A Scalable Framework to Decode Neural Representation of Social Interaction

Michael (Yuhao) Zhu

Masters in Computational Social Science

Introduction

In our daily lives, we constantly observe people interacting with one another. But how does our brain represent social interactions? To answer this question, previous studies have primarily used the naturalistic paradigm (Carvalho & Lampinen, 2025; Monfort et al., 2020) to investigate which aspects of social interactions are processed by different regions of the brain. For example, one study (McMahon et al., 2023) demonstrated the existence of a social pathway that processes social interaction features in a hierarchical manner—ranging from low-level attributes like agent distance to high-level aspects such as communication—mirroring the organization of basic visual processing (McMahon & Isik, 2023). Another study (Lee Masson et al., 2024) found that higher-level features of social interaction, such as mentalization, are represented in the superior temporal gyrus (STG) and middle temporal gyrus (MTG).

Traditionally, hypotheses are tested by comparing human annotations with neural activity to identify correlations between specific social interaction features and brain regions. The naturalistic paradigm is well-suited for this type of investigation, as its information-rich stimuli allow for the possibility of testing multiple hypotheses simultaneously (Almaatouq et al., 2024; Carvalho & Lampinen, 2025). However, due to financial constraints, human annotations are typically limited to a small set of dimensions, restricting our ability to fully utilize the richness of the stimuli and explore the design space comprehensively. To overcome this limitation, we propose a framework that leverages deep neural networks to efficiently generate hypothesis-based annotations, enabling the exploration of a much broader—potentially infinite—range of hypotheses.

Methods

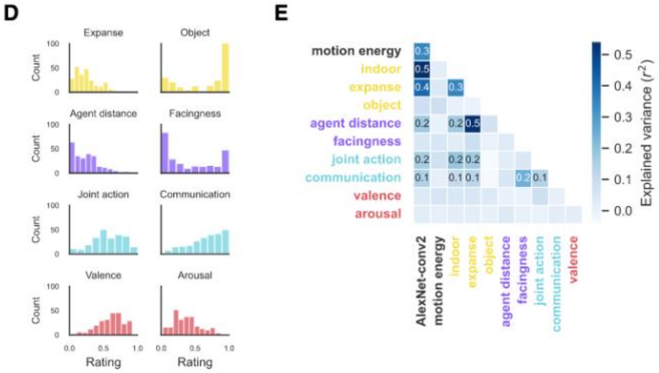
Dataset

Participants & Design

The dataset is from (McMahon et al., 2023). The study included 4 participants (2 females, mean age 25.5 years, range 23–30; 3 Caucasian, 1 Asian), each undergoing four 2-hour fMRI sessions (3T, TR =1.5s). Participants viewed 250[1] 3-second videos depicting dyadic social interactions without sound. The videos were selected from the Moments in Time dataset (Monfort et al., 2020) and annotated across low-, mid-, and high-level features, including visual, social, and affective properties (Figure 1). This small-n, condition-rich design enabled extensive within-subject data collection on neural representations of social interactions.

Figure 1

Video Distribution on Several Human-annotated Dimensions (McMahon et al., 2023)



fMRI data Preprocessing

In the original paper, a systematic preprocessing has been performed using fMRIPrep (Esteban et al., 2019) based on Nipype 1.6.1(Gorgolewski et al., 2011). Key steps included: 1. *Reference Volume & Skull Stripping.* A reference volume was generated by aligning and

[1] As some videos in the downloaded file are damaged, we only have 244 videos.

averaging single-band reference (SBRef) images, followed by skull stripping. 2. *Motion Correction*. Head motion parameters were estimated using mcflirt (Jenkinson et al., 2002), and images were realigned to correct for movement artifacts. 3. *Field Map Correction*. Estimated field maps were aligned with the echo-planar imaging (EPI) reference to correct for distortions. 4. *Slice-Time Correction*. BOLD images were temporally aligned using 3dTshift from AFNI (Cox & Hyde, 1997). 5. *Spatial Co-Registration*. The BOLD reference was co-registered to the T1-weighted structural image using bbregister (FreeSurfer) with boundary-based registration (Greve & Fischl, 2009). 6. *Noise Reduction & Confound Regressors*. Several nuisance regressors were computed, including framewise displacement (FD), DVARS (Power et al., 2014), and global signals from cerebrospinal fluid (CSF), white matter (WM), and whole-brain masks. CompCor (Behzadi et al., 2007) was used to extract physiological noise components. 7. *Motion Outlier Detection*. High-motion frames were identified and flagged based on predefined FD and DVARS thresholds. 8. *Spatial Normalization*. BOLD images were resampled to MNI152NLin2009cAsym space for standardization. 9. *Surface & Volumetric Resampling.* Functional data were mapped onto FreeSurfer's fsnative space and volumetric data were transformed using ANTs' antsApplyTransforms with Lanczos interpolation (Lanczos, 1964).

The regions of interest (ROIs) in this study follow those defined by McMahon et al. (2023). Anatomical ROIs, including EVC and MT, were identified based on structural landmarks. Anatomically constrained functional ROIs, such as pSTS and aSTS, were localized using functional tasks but constrained within anatomically defined regions. Functional ROIs, including biomotion-STS, TPJ, FFA, face-pSTS, PPA, EBA, and LOC, were identified purely through functional localizers without anatomical constraints.

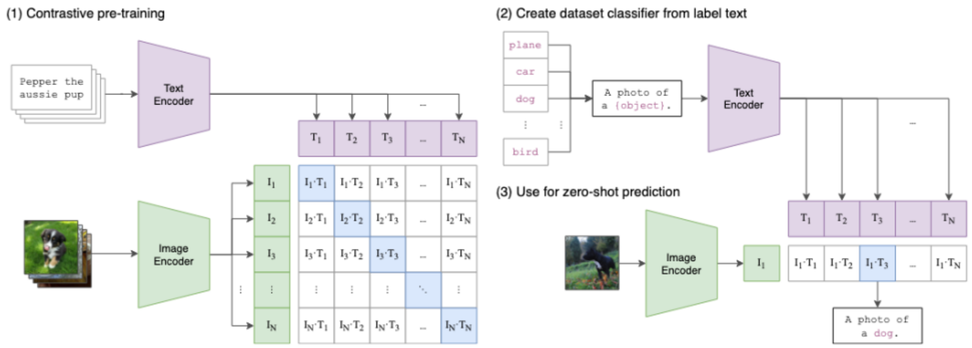
Deep learning

Here we propose a framework utilizing deep learning to develop automated, hypothesis-based annotations for naturalistic stimuli such as videos.

Traditional naturalistic paradigms rely on human annotations to describe social interaction features in videos, but this approach is costly thus restricts the number of hypotheses that can be tested (Almaatouq et al., 2024). To overcome this, we use CLIP model, a multi-modal model trained on paired image-text data (Radford et al., 2021, Figure 2), to generate video annotations automatically, allowing for systematic, scalable, and fine-grained hypothesis testing (Sievers & Thornton, 2024). Instead of directly computing the similarity between video embeddings and raw concepts, this method operationalizes the target concept as a structured set of items, enabling a more nuanced evaluation of video content along specific dimensions (Figure 3).

