



Action features dominate cortical representation during natural vision

Heejung Jung^{a,1}, Hunar Jain^b, Xiaochun Han^a, Ma Feilong^a, Jane Han^a, Deepanshi Shokeen^a, Cara Van Uden^a, Isabella Hanson^a, Andrew C. Connolly^a, Tor D. Wager^a, James V. Haxby^a, and Samuel A. Nastase^{c,1}

Affiliations are included on p. 10.

Edited by Martin S. Banks, University of California, Berkeley, CA; received July 16, 2026; accepted September 23, 2026

Cortical resources are allocated into systems specialized for processing ecologically relevant features of the world. These features are typically studied in isolation using controlled experimental stimuli, making it difficult to assess the relative importance of different kinds of features that tend to overlap during natural vision. In the current study, we evaluated the relative contributions of action, agent, and scene features in predicting cortical activity while participants viewed a 1-h nature documentary. We tested four sets of model features: semantic vectors capturing the observed action, agent, and scene features derived from an annotation of the stimulus, as well as low-level visual motion-energy features. We used banded ridge regression to fit vertex-wise encoding models from the combination of all four sets of features. While each set of features predicted neural activity in the expected areas, the action features predicted more widespread activity. A variance partitioning analysis revealed that the action features captured the most unique variance in neural activity across a much larger extent of cortex than the agent or scene features. Our findings suggest that cortical activity during dynamic, natural vision is dominated by features for understanding the actions of others.

action understanding | cortical organization | fMRI | naturalistic stimuli | voxelwise encoding models

Cortical resources are differently allocated toward processing different features of the world. This division of labor coincides with the behavioral relevance of certain kinds of stimuli. For example, in highly visual species like humans and other primates, a relatively large proportion of cortex is devoted to visual processing, and visual cortex is highly subdivided, relative to other mammals (1–3). Cortical magnification as a function of behavioral importance and complexity is a well-established principle in sensory and motor cortices: foveal vision recruits much more cortex than the periphery for higher-acuity processing (4–8); effectors requiring fine motor control like the hands and mouth occupy disproportionately large portions of sensorimotor cortex (9–13). Cortical magnification may also extend into higher-order perceptual and association cortices, reflecting the relative complexity and ecological importance of different high-level features.

In this vein, functional neuroimaging has revealed a mosaic of functionally specialized cortical areas for processing faces (14, 15), scenes (16, 17), and other salient features of our visual world (18), culminating in a ventral visual pathway for object perception and categorization (19, 20). Results of this kind have informed a number of theories on the large-scale structure of semantic knowledge, particularly object categories, in the human brain (21–26). In these experiments, however, neural responses to static images of faces or scenes are typically contrasted with carefully selected control images matched on certain parameters—making it difficult to assess how different kinds of features compare in terms of their overall cortical footprint.

A largely separate line of work has focused on the cortical systems supporting the perception of dynamic visual features, such as biological motion (27–31) and observed actions (31–35). These experiments typically employ highly controlled stimuli as well: for example, point-light displays of a figure moving naturally versus nonnatural motion (e.g., refs. 36 and 37), moving geometric shapes depicting social interactions (e.g., refs. 31 and 32), or brief videos depicting isolated hands performing simple actions like opening versus closing a container (e.g., refs. 38 and 39). Experiments of this kind have culminated in a lateral visual pathway for dynamic social perception (40, 41) and a frontoparietal action observation network (33, 34, 42). Ventral visual areas, however, are also consistently activated by these dynamic stimuli depicting agentic action (e.g., refs. 27, 31, and 32, suggesting that the putative “object vision pathway” also plays an important role in dynamic social perception and action observation (43, 44). In studies where different

Significance

Human cortex is functionally organized to prioritize the most important features of our external world. In perceiving the natural world, however, these features are often overlapping and intermixed. How can we fairly assess which features of the world are most important to the brain? Using a naturalistic neuroimaging paradigm in which participants freely viewed a nature documentary, we modeled the relative contributions of observed action, agent, scene, and low-level visual motion features in predicting cortical activity. We found that observed actions occupy a disproportionately large amount of cortical real estate—capturing considerably more unique variance than agent and scene features. This cortical magnification of action features suggests that the brain prioritizes action understanding during natural vision.

Author contributions: H. Jung, J.V.H., and S.A.N. designed research; S.A.N., A.C.C., and I.H. performed research; H. Jung, X.H., M.F., J.H., D.S., C.V.U., and T.D.W. contributed new reagents/analytic tools; H. Jung, X.H., and S.A.N. analyzed the data; S.A.N., J.V.H., and T.D.W. supervised the project; and H. Jung, S.A.N., and J.V.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2026 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

¹To whom correspondence may be addressed. Email: heejung.jung@dartmouth.edu or snastase@usc.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2625018123/-DCSupplemental>.

Published October 23, 2026.

classes of stimuli are compared (e.g., refs. 36, 45, and 46, many regions exhibit mixed responses to different classes of stimuli, suggesting that functional specialization may be more complex than some experiments would suggest.

These cortical pathways for static object perception and dynamic action perception emerged separately, in isolated experiments. Taking these findings into consideration, we might surmise that, while cortex has functionally specialized subdomains for certain features of the world, these features are roughly commensurate—each of these features is similarly important to cortex. This conclusion may not be warranted, however, for several reasons. While targeted experimental manipulations are useful for disentangling different features, this piecemeal approach limits our ability to quantify how different features are processed in cortex under naturalistic conditions, when the field is cluttered, dynamic, and embedded in a temporal and semantic context. In natural vision, features are not compartmentalized—they are often superimposed and may interact in a number of ways. For instance, different agents may perform semantically similar actions in kinematically different ways (a monkey eating nuts, a seal eating penguins); different actions may correlate with certain background features (e.g., a cheetah sprinting across the savanna, a dolphin swimming past a reef). While these might be considered confounds in experimental design, the brain may both capitalize on these statistical associations (47–51) and learn invariances across them: despite the fact that a caterpillar, a hummingbird, and a chimpanzee eat in quite different ways, the brain encodes semantic features of the eat action that generalize across different agents (52). The use of naturalistic stimuli allows us to more holistically evaluate the relative contribution of different features that are typically studied in isolation (43, 53–55).

Recent work has begun using movies and spoken narratives to explore how different features are encoded in neural activity under more ecological conditions. Movies have been shown to drive reliable activity across much of cortex (56–65) and are thought to sample features of the world (and their intercorrelations) in a more “representative” (66) way than typical experimental stimuli (55, 67). For example, Huth and colleagues (2012; 68) used an annotation of a natural movie stimuli to demonstrate that different semantic features are intermixed and represented broadly across cortex. Unexpected results have been found in macaques, where dynamic, socially relevant features of naturalistic stimuli were found to drive activity in areas typically associated with face processing in experiments using static image stimuli (69, 70). Given the complexity inherent to naturalistic stimuli, how can we assess what features best capture cortical activity? Researchers have developed a modeling framework where the stimulus is decomposed into different feature spaces based on explicit models, which are then tested against the neural activity using joint and nested regression models (71–73). For example, de Heer and colleagues (2017; 72) compared three different sets of features for speech comprehension in spoken narratives: spectral, articulatory, and semantic features. Using a variance partitioning analysis, they found that semantic features capture much more variance in neural activity across a large portion of cortex, whereas spectral and articulatory features of speech were encoded in relatively focal auditory regions. This framework allows us to quantitatively link complex neural activity driven by naturalistic stimuli to interpretable features, and compare the relative contributions of different types of features.

In the current study, we asked whether certain higher-level features of the visual world are “magnified” across cortex relative

to other seemingly commensurate features. We tested several different kinds of ecologically relevant visual features to determine whether any particular set of features recruits a disproportionately large amount of cortical resources. We measured brain activity using fMRI while subjects viewed a 1-h movie—the “Life” nature documentary narrated by David Attenborough—depicting different animals behaving in their natural environments. Based on predominant trends in cortical organization emerging from the literature, we tested three different sets of features capturing the semantic content of the actions (e.g., swimming, running), agents (e.g., dolphin, cheetah), and scenes (e.g., ocean, savanna) visually depicted over the course of the film. We also include a motion-energy model to capture the low-level visual motion of the stimulus and better isolate higher-level semantic features. To compare these features, we estimated vertex-wise encoding models jointly across all four feature spaces, then assessed the relative contribution of each feature space. We found that the action features predominate in dorsal regions, with strong encoding performance extending from inferior parietal regions to premotor cortex. The agent features yielded the strongest encoding performance in ventrolateral temporal cortex, extending from lateral fusiform to the posterior lateral occipitotemporal cortex. The scene features were strongest in ventromedial occipital and temporal cortex. A variance partitioning analysis, however, revealed that the action features uniquely predicted activity across a much greater proportion of cortex than the competing models.

Results

We tested the relative contributions of different semantic feature spaces to cortical activity while participants viewed a nature documentary (Fig. 1). Semantic features for actions, agents, and scenes were constructed from a detailed behavioral annotation of the visual content of the film (68). For each word, we obtained a 300-dimensional word embedding from the GloVe model (74). In this model, semantic relationships are encoded in the geometric relationships between vectors, such that more similar words are located nearer to each other in the embedding space, based on their co-occurrence in large corpora of text. We also tested a biologically inspired motion-energy model capturing dynamic, low-level visual content (75–77). To assess these different models, we constructed vertex-wise encoding models: we fit a linear model mapping from the semantic vectors to the response time series at each vertex (78). To fairly assess the contribution of each model, we first used principal component analysis (PCA) to reduce each feature space to a matching dimensionality (40 dimensions, explaining over 95% of the variance in each of the three feature sets of interest; *SI Appendix, Fig. S1*). Second, we combined all four feature spaces and estimated joint encoding models using banded ridge regression (79). We evaluated the relative contributions of these models using four-fold, leave-one-run-out cross-validation combined with leave-one-subject-out cross-validation; that is, models were evaluated in a left-out segment of the stimulus in a left-out subject. To better align cortical-functional topographies across individuals, we estimated individual-specific transformations using whole-brain surface-based searchlight hyperalignment (80) from the three training runs, then applied these transformations to the left-out fourth run. We quantified encoding performance as the correlation between the model-predicted time series and the actual fMRI time series for the test run in the test subject.

