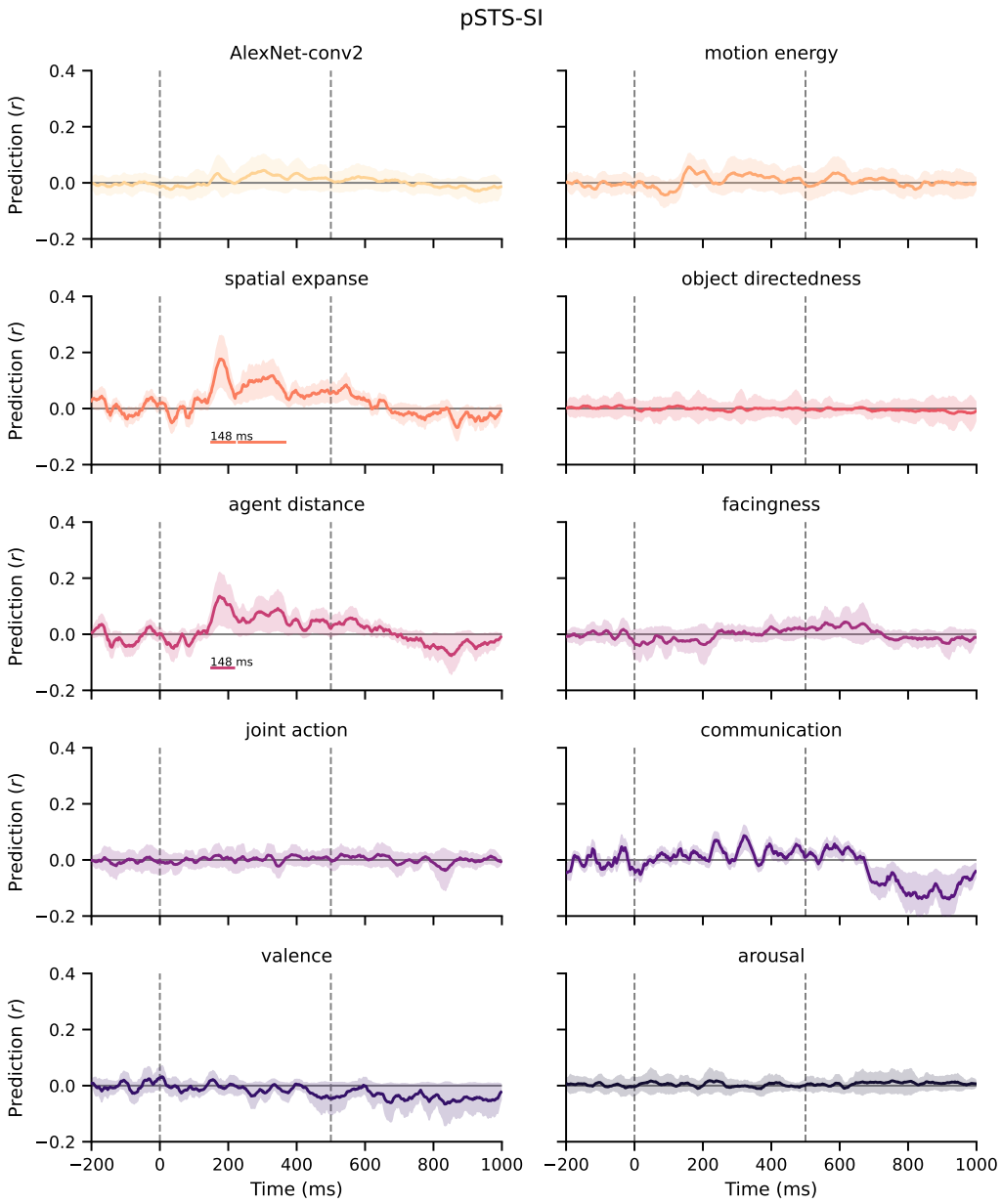

Figure 15: **Time course of joint EEG-feature fMRI encoding of each feature in pSTS-SI** Related to main text Figure 4.