Figure 2

Conceptual Introduction: Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021)


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T₁ T₂ T₃ ... T_N
Image Encoder
I₁ I₂ I₃ ⋮ I_N
I₁·T₁ I₁·T₂ I₁·T₃ ... I₁·T_N
I₂·T₁ I₂·T₂ I₂·T₃ ... I₂·T_N
I₃·T₁ I₃·T₂ I₃·T₃ ... I₃·T_N
I_N·T₁ I_N·T₂ I_N·T₃ ... I_N·T_N
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I₁
I₁·T₁ I₁·T₂ I₁·T₃ ... I₁·T_N
A photo of a dog.

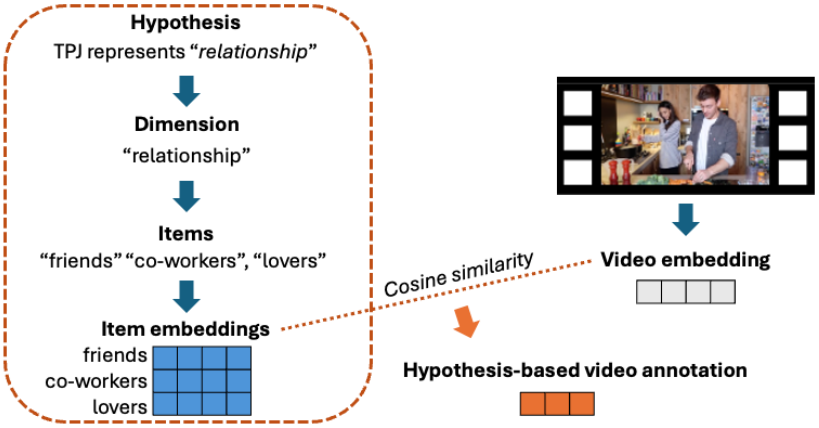
For a given video, we extract a video embedding vector by first processing individual frames and then aggregating their embeddings. Each video is sliced into a sequence of frames, and each frame is passed through CLIP's vision encoder to obtain an embedding vector. The frame embeddings are then aggregated using a mean pooling operation to generate a single video embedding vector.

Inspired by the idea of questionnaire-based psychological measurement, to annotate a video based on a hypothesis regarding certain dimension of social interaction, we first operationalize the hypothesis into a structured set of items. For example, if the hypothesis is about the "agents' relationship" dimension of social interactions, we define specific relationship types such as "friends", "coworkers" as items to define this dimension. Each item is converted into an embedding vector using CLIP's text encoder.

Once both video embeddings and hypothesis item embeddings are obtained, we compute how well the video aligns with each item under the hypothesis dimension using cosine similarity. The similarity score between the video embedding and each item embedding is calculated, resulting in a score vector. These scores are then normalized across all items so that they sum to one, ensuring a relative weighting of how the video fits within the hypothesis dimension. This normalized vector provides a structured description of the video under a specific hypothesis focus.

Figure 3

Hypothesis-based Annotation Based on CLIP



After obtaining an annotation vector for each video, we can compare videos within a hypothesis dimension by computing the similarity between their annotation vectors. Given two videos, their similarity is computed using 1- Euclidean distance, which measures how similarly two videos are annotated within the hypothesis dimension. If two videos have high similarity, it suggests that they are similar under the given hypothesis. If they differ in certain items, such as one scoring high on “friends” and another on “coworkers,” they are more distinct in the “agents’ relationship” dimension.

This method automates video annotation, replacing costly human annotations with a scalable, hypothesis-driven labeling process. It captures video properties in a structured manner rather than relying on generic embedding semantic similarity. With precise comparisons between videos based on specific dimensions of interest, we can test different hypotheses, such as determining which representation structure, "agent relationship" or "emotional value," can better explain the neural activity of a brain region, by comparing the similarity matrices derived from two annotation dimensions.

In this demonstration, we selected several dimensions to describe social interactions from literature (Cheng et al., 2023; Hadley et al., 2022; McMahon & Isik, 2023). *Valence* captures the emotional tone of the interaction (e.g., positive, negative). *Relationship* defines the social connection between individuals (e.g., parent-child, couple, coworkers, friends, neighbors). *Activity* refers to the specific actions people are engaged in (e.g., people dancing, people playing sports, people playing instruments, people cooking, people fishing). *Communication modality* distinguishes between different ways of information flow (e.g., verbal communication, non-verbal communication). *Demographics* account for the identity of the individuals involved (e.g., male, female, child). *Context* describes the environment in which the interaction takes place (e.g., indoor, yard, wild). *Joint action* differentiates whether individuals coordinate their actions or act independently (e.g., joint action, independent action). *Transitivity* describes whether people interact with objects, other people, or act alone (e.g., people interacting with objects, people interacting with other people, people acting independently). *Engagement* reflects the level of involvement in the interaction (e.g., people showing high engagement, people showing low engagement).

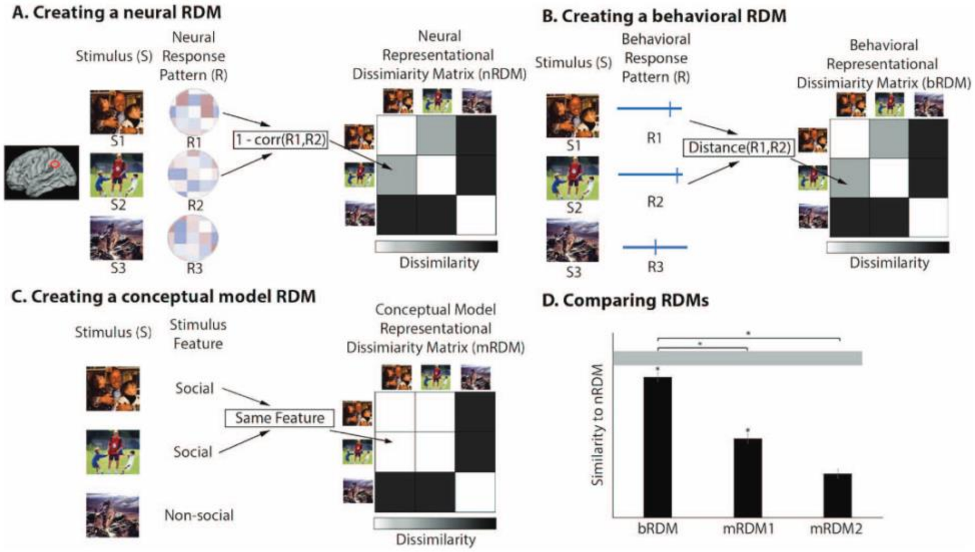
For comparison purpose, we also extract the original embedding of the CLIP model (model name: ViT-bigG-14-quickgelu, pretrained weights: metaclip_fullcc, see https://github.com/facebookresearch/MetaCLIP/tree/main), with a length of 1280. In addition, we extract the global average pooling layer embeddings from several visual models, two ResNet50 variants (see https://github.com/zhoubolei/moments_models), ResNet3d50, an inflated version to deal with video (Carreira & Zisserman, 2017), and multi-ResNet3d50, a version trained on multiple overlapping actions using the Broden dataset and wLSEP loss (Lee Masson et al., 2024). The visual models have embeddings with length 2048.