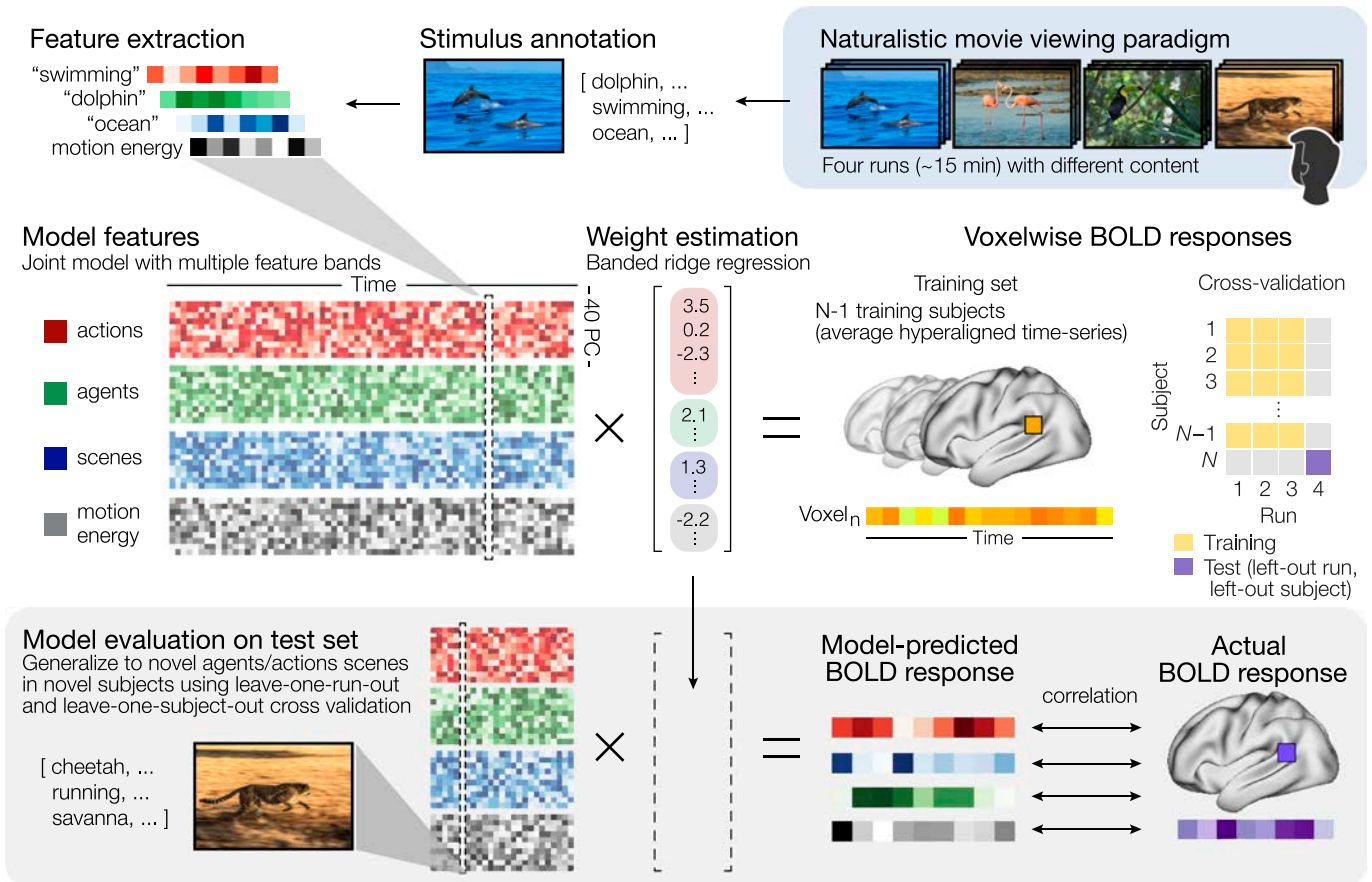


Fig. 1. Encoding multiple feature spaces using banded ridge regression. Participants watched a movie stimulus—a nature documentary featuring a variety of behaving animals in their natural environments—during fMRI acquisition. The stimulus was split into four separate 15-min runs; each run was unique and contained multiple scenes depicting different animals in different environments. The stimulus was annotated at roughly 2-TR (5-s) intervals to describe in words what was visually depicted in the film. The annotation was then split into words describing the agents (e.g., dolphin), actions (e.g., swimming), and scene (e.g., ocean). For each word, we extracted GloVe embeddings capturing the semantic content of the words. We also extracted dynamic, low-level visual features of the stimulus using a motion-energy model. The embeddings for agent, action, and scene words, as well as the frame-by-frame motion energy features, were downsampled to the TR and concatenated into a joint encoding model. We used banded ridge regression to estimate a linear mapping from the joint model to the BOLD response time series at each vertex. Vertex-wise encoding models were trained on the average time series from $N - 1$ participants aligned into a shared space using searchlight hyperalignment (leave-one-subject-out cross-validation) for three of the four stimulus segments (leave-one-run-out cross-validation). We used the trained model to generate predicted BOLD response time series for the test segment. Model-based predictions were generated jointly or separately for each feature space by discarding (zeroing out) weights for other feature spaces. Encoding model performance was evaluated by computing the correlation between the model-predicted BOLD time series and the actual BOLD time series for the left-out test run in the left-out test subject (also aligned to the shared space using hyperalignment transformations based on the training runs).

Vertex-Wise Encoding Performance for Actions, Agents, and Scene Features. To quantify the relative contributions of each feature space, we fit joint vertex-wise encoding models, then generated four sets of predictions based on the subset of weights corresponding to each feature space. We computed the correlation between the predictions for each feature space and the actual vertex time series separately (see *SI Appendix*, Fig. S2 for the full model performance). The joint model captures a subset of reliable variance in neural activity across subjects as estimated using an intersubject correlation (ISC) analysis (*SI Appendix*, Fig. S3). All four feature spaces yielded significant predictions (t test, FDR $P < 0.05$) with different cortical topographies (Fig. 2). The action features yielded significant predictions throughout the action observation network extending from lateral occipitotemporal cortex (LOC), anterior intraparietal (AIP) and postcentral cortex, and premotor cortex (Fig. 2A). Predictions based on the agent features were strongest in LOC, posterior superior temporal cortex, as well as ventral temporal (VT) cortex (Fig. 2B). The scene features yielded significant, but weaker, predictions in LOC, posterior parietal cortex (PPC), retrosplenial cortex, and ventral temporal cortex extending medially into parahippocampal

cortex (Fig. 2C). Finally, the motion-energy features yielded significant performance primarily in early visual areas, including primary visual cortex (V1) and posterior VT (Fig. 2D). In general, these cortical topographies are consistent with the functional organization reported by studies using experimental manipulations to target particular specific features in isolation. See *SI Appendix*, Fig. S4 for direct statistical comparisons between models.

Unique Variance Explained by Each Set of Features. In the previous analyses, we fit a joint model with four feature bands and then generated predictions from the weights corresponding to each feature space. While this approach allows us to visualize relative contributions of different feature spaces (73), the different feature bands may nonetheless predict overlapping portions of variance in neural activity—in a naturalistic context, different features may be highly collinear. That is, the prior analyses do not tell us how much variance in neural activity is uniquely predicted by a given feature space. To quantify the unique variance predicted by each feature space, we performed a variance partitioning analysis (71, 72). We performed a nested