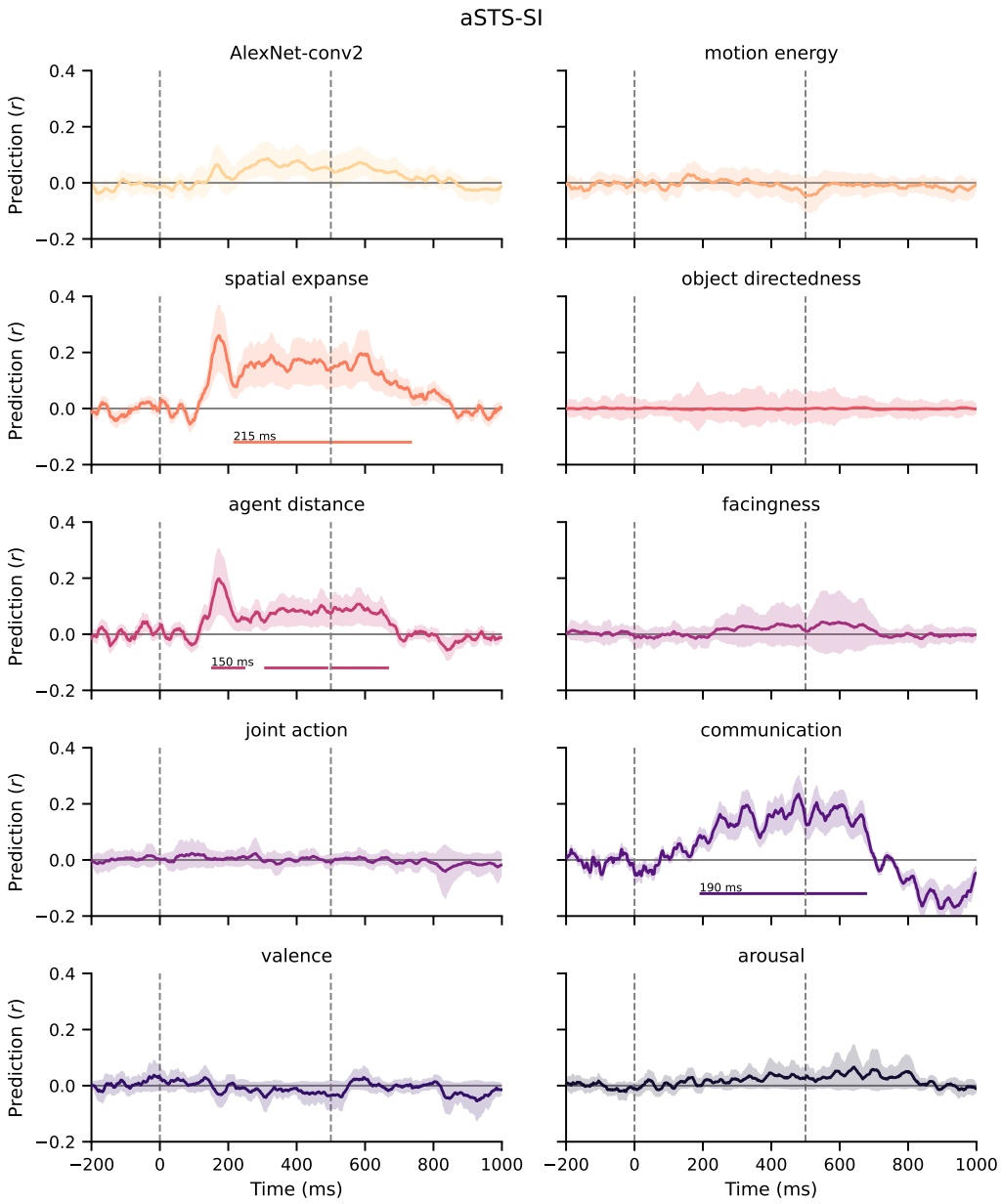

Figure 16: **Time course of joint EEG-feature fMRI encoding of each feature in aSTS-SI** Extended from main text Figure 4C for all features.

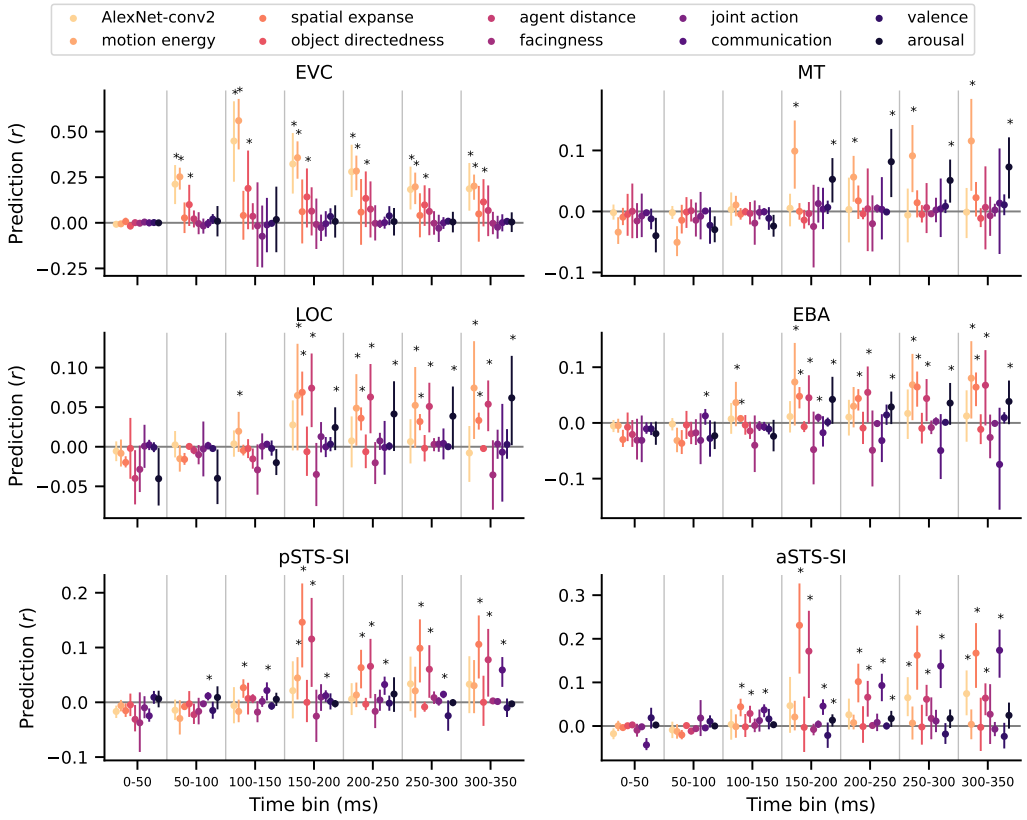

Figure 17: **Latency of feature prediction in ROIs.** Extended from main text Figure 4D–F for all features and all ROIs.