Representational similarity analysis

Representational Similarity Analysis (RSA) was used to compare neural activity patterns with different computational candidates (Kriegeskorte, 2008; Nili et al., 2014; Popal et al., 2019). The core idea is to quantify the similarity structure of representations by constructing representational similarity matrices (RSMs) and then compare candidate RSMs (models) with reference RSM (neural), see Figure 4. This approach allowed us to examine how representational patterns in different model-derived representations aligned with patterns in ROIs. Specifically, we constructed video-by-video RSMs across three levels: neural activity (reference), network embeddings, and hypothesis-driven CLIP annotations (Figure 5).

Figure 4

Conceptual Introduction: Representational Similarity Analysis (Popal et al., 2019)

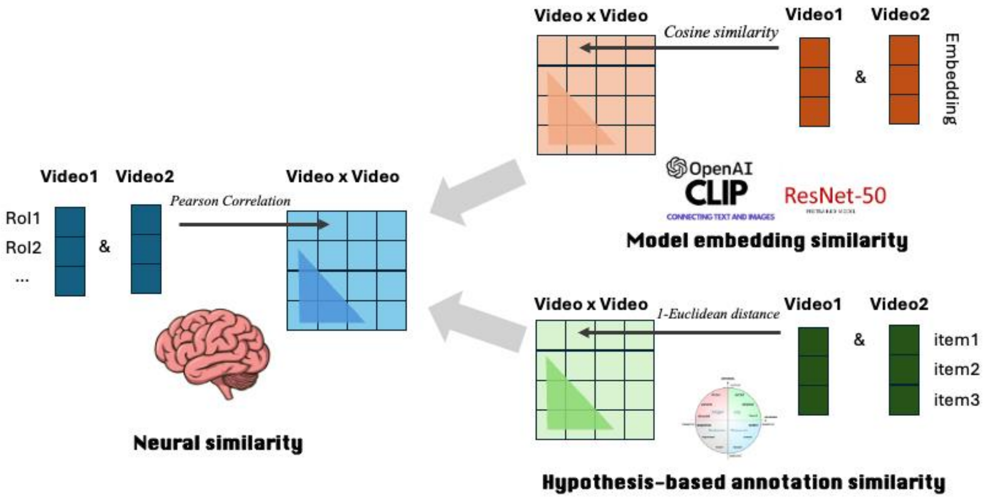

A. Creating a neural RDM
Stimulus (S)
Neural Response Pattern (R)
S1
S2
S3
R1
R2
R3
1 - corr(R1,R2)
Neural Representational Dissimiarity Matrix (nRDM)
Dissimilarity
B. Creating a behavioral RDM
Stimulus (S)
Behavioral Response Pattern (R)
Distance(R1,R2)
Behavioral Representational Dissimiarity Matrix (bRDM)
C. Creating a conceptual model RDM
Stimulus Feature
Social
Social
Non-social
Same Feature
Conceptual Model Representational Dissimiarity Matrix (mRDM)
D. Comparing RDMs
Similarity to nRDM
bRDM
mRDM1
mRDM2

For the neural RSM, we calculated the representational structure of each participant's brain activity during video viewing. This was done by extracting the voxel-level β-value sequences (activation levels, organized by the original paper) for each two videos in a given ROI and computing the pairwise Pearson correlation between these activation patterns. This resulted in a 244 × 244 similarity matrix for each participant, reflecting how similarly the ROI responded to each pair of videos. At the group level, we averaged across participants' RSMs to obtain a group neural RSM (Sartzetaki et al., 2024). To ensure statistical robustness, we applied Fisher's z-transformation (Silver & Dunlap, 1987) before averaging and converted the results back to correlation values.

To compare neural RSMs with computational models, we computed candidate RSMs of two kinds. The first kind was based on network embeddings, where we extracted high-dimensional representations for each video using pre-trained deep learning models. Since embedding spaces are typically non-Euclidean (Wulff & Mata, 2025), we computed cosine similarity between video embeddings to construct an embedding RSM of size 244 × 244. The second kind relied on hypothesis-driven annotations, where we extracted semantic annotations for each video using CLIP model. Since these annotation values were approximately normally distributed, we computed 1 - Euclidean distance between annotation vectors to measure similarity, yielding another 244 × 244 RSM.

Figure 5

Representational Similarity Analysis Roadmap



To assess the relationship between neural patterns and model-based representations, we computed Spearman correlations between the lower triangles of the neural RSM and each candidate RSM (Kriegeskorte, 2008; Popal et al., 2019). This was done for each participant individually as well as at the group-level averaged RSMs. To determine statistical significance, we performed permutation testing of 1000 times by shuffling similarity values within the RSM (Nili et al., 2014). This resulting the correlation ROI x Candidate matrices, for each subject or at the group-level.

To further quantify how different candidate RSMs explained neural similarity, we performed a regression analysis (Parkinson et al., 2017). First, we extracted the lower triangular part of each RSM, resulting in feature vectors of approximately 29,646 values per participant. These vectors included video-pairwise neural similarity, model-based similarity, and annotation-based similarity. We then modeled neural similarity as a function of the candidate RSM similarities using linear mixed-effects models, with random intercepts for each participant (not including group-level RSM) to account for individual variability (Stolier et al., 2020). To ensure comparability across different similarity metrics, we rank-transformed similarity values within each vector before running the regression, similar to the logic of Spearman correlation in RSA (Parkinson et al., 2017).

The regression analysis provided β coefficients quantifying the contribution of each candidate model to explaining neural representational structure. By comparing these β values, we assessed which candidate RSM best explained neural activity patterns while controlling for the effects of other candidates. Through this combination of RSA and regression modeling, we were able to systematically examine the alignment between neural representations, deep learning models, and hypothesis-driven annotations, offering insights into how different levels of computational representation relate to neural encoding of social interactions.

All statistical significances are obtained after the stringent Bonferroni multiple comparison correction (Bland & Altman, 1995).

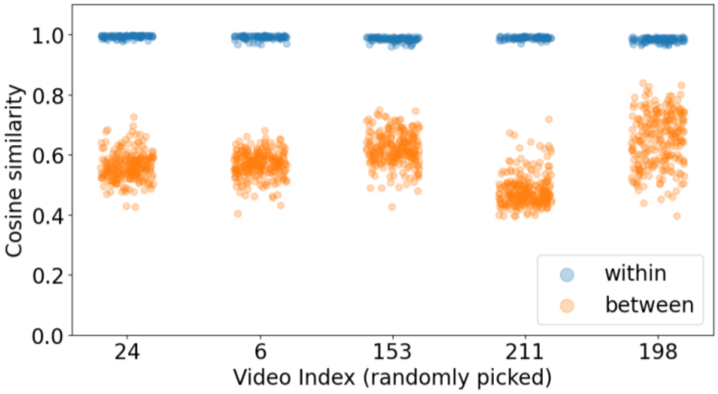
Results

Deep learning

We first evaluated the validity of aggregating frame embeddings into video embeddings. As shown in the Figure 6, the within-video similarity, measured as the cosine similarity between each frame and its video centroid, is consistently high (all above 0.95), while the between-video similarity, calculated as the cosine similarity between the centroids of different videos, is relatively lower and has larger variance across videos. This clear separation indicates that cross-frame aggregation should preserve video-level representation.