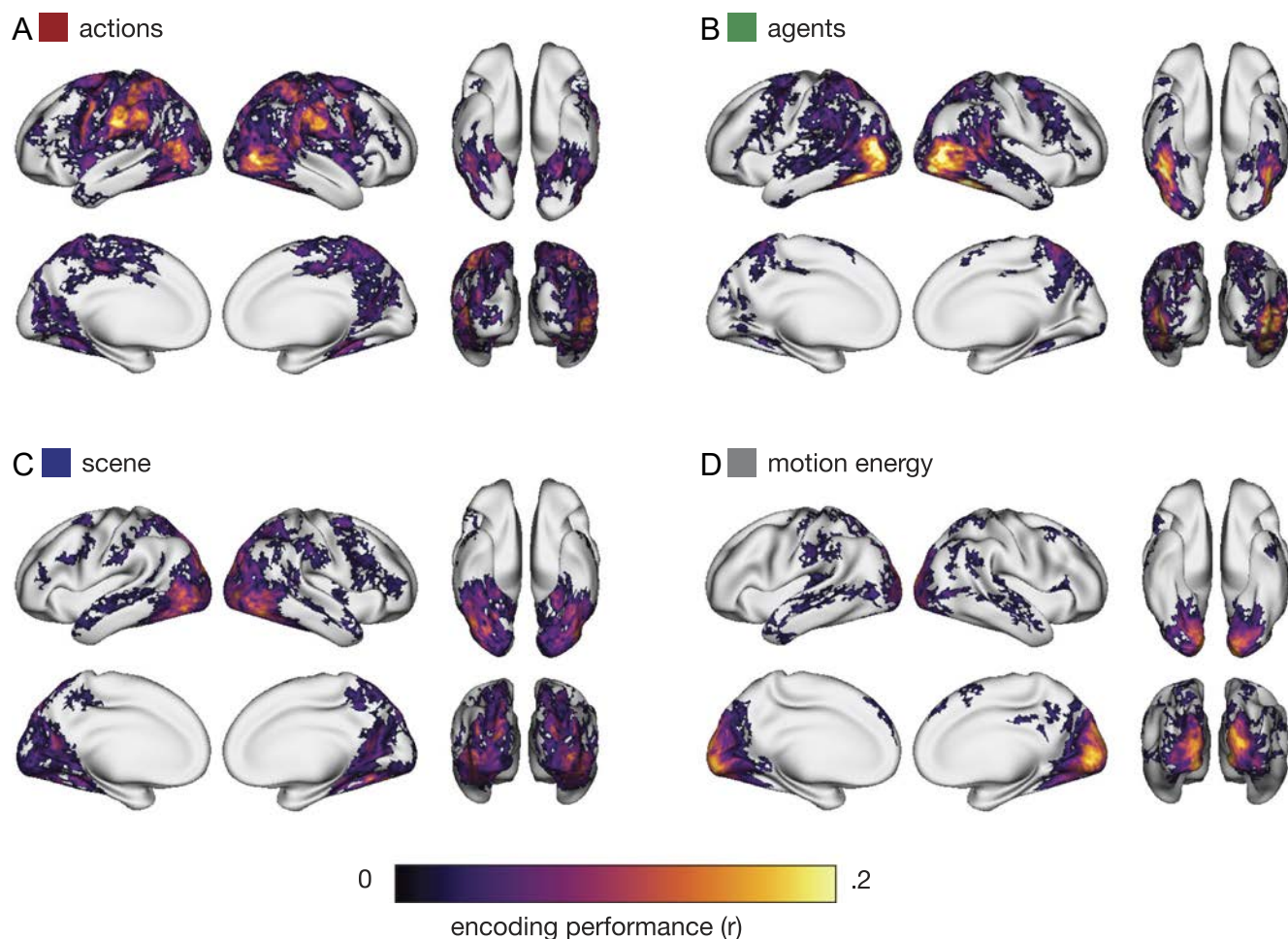


Fig. 2. Encoding performance for each feature space. Whole-brain vertex-wise encoding performance for four feature spaces: (A) actions, (B) agents, (C) scene, and (D) motion energy. Joint encoding models were estimated combining all four feature spaces using banded ridge regression, to allow each feature space to compete for variance. Predictions were then generated from the learned weights for each feature space separately. Encoding performance is quantified as the correlation between model-predicted and actual neural activity in a left-out run in a left-out subject. Correlations are averaged across test runs and test subjects, and thresholded for statistical significance (t test across subjects, FDR $P < 0.05$).

regression analysis in which we constructed a joint encoding model combining all four feature spaces, then fit four nested models, each excluding one feature space of interest. We quantify the unique variance explained by a feature space as the difference in encoding performance between the nested model where that feature space is excluded and the joint model where that feature space is included. This analysis effectively controls for collinearity between different sets of features; if a given feature set is fully collinear with other feature sets, it will not capture any unique variance in this analysis. We first quantified the unique variance explained by each of the four feature spaces and statistically evaluated these maps across subjects (Fig. 4).

We found that the action features uniquely predict neural activity across a large swath of cortex (Fig. 4A), including LOC, AIP, and vPM, as well as anterior VT, excluding the fusiform gyrus (bootstrap hypothesis test, FDR $P < 0.05$). The agent features uniquely predicted a much more restricted, punctate set of vertices localized to posterior LOC, STS, and fusiform gyrus (Fig. 4B). The scene features uniquely predicted neural activity in posterior VT and medial visual areas like retrosplenial cortex (Fig. 4C). The motion-energy features captured unique variance in early visual areas (Fig. 4D).

Last, we sought to quantify how much of the cortical sheet was uniquely predicted by different partitions of variance across

all four feature spaces. To do this, we first constructed encoding models for every combination of the four feature spaces: each individual feature space, each pair, each triplet, and the full model comprising all four feature spaces (15 combinations in total). We averaged encoding performance across subjects and runs for each combination of models. We then quantified the variance explained uniquely by each feature space, and each intersection of two or more feature spaces. This resulted in 15 values corresponding to each partition of a four-way Venn diagram. Finally, we determined which of these partitions explained the most variance in neural activity at each vertex (SI Appendix, Table S1). Although we included the models containing motion-energy features in the variance partitioning analysis, we focus on the three semantic feature spaces (i.e., actions, agents, scene).

First, we found that each model uniquely predicted some set of vertices better than all other models (and combinations of models), suggesting that the models themselves are comparable in quality to one another. For example, agent features captured more unique variance than all other combinations in 475 vertices (Fig. 5). Critically, however, we found that action features uniquely predicted neural activity better than all other combinations of features across an order of magnitude more vertices than agent and scene features: 4,984 vertices for actions versus 475 vertices for agents and 495 vertices for scene features (Fig. 5).

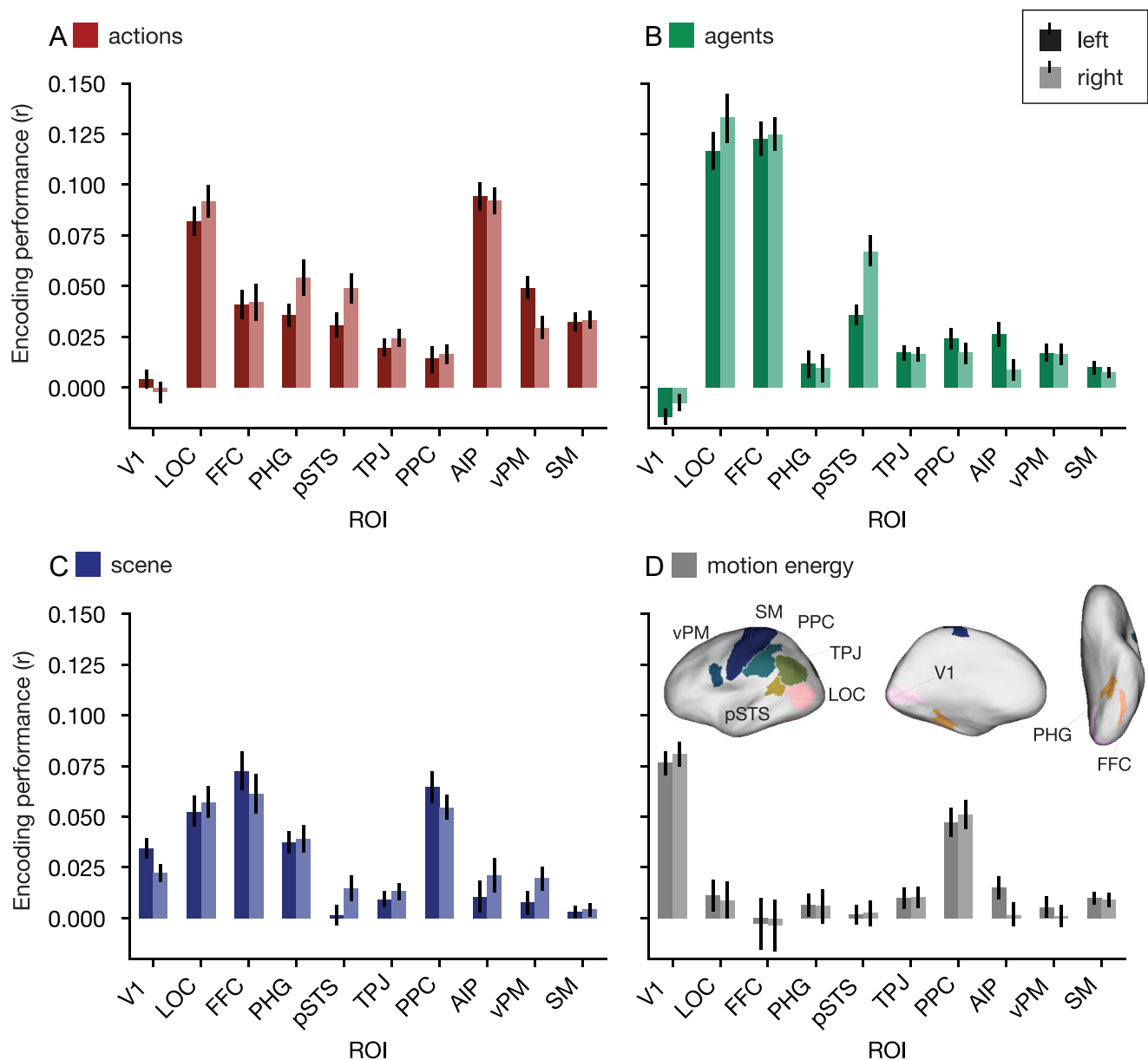


Fig. 3. Encoding performance across regions of interest. To summarize encoding results, we averaged encoding performance values across vertices within each of 10 ROIs ranging from low-level visual areas to higher-level cortical areas. Bars indicate encoding performance averaged across left-out runs and left-out subjects for the four feature spaces: (A) actions, (B) agents, (C) scene, and (D) motion energy. Homologs in the left hemisphere (dark colors) and right hemisphere (light colors) are shown separately. Error bars indicate 95% bootstrap CIs (resampling subjects).

This indicates that a much larger proportion of human cortex is dedicated to understanding observed actions than to other commensurate features. The unique motion-energy partition, on the other hand, best explained 5,586 vertices, primarily in early visual areas, and the intersection of motion energy and semantic features best explained many vertices progressing anteriorly into extrastriate cortex. Many regions exhibited mixed selectivity for multiple semantic feature spaces. For example, 1,204 vertices, primarily in lateral and ventral occipitotemporal areas, were best explained by the intersection of all three semantic feature spaces. The intersection of actions and agents best explained 1,482 vertices, including in right pSTS. In fact, the intersection of all four feature spaces best explained the largest number of vertices overall: 15,239 vertices throughout higher-level visual areas, like VT, LOC, STS, and PPC.

Generalizing to a Broader Sample of Naturalistic Stimuli. The preceding analyses focused on a naturalistic stimulus primarily featuring animals in their natural habitats. While the “Life” documentary has the benefit of including a large variety of animals (agents) across many diverse environments (scenes), it is not necessarily representative of our daily experience, which includes interacting with other humans and human-made objects/environments. With this in mind, we next aimed to test whether our findings generalize to a more diverse set of naturalistic stimuli featuring humans interacting with each other and their environments. To this end, we replicated our analyses on a completely separate fMRI dataset where $N = 72$ subjects viewed 12 different audiovisual short films (31 min in total) spanning a large variety of themes and styles (the “Spacetop” dataset; 81).