Figure 6

Verification: Cross-frame aggregation for obtaining video embeddings.



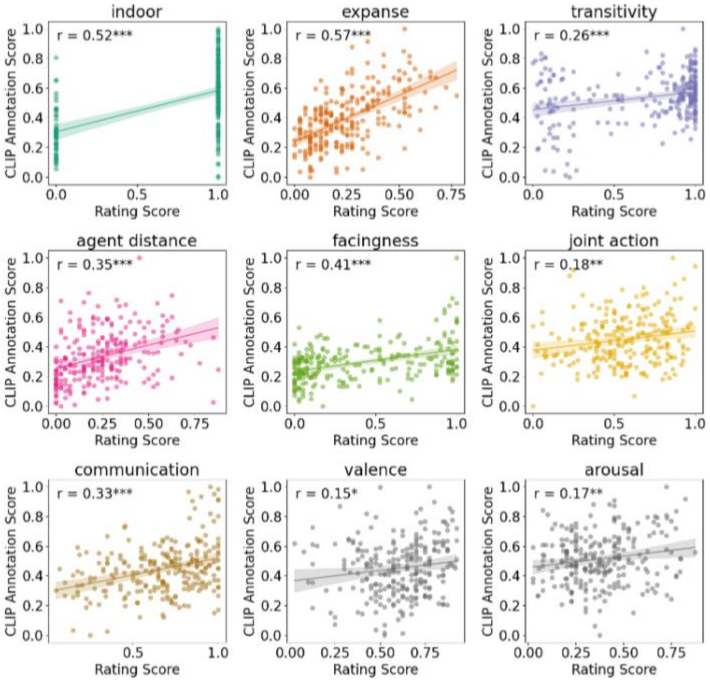
Note. This scatter plot (jittered by x-axis) illustrates the cosine similarity of within-video (blue) and between-video (orange) comparisons. The x-axis represents randomly selected video indices, while the y-axis shows cosine similarity values.

After obtaining the video CLIP embeddings, we examined their consistency with human ratings (Figure 7). The results indicate that CLIP embedding effectively captures meaningful social interaction attributes, showing strong correlations in dimensions such as expanse, indoor, and agent distance (p<0.001), while more abstract dimensions like joint action (p<0.01), valence and arousal (p < 0.05) exhibit weaker correlations, but all of them are statistically significant.

This further validates the reliability of the video embeddings and demonstrates CLIP's potential for capturing semantic information.

Figure 7

Correlation Between Human Ratings and CLIP Annotations Across Dimensions



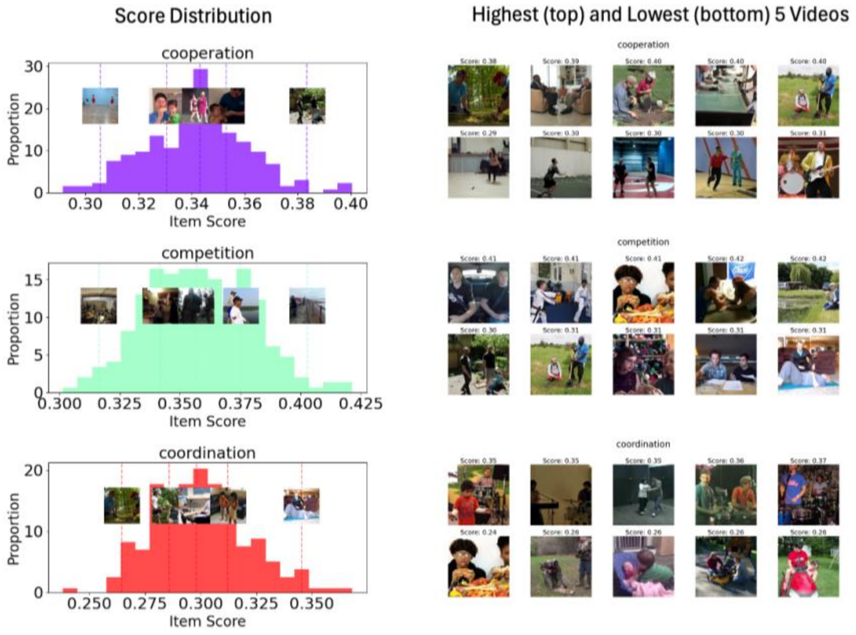
Note. This figure shows the correlation between human rating scores and CLIP-derived annotation scores across various dimensions (McMahon et al., 2023). The annotation scores were obtained by using description of each dimension in original paper to generate embeddings, which were then compared to video embeddings to generate similarity scores. Each subplot represents a different dimension, with scatter points showing individual video scores and trend lines indicating correlation strength (shadow ribbon as 95% confidence interval). Asterisks (* <0.05, ** <0.01, *** <0.001) denote levels of statistical significance.

Finally, we employ the framework proposed before to extract hypothesis-based annotation for videos. To have an intuitive sense of the dimensional annotation, Figure 8 is an example of annotations within the "game theory dynamics" dimension, with three items:

cooperation, competition, and coordination. The distribution of scores for each item, along with the highest- and lowest-scoring videos, demonstrates that the annotation process generally aligns with human intuition. The most striking point lies in the distinction between cooperation and coordination: High-scoring cooperation videos predominantly feature people working together towards a shared goal (e.g., planting trees), whereas coordination captures synchronized actions requiring precise temporal alignment (e.g., dancing, playing in a band). This distinction highlights the framework's ability to capture subtle semantic differences, reinforcing its potential for video annotation. More annotation examples are exhibited in the Appendix.

Figure 8

Example of "Game Theory Dynamics" Dimension Annotation



Note. This figure presents the annotation results of the "game theory dynamics" dimension, which includes three items: *cooperation, competition, and coordination*. The left column displays score distributions for each item, with example video frames overlaid at different percentiles along the distribution (2nd, 25th, 50th, 75th, 98th). The right column shows five highest-scoring videos (top row) and five lowest-scoring videos (bottom row) for each item, providing visual examples of how videos vary along these items.

Representational similarity analysis

In this section, we examine the relationship between different candidate RSMs and neural RSMs. Given the large number of videos, the number of video pairs is substantial, resulting in an RSM with a vast number of observations in its lower triangle. This high volume of data means that even minor effects can appear statistically significant (Ranganathan et al., 2015). For instance, at the group level, the neural-candidate correlations yielded 88.1% permutation-based significance, even after Bonferroni multiple comparison correction. Therefore, in interpreting and visualizing these results, we focus primarily on effect sizes—the magnitude of correlation and regression coefficients—rather than statistical significance.