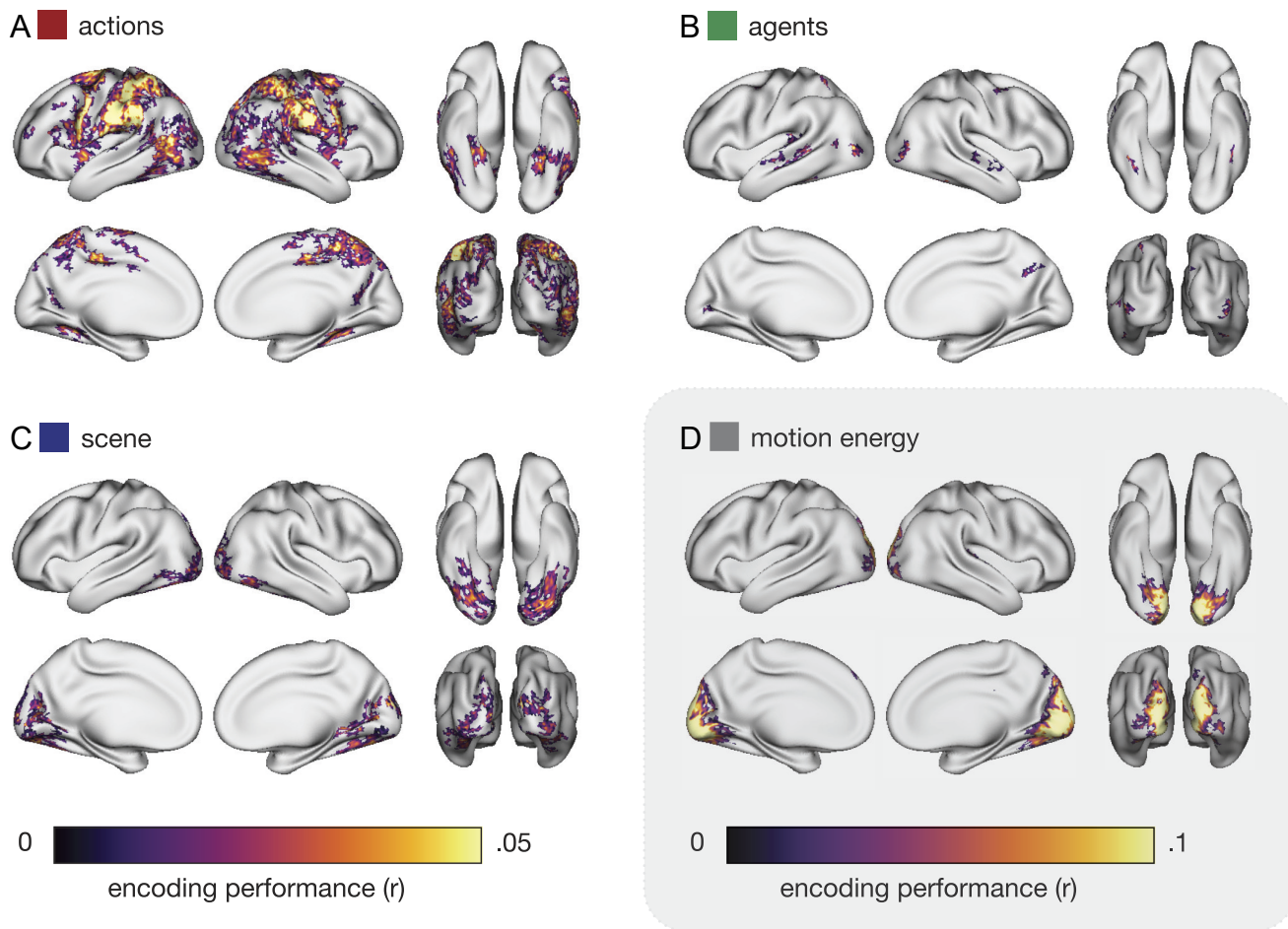


Fig. 4. Unique variance explained by each feature space. Whole-brain maps of unique variance explained by (A) action, (B) agent, (C) scene, and (D) motion-energy features. Unique variance explained by a given feature space was estimated by computing nested regression models, then subtracting the performance of a nested model excluding the features of interest from a full model containing the features of interest. Vertex-wise unique variance was estimated within each run and subject, then averaged across runs and subjects, and thresholded for statistical significance (bootstrap hypothesis test resampling subjects, FDR $P < 0.05$).

We used artificial intelligence in a semiautomated pipeline to extract text annotations from the video frames at each time point of the fMRI data (*Materials and Methods*). Notably, these videos also contained a larger number of objects, tools, and other human artifacts than the “Life” documentary. To account for this, we also extracted an object annotation in the same fashion as the action, agent, and scene annotations. For each word in each of the four annotations, we extracted the corresponding GloVe embeddings and averaged these embeddings within each TR, yielding one embedding per TR for each of the four types of features.

We next estimated within-subject voxelwise encoding models using banded ridge regression with semantic features describing the actions, agents, objects, and scenes, as well as motion energy. We used four-fold cross-validation to ensure that encoding models were always tested on a left-out set of movie stimuli unrelated to those used in training. We evaluated the predictions of each feature space (*SI Appendix, Fig. S5*) as well as the unique variance explained by each feature space (*SI Appendix, Fig. S6*) in the same way as in previous analyses. In keeping with the preceding results, we found that action features captured a broad swath of cortex, including lateral occipitotemporal cortex, parietal cortex, and premotor cortex, as well as ventral temporal cortex. The agent and scene features captured much more punctate regions of cortex. Notably, the object features also performed well in lateral occipitotemporal and parietal regions,

but still did not yield as widespread of coverage as the action features.

Across all vertices, we found that the action features, relative to other semantic features, captured the largest proportion of unique variance in 30.1% of vertices, comparable to motion energy at 32.4% of vertices; object, scene, and agent features captured the most unique variance in 20.1%, 10.9%, and 6.5% of vertices, respectively. Action features captured a significant portion of unique variance in 15,702 vertices (t test across subjects, $P < 0.05$, uncorrected); motion energy, objects, scenes, and agents explained a significant portion of unique variance in 13,403 vertices, 10,361 vertices, 1,893 vertices, and 795 vertices, respectively. Overall, these results qualitatively recapitulate our core findings in a completely independent sample with a more diverse array of more human-oriented stimuli. Action features are a particularly strong predictor of cortical activity and appear to have a particularly large cortical footprint.

Discussion

Human cortex prioritizes certain ecologically relevant features of the world for processing. Natural vision is a holistic family of processes conspiring to understand natural scenes, their constituent agents and objects, and the behavior of those agents, all in service of our behavioral goals (82, 83). Our understanding

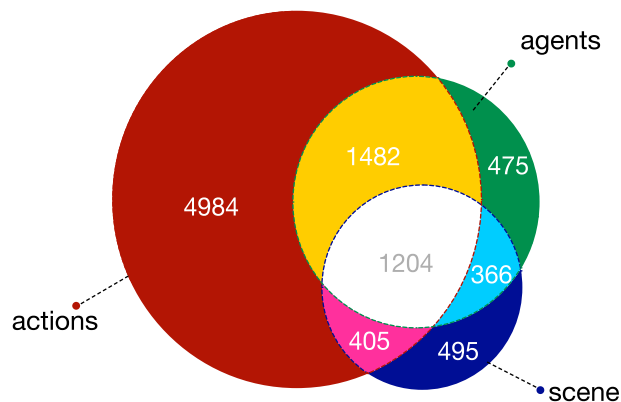


Fig. 5. Largest partition of variance at each vertex across cortex. For each vertex, we partitioned the variance explained into components comprising the unique contributions of each feature space and the intersections of two or more feature spaces. We then assessed which of these partitions corresponded to the largest component of variance at each vertex. The Venn diagram depicts, for each partition, the number of vertices for which that partition was larger than all other partitions. For example, in 4,984 vertices, action features uniquely accounted for more variance in neural activity than any other partition. The motion-energy model was included when quantifying unique variance explained, but is excluded for visualization purposes.

of the functional organization of cortex, however, stems from experiments designed to isolate particular features, typically using highly controlled experimental stimuli (e.g., refs. 18 and 84). In the current study, we used a naturalistic movie-viewing paradigm combined with a specialized modeling framework to more fairly assess the relative contributions of different features to cortical activity. We tested three different sets of semantic features reflecting the agents, actions, and scenes depicted in the stimulus, as well as a model of low-level visual motion. We found that the actions model in particular provided strong predictions across a large portion of cortex, extending from occipitotemporal areas into parietal and frontal cortices. For the “Life” stimulus, the action features captured unique variance in neural activity across an order of magnitude more cortical vertices than the agent or scene features. Our findings indicate that observed actions occupy a privileged role in cortical processing across a diverse range of clips, and in particular for the longer-form “Life” documentary.

Each of the four models we tested outperformed the other models in certain cortical areas typically associated with the respective model features (Figs. 2–4), suggesting that the different features are relatively well-matched; no feature set is degenerate or completely redundant with the others (SI Appendix, Fig. S1). The motion-energy model was the best model in early visual areas (77), but was overtaken by the semantic models in higher-level areas (68). We note that in contrast to both prior experimental work (e.g., refs. 85–87 and certain movie-viewing experiments (68, 77), we did not enforce central fixation and allowed participants to move their eyes naturally while viewing the documentary. Interestingly, while the motion-energy model yielded significant encoding performance in the vicinity of human area MT/V5 (Fig. 2D), MT encoding was largely eclipsed by the semantic models in the variance partitioning analysis (Fig. 4D). This is consistent with recent voxelwise encoding work by Dupré la Tour, Visconti di Oleggio Castello and colleagues (2025; 88) comparing motion energy with a different semantic model (WordNet) during a naturalistic movie-viewing task. Overall, this suggests that, during natural vision, activity in human area MT tends to be strongly correlated with higher-level semantic features of the stimulus.

When considering the relative contributions of different semantic features, we found that agent features were strongly encoded in ventral temporal cortex, particularly the fusiform gyrus (14, 15, 89, 90), and the right pSTS (91–94). Agent features were also a strong predictor in LOC, yielding the highest encoding performance across all models and areas in the ROI analysis (Fig. 3B). Agent features may covary with face (95, 96), body-part (97–99), and social (100) features encoded in lateral occipital and temporal areas. Scene features performed well in ventral temporal areas, including parahippocampal cortex (16, 17, 101), as well as posterior occipitoparietal cortex (102), and retrosplenial cortex (103, 104). Action features performed well throughout the action observation network, including LOC (35, 100, 105–110), as well as anterior parietal and premotor areas (33, 34, 42, 111, 112). In general, our findings align with previous controlled experiments, but extend these trends of cortical functional organization into a more naturalistic setting. We note that eye movements may play a role in our modeling results: for example, gaze may track with observed actions and contribute to encoding performance in AIP and the frontal eye fields. However, if eye movements were to bias our encoding analyses toward action features, they must correlate with action features in a way that generalizes across diverse movie clips and, for the analysis of the “Life” movie, across individual subjects (due to the structure of our cross-validation procedure). Future work may be able to isolate this eye-movement component using a similar model comparison framework (e.g., refs. 44, 113, and 114). We note that eye movements are an integral part of natural vision, however, and if they tend to track dynamic action features more so than other features, this is further evidence that action features are of particular importance to the brain.