We start by calculating the Spearman correlation between neural and candidate RSMs (Figure 6). For the network embedding results (right 3 columns), CLIP embedding similarity patterns correlate a broader range of ROIs compared to purely visual models (ResNet variants). However, while CLIP embeddings outperform vision-based embeddings in explaining neural patterns, they still lack interpretability—making it difficult to pinpoint the specific dimensions driving these neural correlations. This highlights the value of dimensional annotations, which provide a more hypothesis-based, fine-grained understanding of the underlying representations.

In contrast to the broad associations found in original CLIP embedding, the dimensional annotation RSMs reveal selective neural correlations, where specific dimensions are linked to distinct brain regions rather than broadly distributed.

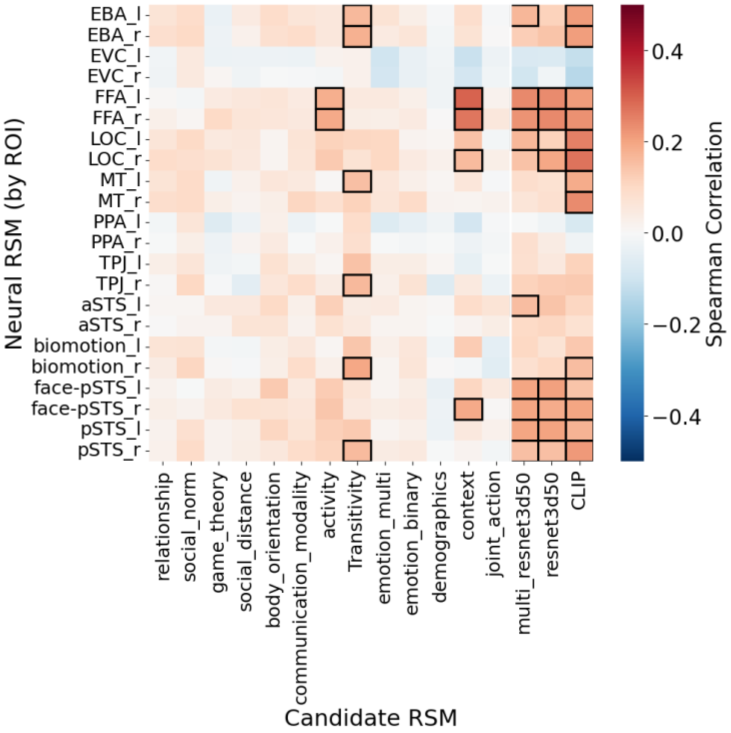
Transitivity (i.e., whether people interact with objects, others, or act independently, see Appendix, Figure S2a) shows above-threshold correlations (r>0.15) with both sides of EBA (extrastriate body area), left MT (middle temporal visual area), right biomotion-STS (superior temporal sulcus) and right pSTS (posterior superior temporal sulcus). Given their well-established roles in body perception and action understanding, these regions appear to be particularly sensitive to how an agent engages with its environment (Wang et al., 2015; Deen et al., 2015). Additionally, right TPJ (temporoparietal junction), a region often associated with theory of mind and perspective-taking (Saxe & Kanwisher, 2003), also shows notable correlation, suggesting a possible role in higher-order interpretation of social engagement.

Context (e.g., indoor vs. outdoor, see Appendix, Figure S2b) dimension shows remarkable correlations with visual scene-processing regions, particularly both sides of FFA (fusiform face area), right LOC (lateral occipital cortex) and right face-pSTS. Activity (e.g., dancing, cooking, see Appendix, Figure S2c) also shows moderate correlation with both sides of FFA. Interestingly, FFA, despite its primary role in face recognition (Kanwisher & Yovel, 2006), demonstrates strong correlations with scene-related dimensions. In contrast, PPA (parahippocampal place area), which is conventionally associated with scene processing (Julian et al., 2012), does not exhibit strong correlations with context.

A possible explanation for the FFA's involvement is that different activities and contexts in these videos introduce substantial variations in facial visibility and orientation: *"Scene features in our dataset are also heavily confounded with the size and visibility of faces in our videos"* (McMahon et al., 2023). For example, in fishing scenes (Appendix, Figure 2c), individuals are often seen in profile, facing away from the camera, and appearing smaller in the frame, which could drive scene-related variability in face perception. In contrast, the lack of significant PPA correlations across both network embeddings and dimensional annotations suggests that its weak response is unlikely to be an artifact of the annotation framework. Instead, it may reflect a more fundamental limitation in the ability of deep models to capture PPA processing patterns.

Figure 9

Correlation between Candidate and Group-level Neural RSMs



Note. This heatmap presents the Spearman correlation between ROI RSMs(y-axis) and candidate RSMs (x-axis). Candidate RSMs were derived from hypothesis-based annotations (the left section) network embeddings (the right section). Warmer colors indicate positive correlations, while cooler colors indicate negative correlations. Since most (88.1%) of correlations are significant through permutation test of 1000 times, we did not highlight based on significance, instead based on the absolute value of the correlation coefficient. Correlations with magnitude larger than 0.15 are highlighted with black-bordered squares.

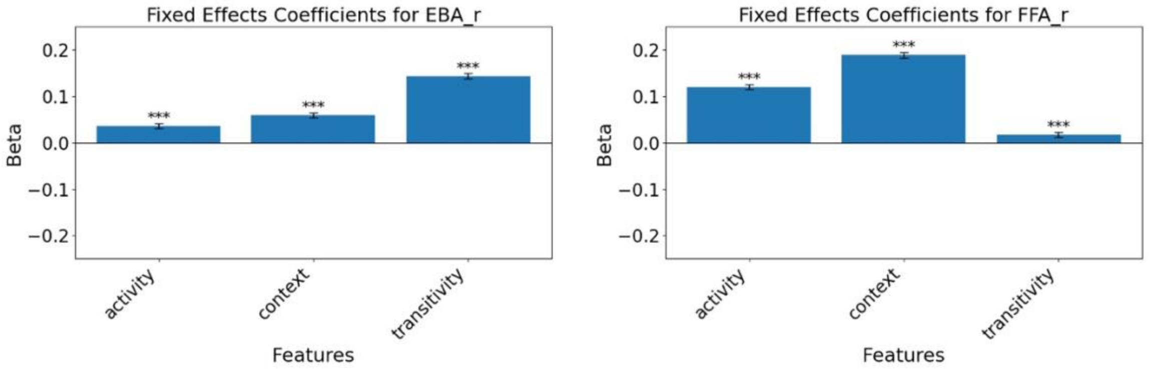
To further investigate the selectivity of neural responses to different annotation dimensions, we employed a linear mixed-effects model (Stolier et al., 2020) to predict neural activity based on selected annotation features, with subject variability included as a random effect. This analysis focuses on dimensions that showed strong associations with specific brain regions in the previous results.

The findings reveal distinct selectivity across brain regions. For example, in the right EBA (extrastriate body area), transitivity emerges as the primary predictor, suggesting heightened sensitivity to whether actions involve objects or social interactions — consistent with EBA's role in processing human body movements and action perception (Deen et al., 2015).

In the right FFA, context is the most influential feature, followed by activity, despite FFA's well-established role in face processing. This likely reflects the substantial variability in facial visibility and orientation across different activities and environments, as discussed earlier.