For the preceding results, we evaluated each feature space after fitting all features jointly, allowing each set of features to compete for variance in neural activity (73, 79, 115). Although that approach allows us to assess the relative contributions of each feature space to joint encoding performance, it does not allow us to quantify how much variance in neural activity is uniquely predicted by each set of features. To quantify the unique variance explained by each set of features—more strictly controlling for all other features, including motion energy—we used a variance partitioning analysis (71, 72; Fig. 4). We found that, while the agent and scene features still accounted for unique variance in expected regions, cortical coverage was much more focal. For example, cortical vertices uniquely captured by scene features were relatively restricted to posterior ventral temporal regions, occipitoparietal cortex, and medial visual areas like parahippocampal cortex and retrosplenial cortex. Cortical vertices uniquely captured by agent features were even sparser, with localized clusters of vertices in the bilateral fusiform gyrus, lateral occipital cortex, and STS. Action features, on the other hand, captured unique variance across a much larger portion of cortex, including VT, medial parietal cortex, and canonical action observation areas extending from LOC to AIP and VPM.

One of the principal challenges of naturalistic neuroscience stems from the inherent collinearity among stimulus features in real-world contexts. For example, agent and action features tend to co-occur in natural vision, leading to intertwined features in the regression model. As a result, many cortical areas can be predicted by both sets of features (Fig. 2 A and B). However, under a more strict variance partitioning regime, we find that the agent features perform poorly and much of this variance is assigned to the action features instead (Fig. 4 A and B). This observation suggests that agent features (and to a lesser extent

scene features) are correlated with action features, but that neural activity predicted by agent and scene features tends to be better explained by action features. By including motion-energy features in our variance partitioning analysis, we ensure that this effect is not driven by low-level visual motion. That said, our variance partitioning analysis also revealed that a large portion of cortical vertices exhibit heterogeneous selectivity: activity at these vertices is best explained by the intersection of different feature spaces (*SI Appendix, Table S1*).

Note that in our annotation, action features typically correspond to verbs, whereas agent and scene (as well as object) features typically correspond to nouns, which are thought to rely on dissociable neural systems (116, 117). The cortical topographies we observe are generally consistent with prior neuroimaging work comparing verbs versus nouns (e.g., refs. 108, 118, and 119). We note, however, that our core encoding results do not boil down to a simple contrast between verbs and nouns. To obtain high encoding performance, the semantic action features in our regression model must predict changes (i.e., variance) in voxel activity across different actions present in the stimulus. If a voxel's activity was uniformly high across the actions present in the stimulus, it would not show high encoding performance for the action features due to a lack of variance across actions. Our findings provide unique insights into the relative strength with which these different features are represented in cortical activity during natural vision. In contrast to nouns, verbs may key into neural representations of abstract events (107) or the contextual/relational structure between agents and objects (e.g., refs. 120 and 121). Future work should further investigate whether the relational structures of language can provide insights into the organization of high-level visual representations (122–125).

While our encoding models were designed to compare certain theoretically motivated visual and semantic features, there are important elements of the stimulus that our models do not capture. Comparing our encoding model performance maps to the total stimulus-driven variance as estimated by an intersubject correlation (ISC) analysis (*SI Appendix, Fig. S3*) is instructive. First, even in areas where the encoding models perform well, they still only capture a portion of the variance revealed by the ISC analysis. Second, the present encoding models were derived from an annotation of the visual content of the stimulus, which excluded David Attenborough's voice-over narrative by design. The ISC map, however, reveals very strong correlations in superior temporal areas, presumably driven by the spoken narrative. Finally, even a nature documentary is designed to create narratives with suspense, resolution, comedy, and tragedy (126). We observe modest ISC in prefrontal cortex and areas of the default-mode network (DMN), such as the temporoparietal junction (TPJ) and precuneus, regions that are not well-predicted by any of our encoding models. Models that better capture the contextual structure of the narrative (e.g., ref. 127) or the social interactions and intentions of the agents in the film (e.g., ref. 128) may be required to key into these higher-order regions.

Although our encoding results recapitulate familiar maps of cortical functional organization, the relative differences in variance explained are instructive. Our core findings indicate that—for a longform nature documentary—roughly 10 times as much of cortex is allocated to representing observed actions as compared to agents and scene features (Fig. 5). When viewing agents behaving in their natural environments, the “behavior” appears to be what matters most to cortex. This aligns with recent work suggesting that the dynamic features of observed actions play an outsized role in cortical activity, including in

unexpected cortical areas like ventral temporal cortex (44, 52, 69). We contend that this primacy of action representation in vision has been historically underappreciated due to the field's reliance on highly controlled experimental manipulations and static image stimuli (43, 54, 55).

Our findings raise a difficult question: Why are cortical resources so heavily skewed toward understanding the actions of others? Why are these action features “magnified” across the cortical surface? While processing natural scenes and the agents occupying them is certainly important, the actions of an agent are likely to change most meaningfully over time. As an agent pursues its goals, its behavior dynamically evolves over time—these actions are laden with intentions, oftentimes social content, and sometimes behavioral ramifications for the observer. Humans are a remarkably social species and understanding the actions and intentions of others is the bedrock of social cognition. From an ecological perspective (129), all of these features of the world afford certain adaptive behaviors and may “compete” to capture cortical resources over the course of evolution and development (51, 130, 131). We speculate that the actions of other agents are particularly important for our own behavior, and have therefore secured a remarkably large proportion of cortical real estate.

Materials and Methods

Participants. Eighteen right-handed adults (10 reported female, mean age $\pm$ SD: 25.4 ± 2.6 y) participated in the experiment. All participants had normal or corrected-to-normal vision, normal hearing, and reported no neurological conditions. All participants provided written, informed consent prior to participating in the study. The study was approved by the Institutional Review Board of Dartmouth College. These data have been analyzed in previous publications (52, 132).

Stimuli. Participants viewed four audiovisual clips from the “Life” nature documentary narrated by David Attenborough during MRI scanning. The stimulus was divided into four 15-min runs (15.3, 14, 15.4, and 16.5 min), totaling 63 min. The video stimulus was backprojected on a screen at the end of the scanner bore and viewed with a mirror attached to the head coil. The audio track was delivered using MRI-compatible fiber-optic electrodynamic headphones. Participants freely viewed the movie and were instructed to remain still. The stimulus was presented using PsychoPy (133).

Data Acquisition and Preprocessing. MRI data were acquired using a 3T Philips Intera Achieva MRI scanner with a 32-channel phased-array head coil: TR/TE = 2500 ms, resolution = 3 mm isotropic voxels. MRI data were preprocessed using fMRIPrep (1.0.0-rc5; (134)) followed by confound regression: six head motion parameters, framewise displacement, five principal components from a conservative mask of cerebrospinal fluid and white matter, linear detrending, and a band-pass filter (0.00667 to 0.1 Hz). Whole-brain surface-based searchlight hyperalignment was used to align fine-grained functional topographies across individual brains prior to vertex-wise encoding analysis (43, 60, 80). See *SI Appendix* for further acquisition and processing details.

Stimulus Annotation. In order to quantify the semantic content of the movie stimulus, we first assigned word labels to each moment of the film (68, 132). A human annotator assigned words describing the visual content of the stimulus in roughly 2-TR segments averaging 5.17 ± 3.85 s (run 1 = 5.83; run 2 = 4.61; run 3 = 5.59; run 4 = 4.67; certain segments—e.g., the introductory sequence—were longer than others). The audio track and voice-over narrative were not presented during this annotation task. After the annotation was complete, the words in each segment were then manually filtered by a human annotator into three different types of descriptive words: words describing the actions taken by animals in the film (e.g., “swimming”), words describing the agents themselves

(e.g., "dolphin"), and words describing the scene or background (e.g., "ocean"). Words from each segment were then duplicated to effectively upsample the roughly 5-s annotation windows to the 2.5-s TR of the fMRI acquisition. The choice of action-, agent-, and scene-related words was motivated by major organizational features of high-level visual cortex reported in the neuroimaging literature (e.g., ref. 18). Note that words describing actions will tend to correspond to verbs, while words describing agents or scenes will tend to correspond to nouns. The annotations are available at the accompanying GitHub repository: <https://github.com/jungheejung/life-encoding>.

Semantic Feature Spaces. To quantify the semantic content of the visual stimulus, we extracted 300-dimensional semantic vectors from GloVe (74). GloVe embeddings have been widely used to capture semantic content encoded in neural activity (e.g., refs. 127 and 135–139). GloVe is an unsupervised learning algorithm that represents words as vectors in a high-dimensional embedding space based on their co-occurrence statistics in large corpora of text. Semantic structure is encoded in the geometric relationships between word embeddings in this high-dimensional vector space: certain semantic relationships are encoded along particular directions and semantically similar words are nearer to each other along certain dimensions (140, 141). Each word receives the same embedding regardless of its local context or usage in our annotation. For each set of annotations (actions, agents, and scene words), we compiled the embeddings into TRs and averaged embeddings for multiple words occurring within a single TR. This resulted in one embedding for actions, one embedding for agents, and one embedding for scene words per TR. Any TRs containing no words were assigned a 300-dimensional vector of zeros.

Motion-Energy Model. Prior work suggests that visual motion has a large impact on cortical activity in response to dynamic, naturalistic visual stimuli like movies (68, 77). To account for low-level visual motion, we submitted our movie stimuli to a motion-energy model (75, 76) implemented in pymoten (142, 143). Video frames were first converted to luminance values and downsampled to 192×108 pixels, yielding a 2,530-parameter pyramid of motion energy filters tiling the stimulus (according to the default parameters in pymoten). The resulting motion-energy features were log-transformed and downsampled from the original stimulus frame rate to match the TR of the fMRI data.