Figure 10

Linear Mixed Model Coefficients for Two Example ROIs



Note. These bar plots display the fixed effects coefficients from the linear mixed model predicting neural responses in two selected brain regions: right EBA (left panel) and right FFA (right panel). The x-axis represents annotation dimensions that exceeded a predefined correlation threshold of 0.15 in the correlation analysis before. The linear mixed-effects model was fitted using the python *statsmodels* package (Seabold & Perktold, 2010), with subject as the random intercept. Error bars indicate bootstrap 95% confidence intervals, and asterisks (* <0.05, ** <0.01, *** <0.001) denote levels of statistical significance.

However, when we explicitly included face-related annotations such as *face orientation* (items: face looking directly at the camera, face looking to the left, face looking to the right, face looking down, face looking up), *face visibility* (items: face fully visible, face partially visible, face not visible), and *face distance* (items: face in close-up, face at medium distance, face far away), their correlations with FFA activity remained weak. Except for participant 1's right FFA and face orientation (r = 0.102), none of other features showed a correlation exceeding 0.1. Future detailed investigation is expected.

Overall, the RSA findings demonstrate that neural responses are selectively tuned to specific social and contextual dimensions rather than uniformly reflecting all perceptual attributes. This selectivity highlights the utility of hypothesis-driven annotations, which provide a more interpretable framework for understanding neural representations compared to the generic insights offered by network embeddings.

Discussion

Driven by the question of which aspects of social interactions are represented in the brain, this study introduces a scalable, hypothesis-driven framework that automates video annotation, improving efficiency while maintaining interpretability. By integrating CLIP model embeddings with structured annotation design, this approach enables systematic exploration of the hypothesis space. Employing Representational Similarity Analysis, this study examines how neural responses align with different candidate representations, including CLIP and ResNet variants embeddings, as well as hypothesis-based annotations. The results reveal selective neural tuning, with distinct brain regions responding preferentially to specific features.

Several directions can further refine this framework. First of all, although the alignment between CLIP annotations and neural activity as well as human ratings has been observed, a more comprehensive comparison with human judgments is needed to fully assess validity. Second, ROI definition in this study followed the original paper, which primarily focused on visual processing regions, potentially overlooking higher-order cognitive areas involved in social reasoning. Searchlight (Kriegeskorte et al., 2006) analysis could provide a more comprehensive neural mapping. Third, while this study used predefined annotation dimensions, the framework can be extended to automatically identify relevant dimensions using data-driven methods such as active learning within the hypothesis space (Awad et al., 2018; Huang, 2025; Peterson et al., 2021), which leverages previously explored hypotheses' results to guide the sampling of new dimensions (ongoing project, see Appendix). Finally, this framework can be applied beyond social interaction, providing interpretable insights into broad perception and evaluation of complex stimuli, for example, which features of an image makes it memorable, or what factors drive the popularity of a YouTube video.

Appendix

Automated hypothesis discovery and testing to investigate Neural representation

This framework provides an automated workflow to discover and test hypotheses about which dimensions of a complex stimuli (e.g., social interaction) correspond to their neural representations, shedding light on how human process how humans perceive and encode such information. It consists of two processes:

Constructing the design space

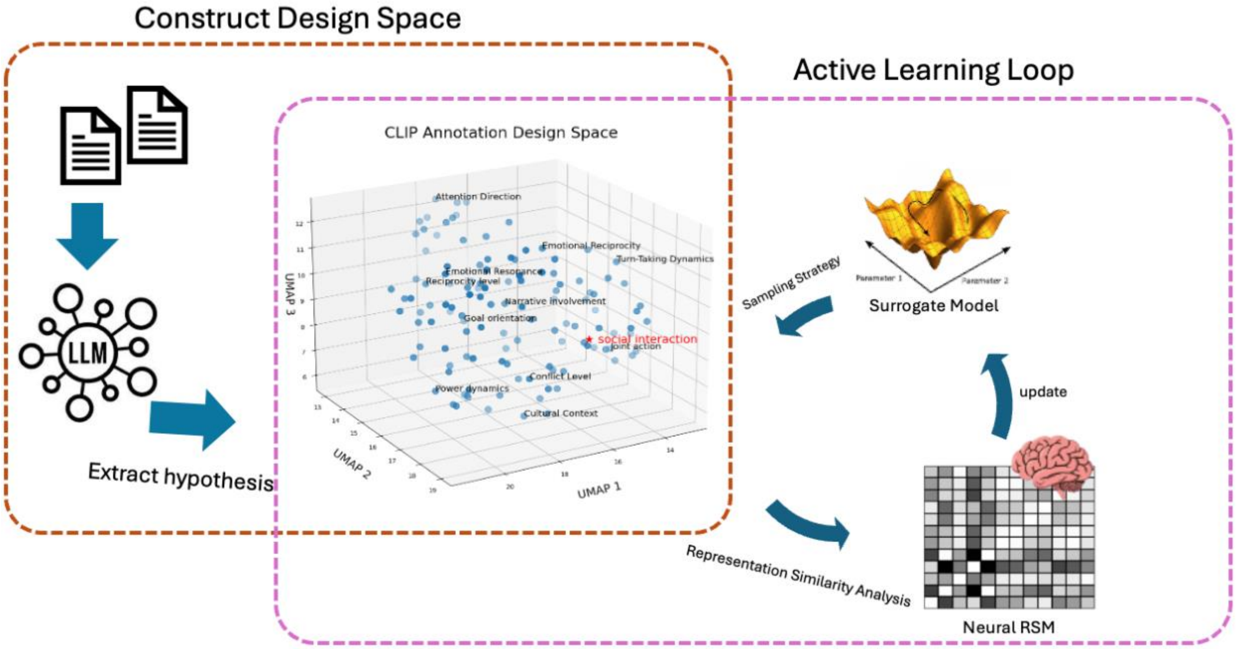
This process begins with relevant seed literature, which serves as the foundation for hypothesis generation. These sources are used as prompt examples for a Large Language Model agent, which undergoes multiple rounds of actor-critic refinement (Zhang et al., 2023) to iteratively expand and refine a hypothesis set. The final output is a structured list of potential dimensions, which are then embedded into a semantic space, forming the design space for further exploration.

Exploring the design space efficiently through an active learning loop

Initially, a subset of hypotheses is randomly sampled from the space, and Representational Similarity Analysis (RSA) is used to evaluate their correlation with neural activity patterns. Based on these results, a surrogate model is updated to estimate how different areas of the embedding space might relate to neural representations. A sampling strategy then selects the next set of hypotheses to test, balancing exploration (searching under-explored regions) and exploitation (prioritizing areas with strong neural correlations). This process repeats iteratively until the framework either achieves a comprehensive understanding of the entire space or identifies certain dimensions that exhibit strong and meaningful neural correlations.