Vertex-Wise Encoding Models. To quantify which features of the movie stimulus are encoded in neural activity—e.g., low-level motion features or higher-level semantic features—we constructed vertex-wise encoding models (68, 78). Vertex-wise encoding models estimate a linear mapping from a representation of the stimulus in a particular to feature space onto the time series of neural activity. To more equitably compare different feature spaces, we use banded ridge regression to fit all feature spaces simultaneously in a joint model (79, 144). We quantify how well these features map onto vertex-wise activity by 1) evaluating the predictions from each feature space separately and 2) evaluating the unique variance explained by each feature space (and intersection of feature spaces). We describe the vertex-wise encoding analysis in greater detail in *SI Appendix*. To evaluate the relative contributions of each feature space—actions, agents, scene, and motion energy—we generated model-based predictions using the regression weights for each feature band separately (effectively zeroing out weights for other bands; 73). We then computed the correlation between predicted time series and actual fMRI time series in the left-out run and left-out test subject for each feature space. This results in four separate encoding performance maps, one for each feature space.

Variance Partitioning. We performed two separate variance partitioning analyses: 1) we computed the unique variance explained by each of the four models and assessed the statistical significance of these unique variance values (Fig. 4); 2) we computed all 15 partitions comprising the unique contributions and intersections between two or more sets of features, then determined which of these 15 partitions was largest for each vertex (Fig. 5). In the first analysis, we fit a full model comprising all four sets of features, then fit nested models excluding the feature space of interest. For a given feature space of interest, we quantified the unique variance explained by that feature space as $r_{\text{unique}} = r_{\text{full}} - r_{\text{nested}}$, where r_{full} corresponds to the encoding

performance for the full model with all four sets of features, and r_{nested} corresponds to the encoding performance for the nested model excluding the model of interest. We computed these vertex-wise r_{unique} values for each subject, then assessed statistical significance across subjects, resulting in a whole-brain map of statistically significant r_{unique} values for each of the four feature spaces (Fig. 4). In the second analysis, we fully partitioned the variance explained by each combination of feature spaces using commonality analysis (71, 72, 145, 146). We ran encoding models for all possible combinations of feature spaces, resulting in 15 models (e.g., single features, pairs, triplets, and all four features: $4C_1 + 4C_2 + 4C_3 + 4C_4$). We evaluated these models in the same way as previous models, by computing the correlation between model-predicted and actual vertex-wise time series (73) for left-out runs in left-out subjects. To simplify the commonality analysis, for each of the 15 models, we first averaged correlations (with z-transformation) across both runs and subjects, to obtain a single encoding performance score for each model. We used set theory to arithmetically determine the variance assigned uniquely to each of the four feature spaces as well as each intersection of two, three, or all four features for every vertex. This allowed us to quantify the variance uniquely explained by one feature space by isolating the variance specific to that feature space while excluding all other feature spaces and their intersections. In keeping with prior work (73), we quantified unique variance explained in terms of the out-of-sample correlation between predicted and actual time series. Finally, we identified which partition—i.e., unique feature space or intersection of feature spaces—explained the most variance in vertex-wise activity. This allowed us to quantify, for example, the number of vertices best explained uniquely by action features across cortex, or the number of vertices best explained by the intersection of action and motion-energy features. We computed these vertex counts within a mask of vertices with significant model performance for the joint model combining all four feature spaces (39,031 significant vertices in the mask in total across both hemispheres; bootstrap hypothesis test, $\text{FDR } P < 0.05$, *SI Appendix*, Fig. S2). Motion energy features were included in the variance partitioning analysis, although we focus primarily on the three semantic feature sets. Note that vertices identified in the preceding analyses as having significant unique variance explained by a given feature space (Fig. 4) may be better explained by a different combination of features in this second "winner-take-all" analysis (Fig. 5).

Statistical Evaluation. To determine whether the mean encoding performance (correlation between predicted and actual response time series) across subjects was statistically greater than zero, we performed a one-sample *t* test. To compare encoding performance between models, we submitted the subject-level correlation values for both models to a paired *t* test. Encoding performance values were averaged across the four left-out runs (four different segments of the film) within each subject prior to statistical evaluation. Given that the unique variance estimates are biased toward positive values due to the nested structure of the models, we used a bootstrap hypothesis test to more conservatively assess significance (147). To correct for multiple tests across vertices, we controlled the false discovery rate (FDR; 148) at .05 across both hemispheres (excluding vertices in the medial wall). To simplify visualization of whole-brain model performance, we omitted any clusters of significant vertices containing fewer than 100 contiguous surface vertices (Fig. 2 and *SI Appendix*, Fig. S4); when visualizing unique variance explained (Fig. 4), we omitted clusters with fewer than 50 significant contiguous surface vertices.

Generalization to an Independent Sample. To assess the generalizability of our findings, we aimed to replicate our analyses on a completely separate fMRI dataset with naturalistic movie viewing: the openly available "Spacetop" dataset (81); (<https://openneuro.org/datasets/ds005256/versions/1.1.0>). In the absence of human-generated annotations, we developed a semiautomated pipeline for producing structured, frame-level descriptions of the video using vision language models (VLMs). The video stimuli were decoded using OpenCV, frames were extracted at a fixed sampling rate, and each frame together with its timestamp was cached to disk along with a metadata file recording extraction parameters and frame indices. Frames were then submitted in batches to a vision language model, with the analyzer backend selectable between Anthropic's

Claude 3.5 Sonnet (149) and Google's Gemini 1.5 Flash (150) through a common interface, so that the same frames could be annotated by either provider without changing the downstream code. For each frame we passed the image together with a fixed context prompt describing the labels of interest (for example, the agent present, the action being performed, and the surrounding scene), and the model returned a structured JSON response per frame. API calls were issued concurrently within each batch (default of five in-flight requests) with automatic retries on transient failures, and partial results were written incrementally to disk so that interrupted runs could be resumed without re-extracting frames or requerying frames that had already been successfully annotated. The per frame JSON outputs were then concatenated into a single annotation file per video, with one timestamped entry per sampled frame. These annotations were then downsampled to the TR of the fMRI data in the same manner as the human annotation described previously. Model identity, prompt text, sampling rate, and concurrency settings were stored in a YAML configuration file alongside the outputs to support reproducibility, and a shell wrapper allowed the same pipeline to be run across a directory of stimulus videos to produce one annotation file per video. The code and prompts used to obtain the annotation are publicly available on GitHub: https://github.com/hunarjain09/video_frame_analyzer. The annotations were manually inspected by the authors for quality. Since the Spacetop stimuli sample more typical human-oriented films than the "Life" documentary, they also feature a much larger diversity of human-made objects (e.g., tools, furniture) and human-made environments (e.g., kitchen, alleyway) than the nature documentary. To account for this, we also extracted an annotation describing the objects in each stimulus, in addition to the actions, agents, and scene features. The automated annotation pipeline resulted in a number of words describing the stimulus frames at each TR for each of the four sets of features (actions, agents, objects, scene). Using the same procedure as the preceding analyses, we extracted GloVe embeddings for each of these words, then averaged the GloVe embeddings within a single TR, resulting in four time series of GloVe embeddings matched to the fMRI data for each feature type. We also extracted motion-energy features from the Spacetop stimuli using the same procedure as the preceding analyses. We next reimplemented a voxelwise encoding analysis using banded ridge regression with five different types of features: action, agent, object, scene, and motion energy. For this analysis, we omitted hyperalignment

and ran within-subject encoding models following a more typical pipeline. To mimic the structure of our original analysis, we implemented four-fold cross-validation across the 12 movie stimuli. For each subject, we first randomized the order of the 12 movie stimuli prior to cross-validation, since the 12 stimuli were not evenly distributed across scanning runs. This ensures that a given movie stimulus was never split across folds and that the encoding models were always tested on a nonoverlapping set of unrelated movie stimuli. We ran full models comprising all five feature spaces and evaluated the performance of each feature band separately (SI Appendix, Fig. S5). We also ran five nested models excluding one feature space each to compute a unique variance score in the same fashion as the preceding analysis (SI Appendix, Fig. S6). We then averaged these unique variance values across test folds within each subject. To accommodate individual differences in cortical-functional topography, we applied spatial smoothing with a 10-mm full-width at half maximum (FWHM) Gaussian kernel to the subject-level unique variance scores on the cortical surface. Finally, we computed a *t* test to evaluate the statistical significance of these unique variance scores across subjects, controlling FDR at $P < 0.05$.

Data, Materials, and Software Availability. The functional MRI data generated and analyzed for this study are publicly available via DataLad: <http://datasets.datalad.org/?dir=/labs/haxby/life> (151). The analysis code used for this manuscript is publicly available at the following GitHub repositories: <https://github.com/jungheejung/life-encoding> (152) and https://github.com/hunarjain09/video_frame_analyzer (153).

ACKNOWLEDGMENTS. We would like to thank Courtney Rogers, Terry Sackett, and the Dartmouth Brain Imaging Center for help with data collection, as well as Zaid Zada and Vassiki Chauhan for help with modeling and visualization. We also thank Research Computing and Data Services of Information, Technology & Consulting at Dartmouth College for their support on conducting analyses on the Discovery Cluster high performance computing environment.