Figure S1

Automatic Discovery and Testing of Hypotheses regarding Neural Representations

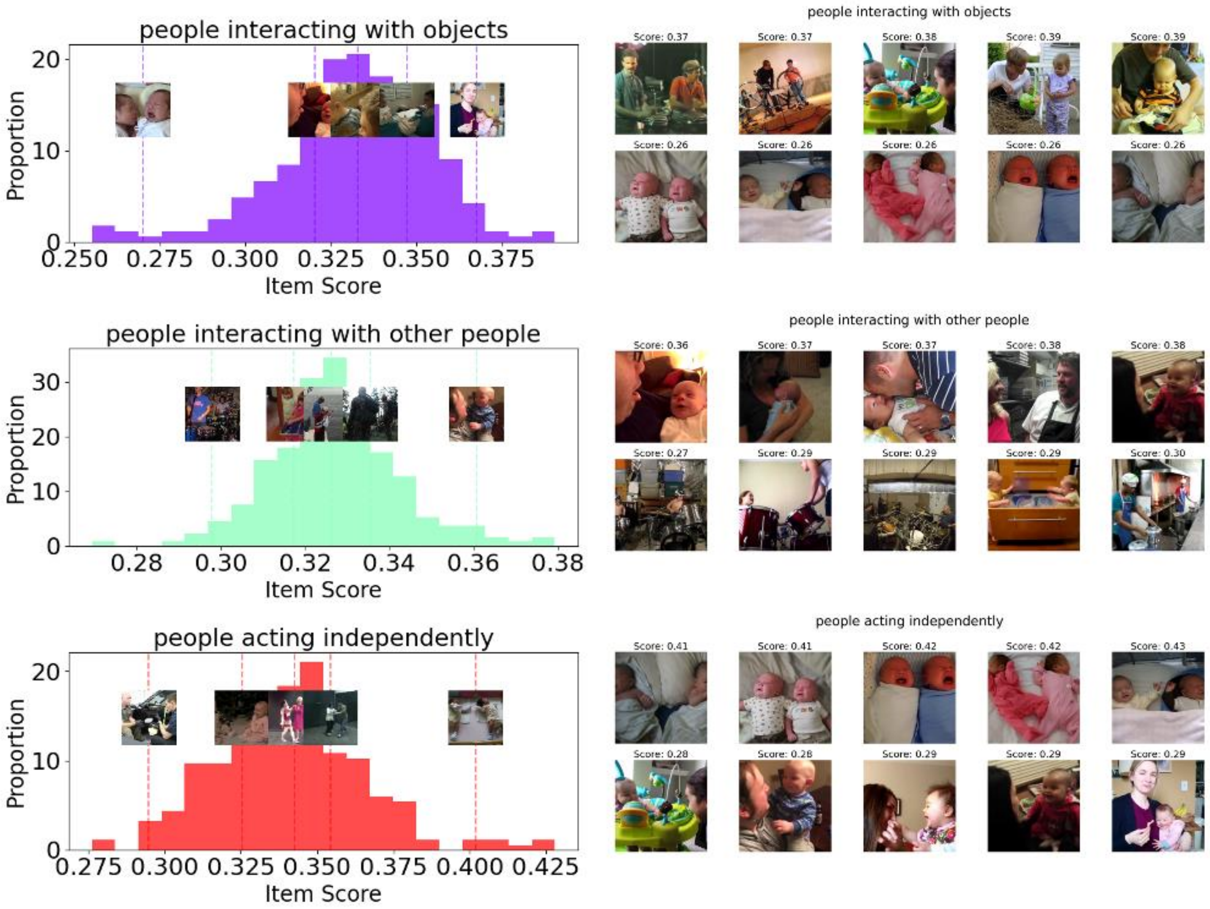


By integrating data-driven hypothesis generation with an iterative search process, this framework enables the automated specification of neural representation dimensions. Instead of relying on predefined annotations, it dynamically refines hypotheses, providing a structured yet flexible way to investigate how our brain represents complex stimuli.