Author affiliations: ^aDepartment of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755; ^bIndependent Researcher, Sunnyvale, CA 94085; and ^cDepartment of Psychology, Center for Computational Language Sciences, University of Southern California, Los Angeles, CA 90089

- J. H. Kaas, Why does the brain have so many visual areas? *J. Cogn. Neurosci.* **1**, 121–135 (1989).
- J. H. Kaas, "The evolution of the visual system in primates" in *The Visual Neurosciences*, L. M. Chalupa, J. S. Werner, Eds. (MIT Press, 2004), vol. 2, pp. 1563–1572.
- D. J. Felleman, D. C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
- P. M. Daniel, D. Whitteridge, The representation of the visual field on the cerebral cortex in monkeys. *J. Physiol.* **159**, 203–221 (1961).
- A. Cowey, E. T. Rolls, Human cortical magnification factor and its relation to visual acuity. *Exp. Brain Res.* **21**, 447–454 (1974).
- J. Rovamo, V. Virsu, R. Näsänen, Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature* **271**, 54–56 (1978).
- R. Malach, I. Levy, U. Hasson, The topography of high-order human object areas. *Trends Cogn. Sci.* **6**, 176–184 (2002).
- R. O. Duncan, G. M. Boynton, Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron* **38**, 659–671 (2003).
- W. Penfield, E. Boldrey, Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* **60**, 389–443 (1937).
- C. N. Woolsey, W. H. Marshall, P. Bard, Representation of cutaneous tactile sensibility in the cerebral cortex of the monkey as indicated by evoked potentials. *Johns Hopkins Hosp. Bull.* **70**, 399–441 (1942).
- W. Penfield, T. Rasmussen, *The Cerebral Cortex of Man: A Clinical Study of Localization of Function* (Macmillan, 1950).
- R. J. Nudo, W. M. Jenkins, M. M. Merzenich, T. Prejean, R. Grenda, Neurophysiological correlates of hand preference in primary motor cortex of adult squirrel monkeys. *J. Neurosci.* **12**, 2918–2947 (1992).
- A. Pascual-Leone, F. Torres, Plasticity of the sensorimotor cortex representation of the reading finger in braille readers. *Brain* **116**, 39–52 (1993).
- N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
- G. McCarthy, A. Puce, J. C. Gore, T. Allison, Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* **9**, 605–610 (1997).
- G. K. Aguirre, E. Zarahn, M. D'Esposito, An area within human ventral cortex sensitive to "building" stimuli: Evidence and implications. *Neuron* **21**, 373–383 (1998).
- R. Epstein, N. Kanwisher, A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
- N. Kanwisher, Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11163–11170 (2010).
- D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, M. Mishkin, The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).
- K. Grill-Spector, K. S. Weiner, The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
- L. W. Barsalou, W. Kyle Simmons, A. K. Barbey, C. D. Wilson, Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci.* **7**, 84–91 (2003).
- A. Caramazza, B. Z. Mahon, The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends Cogn. Sci.* **7**, 354–361 (2003).
- A. Martin, The representation of object concepts in the brain. *Annu. Rev. Psychol.* **58**, 25–45 (2007).
- A. Martin, Grapes-grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychon. Bull. Rev.* **23**, 979–990 (2016).
- K. Patterson, P. J. Nestor, T. T. Rogers, Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* **8**, 976–987 (2007).
- B. Z. Mahon, A. Caramazza, Concepts and categories: A cognitive neuropsychological perspective. *Annu. Rev. Psychol.* **60**, 27–51 (2009).
- E. Bonda, M. Petrides, D. Ostry, A. Evans, Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *J. Neurosci.* **16**, 3737–3744 (1996).
- E. D. Grossman *et al.*, Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.* **12**, 711–720 (2000).
- M. S. Beauchamp, K. E. Lee, J. V. Haxby, A. Martin, fMRI responses to video and point-light displays of moving humans and manipulable objects. *J. Cogn. Neurosci.* **15**, 991–1001 (2003).
- A. Puce, D. Perrett, Electrophysiology and brain imaging of biological motion. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **358**, 435–445 (2003).
- M. I. Gobbi, A. C. Koralek, R. E. Bryan, K. J. Montgomery, J. V. Haxby, Two takes on the social brain: A comparison of theory of mind tasks. *J. Cogn. Neurosci.* **19**, 1803–1814 (2007).
- F. Castelli, F. Happé, U. Frith, C. Frith, Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* **12**, 314–325 (2000).
- S. T. Grafton, A. F. d. C. Hamilton, Evidence for a distributed hierarchy of action representation in the brain. *Hum. Mov. Sci.* **26**, 590–616 (2007).
- N. N. Oosterhof, S. P. Tipper, P. E. Downing, Crossmodal and action-specific: Neuroimaging the human mirror neuron system. *Trends Cogn. Sci.* **17**, 311–318 (2013).
- A. Lingnau, P. E. Downing, The lateral occipitotemporal cortex in action. *Trends Cogn. Sci.* **19**, 268–277 (2015).

36. E. D. Grossman, R. Blake, Brain areas active during visual perception of biological motion. *Neuron* **35**, 1167–1175 (2002).
37. A. P. Saygin *et al.*, Point-light biological motion perception activates human premotor cortex. *J. Neurosci.* **24**, 6181–6188 (2004).
38. N. N. Oosterhof, S. P. Tipper, P. E. Downing, Viewpoint (in)dependence of action representations: An mvpa study. *J. Cogn. Neurosci.* **24**, 975–989 (2012).
39. M. F. Wurm, A. Lingnau, Decoding actions at different levels of abstraction. *J. Neurosci.* **35**, 7727–7735 (2015).
40. D. Pitcher, L. G. Ungerleider, Evidence for a third visual pathway specialized for social perception. *Trends Cogn. Sci.* **25**, 100–110 (2021).
41. E. McMahon, L. Isik, Seeing social interactions. *Trends Cogn. Sci.* **27**, 1165–1179 (2023).
42. G. Rizzolatti, C. Sinigaglia, The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nat. Rev. Neurosci.* **11**, 264–274 (2010).
43. J. V. Haxby, M. I. Gobbini, S. A. Nastase, Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *NeuroImage* **216**, 116561 (2020).
44. J. Han *et al.*, Behaviorally-relevant features of observed actions dominate cortical representational geometry in natural vision. *Nat Commun.* **10**, 1038/s41467-026-76599-w (2026).
45. J. V. Haxby *et al.*, Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
46. M. S. Beauchamp, K. E. Lee, J. V. Haxby, A. Martin, Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* **34**, 149–159 (2002).
47. B. A. Olshausen, D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
48. E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
49. S. V. David, W. E. Vinje, J. L. Gallant, Natural stimulus statistics alter the receptive field structure of v1 neurons. *J. Neurosci.* **24**, 6991–7006 (2004).
50. F. E. Theunissen, J. E. Elie, Neural processing of natural sounds. *Nat. Rev. Neurosci.* **15**, 355–366 (2014).
51. U. Hasson, S. A. Nastase, A. Goldstein, Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
52. S. A. Nastase *et al.*, Attention selectively reshapes the geometry of distributed semantic representation. *Cereb. Cortex* **27**, 4277–4291 (2017).
53. L. S. Hamilton, A. G. Huth, The revolution will not be controlled: Natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* **35**, 573–582 (2020).
54. D. A. Leopold, S. H. Park, Studying the visual brain in its natural rhythm. *NeuroImage* **216**, 116790 (2020).
55. S. A. Nastase, A. Goldstein, U. Hasson, Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage* **222**, 117254 (2020).
56. A. Bartels, S. Zeki, Functional brain mapping during free viewing of natural scenes. *Hum. Brain Mapp.* **21**, 75–85 (2004).
57. U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
58. U. Hasson, R. Malach, D. J. Heeger, Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* **14**, 40–48 (2010).
59. I. P. Jääskeläinen *et al.*, Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *Open Neuroimaging J.* **2**, 14–19 (2008).
60. J. V. Haxby *et al.*, A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
61. C. Baldassano *et al.*, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).
62. D. Kim, K. Kay, G. L. Shulman, M. Corbetta, A new modular brain organization of the bold signal during natural vision. *Cereb. Cortex* **28**, 3065–3081 (2018).
63. H. Richardson, F. Lisandrelli, A. Riobueno-Naylor, R. Saxe, Development of the social brain from age three to twelve years. *Nat. Commun.* **9**, 1027 (2018).
64. E. Yang *et al.*, The default network dominates neural responses to evolving movie stories. *Nat. Commun.* **14**, 4197 (2023).
65. R. Rajimehr *et al.*, Functional architecture of cerebral cortex during naturalistic movie watching. *Neuron* **112**, 4130–4146 (2024).
66. E. Brunswik, Representative design and probabilistic theory in a functional psychology. *Psychol. Rev.* **62**, 193–217 (1955).
67. S. Sonkusare, M. Breakspear, C. Guo, Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
68. A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
69. B. E. Russ, D. A. Leopold, Functional MRI mapping of dynamic visual features during natural viewing in the macaque. *NeuroImage* **109**, 84–94 (2015).
70. B. E. Russ, K. W. Koyano, J. Day-Cooney, N. Perwez, D. A. Leopold, Temporal continuity shapes visual responses of macaque face patch neurons. *Neuron* **111**, 903–914 (2023).
71. M. D. Lescroart, D. E. Stansbury, J. L. Gallant, Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Front. Comput. Neurosci.* **9**, 135 (2015).
72. W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, F. E. Theunissen, The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
73. H. Lee Masson, L. Isik, Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* **245**, 118741 (2021).
74. J. Pennington, R. Socher, C. Manning, “GloVe: Global vectors for word representation” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, W. Daelemans, Eds. (Association for Computational Linguistics, 2014), pp. 1532–1543.
75. E. H. Adelson, J. R. Bergen, Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
76. A. B. Watson, J. Ahumada, A. J., Model of human visual-motion sensing. *J. Opt. Soc. Am. A* **2**, 322–341 (1985).
77. S. Nishimoto *et al.*, Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
78. T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
79. T. Dupré la Tour, M. Eickenberg, A. O. Nunez-Elizalde, J. L. Gallant, Feature-space selection with banded ridge regression. *NeuroImage* **264**, 119728 (2022).
80. J. S. Guntupalli *et al.*, A model of representational spaces in human cortex. *Cereb. Cortex* **26**, 2919–2934 (2016).
81. H. Jung *et al.*, Spacetop: A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks. *Sci. Data* **12**, 1465 (2025).
82. M. V. Peelen, P. E. Downing, Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia* **105**, 177–183 (2017).
83. S. Bracci, H. P. Op, de Beeck, Understanding human object vision: A picture is worth a thousand representations. *Annu. Rev. Psychol.* **74**, 113–135 (2023).
84. R. Saxe, M. Brett, N. Kanwisher, Divide and conquer: A defense of functional localizers. *NeuroImage* **30**, 1088–1099 (2006).
85. S. Zeki *et al.*, A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11**, 641–649 (1991).
86. J. D. Watson *et al.*, Area v5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cereb. Cortex* **3**, 79–94 (1993).
87. R. B. Tootell *et al.*, Functional analysis of human mt and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* **15**, 3215–3230 (1995).
88. T. Dupré la Tour, M. Visconti di Oleggio Castello, J. L. Gallant, The voxelwise encoding model framework: a tutorial introduction to fitting encoding models to fMRI data. *Imaging Neurosci.* **3**, imag_a_00575 (2025).
89. L. L. Chao, J. V. Haxby, A. Martin, Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat. Neurosci.* **2**, 913–919 (1999).
90. N. Kanwisher, G. Yovel, The fusiform face area: A cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **361**, 2109–2128 (2006).
91. C. P. Said, C. D. Moore, A. D. Engell, A. Todorov, J. V. Haxby, Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J. Vis.* **10**, 11 (2010).
92. J. D. Carlin, A. J. Calder, N. Kriegeskorte, H. Nili, J. B. Rowe, A head view-invariant representation of gaze direction in anterior superior temporal sulcus. *Curr. Biol.* **21**, 1817–1821 (2011).
93. D. Pitcher, D. D. Dilks, R. R. Saxe, C. Triantafyllou, N. Kanwisher, Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage* **56**, 2356–2363 (2011).
94. B. Deen, K. Koldewyn, N. Kanwisher, R. Saxe, Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* **25**, 4596–4609 (2015).
95. I. Gauthier *et al.*, The fusiform “face area” is part of a network that processes faces at the individual level. *J. Cogn. Neurosci.* **12**, 495–504 (2000).
96. J. V. Haxby, E. A. Hoffman, M. I. Gobbini, The distributed human neural system for face perception. *Trends Cogn. Sci.* **4**, 223–233 (2000).
97. P. E. Downing, Y. Jiang, M. Shuman, N. Kanwisher, A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
98. T. Orlov, T. R. Makin, E. Zohary, Topographic representation of the human body in the occipitotemporal cortex. *Neuron* **68**, 586–600 (2010).
99. S. Bracci, C. Cavina-Pratesi, M. Ietswaart, A. Caramazza, M. V. Peelen, Closely overlapping responses to tools and hands in left lateral occipitotemporal cortex. *J. Neurophysiol.* **107**, 1443–1456 (2012).
100. M. F. Wurm, A. Caramazza, A. Lingnau, Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *J. Neurosci.* **37**, 562–575 (2017).
101. M. F. Bonner, A. R. Price, J. E. Peelle, M. Grossman, Semantics of the visual environment encoded in parahippocampal cortex. *J. Cogn. Neurosci.* **28**, 361–378 (2016).
102. D. D. Dilks, J. B. Julian, A. M. Pounov, N. Kanwisher, The occipital place area is causally and selectively involved in scene perception. *J. Neurosci.* **33**, 1331–1336 (2013).
103. S. D. Vann, J. P. Aggleton, E. A. Maguire, What does the retrosplenial cortex do? *Nat. Rev. Neurosci.* **10**, 792–802 (2009).
104. S. Park, T. Konkle, A. Oliva, Parametric coding of the size and clutter of natural scenes in the human brain. *Cereb. Cortex* **25**, 1792–1805 (2015).
105. A. Martin, J. V. Haxby, F. M. Lalonde, C. L. Wiggs, L. G. Ungerleider, Discrete cortical regions associated with knowledge of color and knowledge of action. *Science* **270**, 102–105 (1995).
106. J. Decety, J. Grèzes, Neural mechanisms subserving the perception of human actions. *Trends Cogn. Sci.* **3**, 172–178 (1999).
107. J. W. Kable, J. Lease-Spellmeyer, A. Chatterjee, Neural substrates of action event knowledge. *J. Cogn. Neurosci.* **14**, 795–805 (2002).
108. J. W. Kable, I. P. Kan, A. Wilson, S. L. Thompson-Schill, A. Chatterjee, Conceptual representations of action in the lateral temporal cortex. *J. Cogn. Neurosci.* **17**, 1855–1870 (2005).
109. S. Kalénine, L. J. Buxbaum, H. B. Coslett, Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. *Brain* **133**, 3269–3280 (2010).
110. E. McMahon, M. F. Bonner, L. Isik, Hierarchical organization of social action features along the lateral visual pathway. *Curr. Biol.* **33**, 5035–5047 (2023).
111. S. Caspers, K. Zilles, A. R. Laird, S. B. Eickhoff, ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage* **50**, 1148–1167 (2010).
112. L. Tarhan, T. Konkle, Sociality and interaction envelope organize visual action representations. *Nat. Commun.* **11**, 3002 (2020).
113. M. Frey, M. Nau, C. F. Doeller, Magnetic resonance-based eye tracking using deep neural networks. *Nat. Neurosci.* **24**, 1772–1779 (2021).
114. D. Gozukara, N. Ahmad, K. Seeliger, D. Oetinger, L. Geerlings, Neural network-based encoding in free-viewing fMRI with gaze-aware models. *Neurons Behav. Data Anal. Theory* **9** (2026).
115. H. Small, H. Lee Masson, E. Wodka, S. Mostofsky, L. Isik, Ubiquitous cortical sensitivity to visual information during naturalistic, audiovisual movie viewing. *PsyArXiv [Preprint]* (2025). https://doi.org/10.31234/osf.io/b5p4n_v1 (Accessed 20 September 2025).
116. A. Caramazza, A. E. Hillis, Lexical organization of nouns and verbs in the brain. *Nature* **349**, 788–790 (1991).