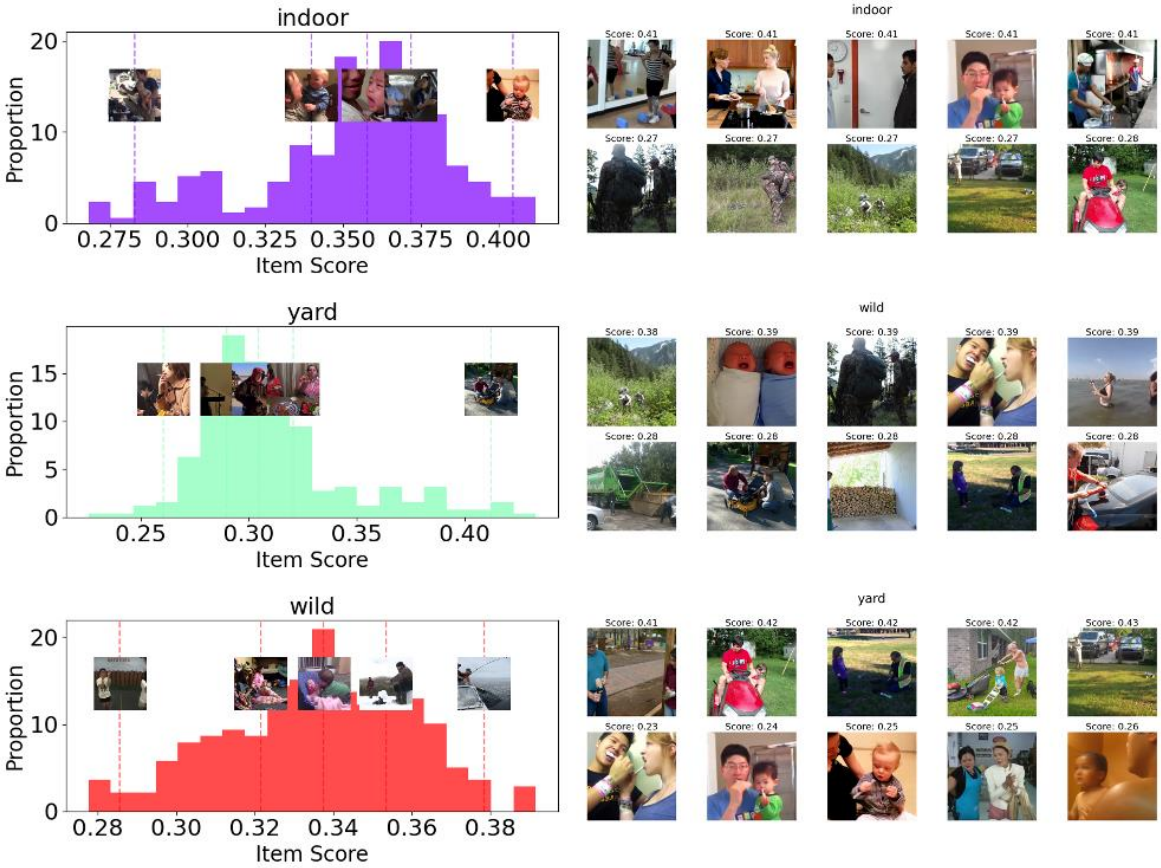
More examples of CLIP annotations

Figure S2

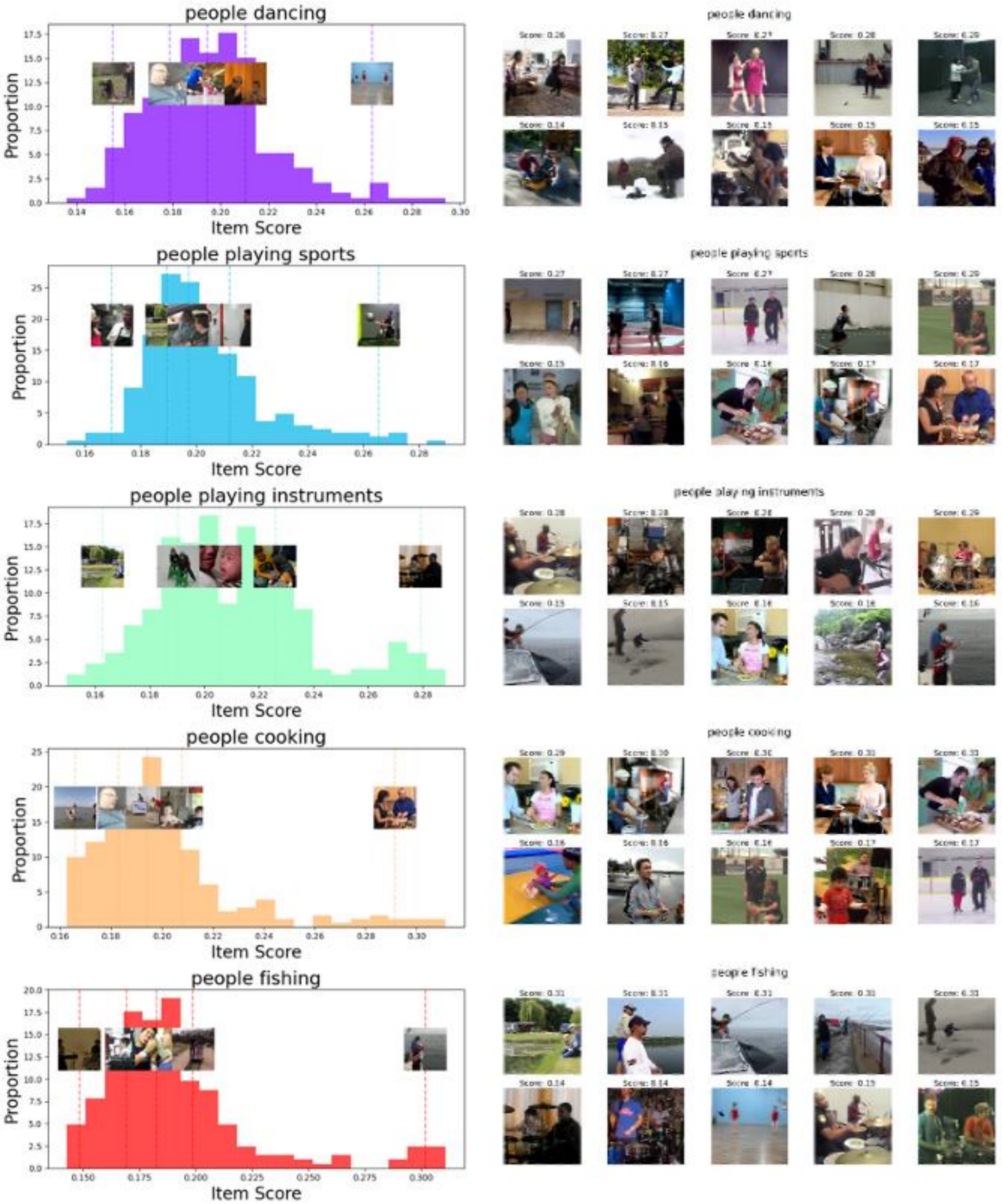
a. *Example of "Transitivity" Dimension Annotation*



b. Example of "Context" Dimension Annotation



c. Example of "Activity" Dimension Annotation



References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*. https://doi.org/10.1017/s0140525x22002874

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, *37*(1), 90101.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *Bmj*, *310*(6973), 170170.

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Arxiv preprint arXiv:1705.07750*.

Carvalho, W., & Lampinen, A. (2025). Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *Arxiv preprint arXiv:2502.20349*.

Cheng, X., Popal, H., Wang, H., Hu, R., Zang, Y., Zhang, M., Thornton, M. A., Cai, H., Bi, Y., Reilly, J., & others. (2023). *The Conceptual Structure of Human Relationships Across Modern and Historical Cultures*.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *10*(4-5), 171178.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 6372.

Hadley, L. V., Naylor, G., & Hamilton, A. F. d. C. (2022). A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, *1*(1), 42–54. https://doi.org/10.1038/s44159-021-00008-w

Huang, L. (2025). Comprehensive exploration of visual working memory mechanisms using large-scale behavioral experiment. *Nature Communications*, *16*(1). https://doi.org/10.1038/s41467-025-56700-5

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Julian, J., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76–85. https://doi.org/10.1137/0701007

Lee Masson, H., Chang, L., & Isik, L. (2024). Multidimensional neural representations of social features during movie viewing. *Social Cognitive and Affective Neuroscience*, *19*(1), nsae030nsae030.

McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, *33*(23), 5035–5047e8. https://doi.org/10.1016/j.cub.2023.10.015

McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*, *27*(12), 1165–1179. https://doi.org/10.1016/j.tics.2023.09.001

Monfort, M., Vondrick, C., Oliva, A., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., & Gutfreund, D. (2020). Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502–508. https://doi.org/10.1109/tpami.2019.2901464

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5). https://doi.org/10.1038/s41562-017-0072

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Popal, H., Wang, Y., & Olson, I. R. (2019). A guide to representational similarity analysis for social neuroscience. *Social cognitive and affective neuroscience*, *14*(11), 12431253.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320341.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Ranganathan, P., Pramesh, C., & Buyse, M. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in Clinical Research*, *6*(3), 169. https://doi.org/10.4103/2229-3485.159943

Sartzetaki, C., Roig, G., Snoek, C. G., & Groen, I. I. (2024). One Hundred Neural Networks and Brains Watching Videos: Lessons from Alignment. *Biorxiv*, 202412.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *NeuroImage, 19*(4), 1835–1842. https://doi.org/10.1016/s1053-8119(03)00230-1

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *SciPy, 7*(1), 92-96.

Sievers, B., & Thornton, M. A. (2024). Deep social neuroscience: the promise and peril of using artificial neural networks to study the social brain. *Social cognitive and affective neuroscience*, *19*(1), nsae014. https://doi.org/10.1093/scan/nsae014

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's z transformation be used? *Journal of applied psychology*, *72*(1), 146146.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, *4*(4), 361–371. https://doi.org/10.1038/s41562-019-0800-6

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-02089-y

Zhang, B., Mao, H., Ruan, J., Wen, Y., Li, Y., Zhang, S., Xu, Z., Li, D., Li, Z., Zhao, R., & others. (2023). Controlling large language model-based agents for large-scale decision-making: An actor-critic approach. *Arxiv preprint arXiv:2311.13884.*