117. A. R. Damasio, D. Tranel, Nouns and verbs are retrieved with differently distributed neural systems. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4957–4960 (1993).
118. H. Damasio *et al.*, Neural correlates of naming actions and of naming spatial relations. *NeuroImage* **13**, 1053–1064 (2001).
119. M. Bedny, A. Caramazza, E. Grossman, A. Pascual-Leone, R. Saxe, Concepts are more than percepts: The case of action verbs. *J. Neurosci.* **28**, 11347–11353 (2008).
120. E. Abassi, L. Papeo, The representation of two-body shapes in the human visual cortex. *J. Neurosci.* **40**, 852–863 (2020).
121. G. L. Quek, M. V. Peelen, Contextual and spatial associations between objects interactively modulate visual processing. *Cereb. Cortex* **30**, 6391–6404 (2020).
122. A. Y. Wang, K. Kay, T. Naselaris, M. J. Tarr, L. Wehbe, Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat. Mach. Intell.* **5**, 1415–1426 (2023).
123. C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, T. Konkle, A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
124. B. Lahner *et al.*, Modeling short visual events through the bold moments video fMRI dataset and metadata. *Nat. Commun.* **15**, 6241 (2024).
125. A. Doerig *et al.*, High-level visual representations in the human brain are aligned with large language models. *Nat. Mach. Intell.* **7**, 1220–1234 (2025).
126. C. Grall, E. S. Finn, Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Soc. Cogn. Affect. Neurosci.* **17**, 598–608 (2022).
127. S. Kumar *et al.*, Shared functional specialization in transformer-based language models and the human brain. *Nat. Commun.* **15**, 5523 (2024).
128. Z. Miao *et al.*, Common and distinct neural correlates of social interaction processing and theory of mind in narratives. *Nat. Commun.* **17**, 4830 (2026).
129. J. J. Gibson, *The Ecological Approach to Visual Perception* (Houghton Mifflin, 1979).
130. A. M. Zador, A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 3770 (2019).
131. M. J. Arcaro, M. S. Livingstone, On the relationship between maps and domains in inferotemporal cortex. *Nat. Rev. Neurosci.* **22**, 573–583 (2021).
132. C. E. Van Uden *et al.*, Modeling semantic encoding in a common neural representational space. *Front. Neurosci.* **12**, 437 (2018).
133. J. W. Peirce, Generating stimuli for neuroscience using psychopy. *Front. Neuroinf.* **2**, 10 (2008).
134. O. Esteban *et al.*, fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
135. F. Pereira *et al.*, Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
136. A. J. Anderson *et al.*, Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J. Neurosci.* **41**, 4100–4119 (2021).
137. L. Fernandino, J. Q. Tong, L. L. Conant, C. J. Humphries, J. R. Binder, Decoding the information structure underlying the neural representation of concepts. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2108091119 (2022).
138. A. Goldstein *et al.*, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
139. G. Grand, I. A. Blank, F. Pereira, E. Fedorenko, Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **6**, 975–987 (2022).
140. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
141. T. Mikolov, W. T. Yih, G. Zweig, "Linguistic regularities in continuous space word representations" in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, K. Kirchhoff, Eds. (Association for Computational Linguistics, 2013), pp. 746–751.
142. S. Nishimoto, J. L. Gallant, A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *J. Neurosci.* **31**, 14551–14564 (2011).
143. A. O. Nunez-Elizalde, F. Deniz, T. Dupré la Tour, M. Visconti di Oleggio Castello, J. L. Gallant, pymoten: scientific python package for computing motion energy features from video. Zenodo. <https://doi.org/10.5281/zenodo.6349625>. Accessed 14 February 2023.
144. A. O. Nunez-Elizalde, A. G. Huth, J. L. Gallant, Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage* **197**, 482–492 (2019).
145. A. M. Mood, Partitioning variance in multiple regression analyses as a tool for developing learning models. *Am. Educ. Res. J.* **8**, 191–202 (1971).
146. D. R. Seibold, R. D. McPhee, Commonality analysis: A method for decomposing explained variance in multiple regression analyses. *Hum. Commun. Res.* **5**, 355–365 (1979).
147. P. Hall, S. R. Wilson, Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762 (1991).
148. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
149. Anthropic, The Claude 3 Model Family: Opus, Sonnet, Haiku (Anthropic, San Francisco, CA, 2024). https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed 15 August 2025.
150. Gemini Team, Google, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv [Preprint] (2024). <https://arxiv.org/abs/2403.05530> (Accessed 16 December 2024).
151. S. A. Nastase, Y. O. Halchenko, A. C. Connolly, J. V. Haxby, Neural responses to naturalistic clips of animals. DataLad. <https://datasets.datalad.org/?dir=/labs/haxby/life> (2017).
152. H. Jung, X. Han, S. A. Nastase, life-encoding. GitHub. <https://github.com/junghheejung/life-encoding>. Deposited 26 May 2020.
153. H. Jain, Video frame analysis system. GitHub. https://github.com/hunarjain09/video_frame_analyzer. Deposited 2 August 2025.