BOLD Moments: modeling short visual events through a video fMRI dataset and metadata

Benjamin Lahner[1,*], Kshitij Dwivedi[2,3,^], Polina Iamshchinina[2], Monika Graumann[2], Alex Lascelles[1], Gemma Roig[3,5], Alessandro Thomas Gifford[2], Bowen Pan[1], SouYoung Jin[1], N. Apurva Ratan Murty[6], Kendrick Kay[4], Aude Oliva[1,+], Radoslaw Cichy[2,+]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

[2] Department of Education and Psychology, Freie Universität Berlin, Germany

[3] Department of Computer Science, Goethe University Frankfurt, Germany

[4] Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota, Minneapolis, MN, USA

[5] The Hessian Center for AI (hessian.AI), Darmstadt, Germany

[6] Department of Brain and Cognitive Science, MIT, Cambridge, MA, USA

[+] denotes equal senior author contribution

[*] correspondence: blahner@mit.edu

Summary

Grasping the meaning of everyday visual events is a fundamental feat of human intelligence that hinges on diverse neural processes ranging from vision to higher-level cognition. Deciphering the neural basis of visual event understanding requires rich, extensive, and appropriately designed experimental data. However, this type of data is hitherto missing. To fill this gap, we introduce the BOLD Moments Dataset (BMD), a large dataset of whole-brain fMRI responses to over 1,000 short (3s) naturalistic video clips and accompanying metadata. We show visual events interface with an array of processes, extending even to memory, and we reveal a match in hierarchical processing between brains and video-computable deep neural networks. Furthermore, we showcase that BMD successfully captures temporal dynamics of visual events at second resolution. BMD thus establishes a critical groundwork for investigations of the neural basis of visual event understanding.

^ work was done prior to joining Amazon

Introduction

Understanding visual events is a hallmark of human intelligence that engages a distributed and functionally diverse network of cortical regions. This complexity provides a compelling model system for ecologically-valid cognition but is challenging to study. For example, consider a short and simple visual stimulus in which a person opens a door. To extract the underlying meaning of this stimulus, the brain must execute the entire visual object recognition processing cascade to identify the most relevant objects (a person, a door) (Carandini, 2005; DeYoe & Van Essen, 1988; DiCarlo et al., 2012; Felleman & Van Essen, 1991; Logothetis & Sheinberg, 1996; Ress & Heeger, 2003). Additionally, the brain must integrate information over time to understand the temporal relationships between objects (the door is being opened, not closed) (Fairhall et al., 2014; Hasson, Yang, et al., 2008; Orlov & Zohary, 2018). Finally, the extracted visual information must be integrated with other cognitive faculties, such as memory and emotion, and incite an appropriate behavioral response (Bainbridge, 2019; Buccino et al., 2004; Bylinskii et al., 2022; Calvo-Merino et al., 2005; Hardwick et al., 2018; Iacoboni et al., 2005; Kanske et al., 2015; Schneider, 2013). For instance, is the person opening the door angry or frightened, why is the person opening the door, and should I follow?

Investigation of visual event understanding requires appropriate experimental measures. Critically, visual stimuli should consist of naturalistic videos in order to engage rich, real-life neural processes (Berkes et al., 2011; Olshausen & Field, 1996a, 1996b; Smyth et al., 2003). While still images are a commonly used form of naturalistic visual stimuli, they do not evoke the same extent (Bartels & Zeki, 2004; Konen & Kastner, 2008; Press et al., 2001; Schultz & Pilz, 2009; Yildirim et al., 2019) or pattern (Buccino et al., 2004; Kret et al., 2011) of neural responses that dynamic visual stimuli evoke. Moreover, still images cannot cover the space of possible visual events (e.g., is the person staring or looking around?) and lack the nuances that are critical to contextualize an event (e.g., the person is looking around in order to find a friend). Longform movies could address these limitations, but their extended length and production edits introduce ambiguous temporal event boundaries, long-term interactions in event understanding, and other complications (Hasson, Landesman, et al., 2008; Roberts et al., 2013). Thus, short naturalistic videos strike an ideal balance between ecological validity and experimental control.

In this paper, we introduce the BOLD Moments Dataset (BMD), an extensive and carefully designed dataset of fMRI responses to a large number (1,102) of short (3 s) naturalistic videos.

The videos are sampled from the Moments in Time video dataset (Monfort et al., 2020, 2022; Newman et al., 2020) to facilitate collaborations with state-of-the-art computer vision models. The videos span a wide range of events that humans may witness, varying in their novelty, levels of motion, closeness to the observer, scenery, and more, in order to evoke a representative range of brain responses. To enhance the potential of BMD, we also collect and prepare a comprehensive set of video annotations (metadata) that detail behaviorally meaningful aspects of visual event understanding.

Below, we detail the high quality of BMD and showcase its suitability for a diverse set of state-of-the-art analysis and modeling approaches. First, we demonstrate that meaningful temporal information can be successfully extracted from the data despite the sluggishness of the BOLD response. We then show how the scale and quality of BMD enables integration with analysis approaches based on deep neural network models. These modeling analyses reveal that sentence-level descriptions of the videos better characterize neural responses compared to simple scene- or action-labels. Finally, we show that the memorability of the videos is correlated with brain responses in high-level visual cortex, indicating that BMD successfully taps into complex cognitive functions involved in visual event understanding.

We contend that BMD's encapsulation of the many aspects defining visual event understanding, typically studied in isolation, enables an unprecedented, unified account of the complex mediating neural mechanisms.

Results

Sampling brain activity for 1,102 distinct visual events

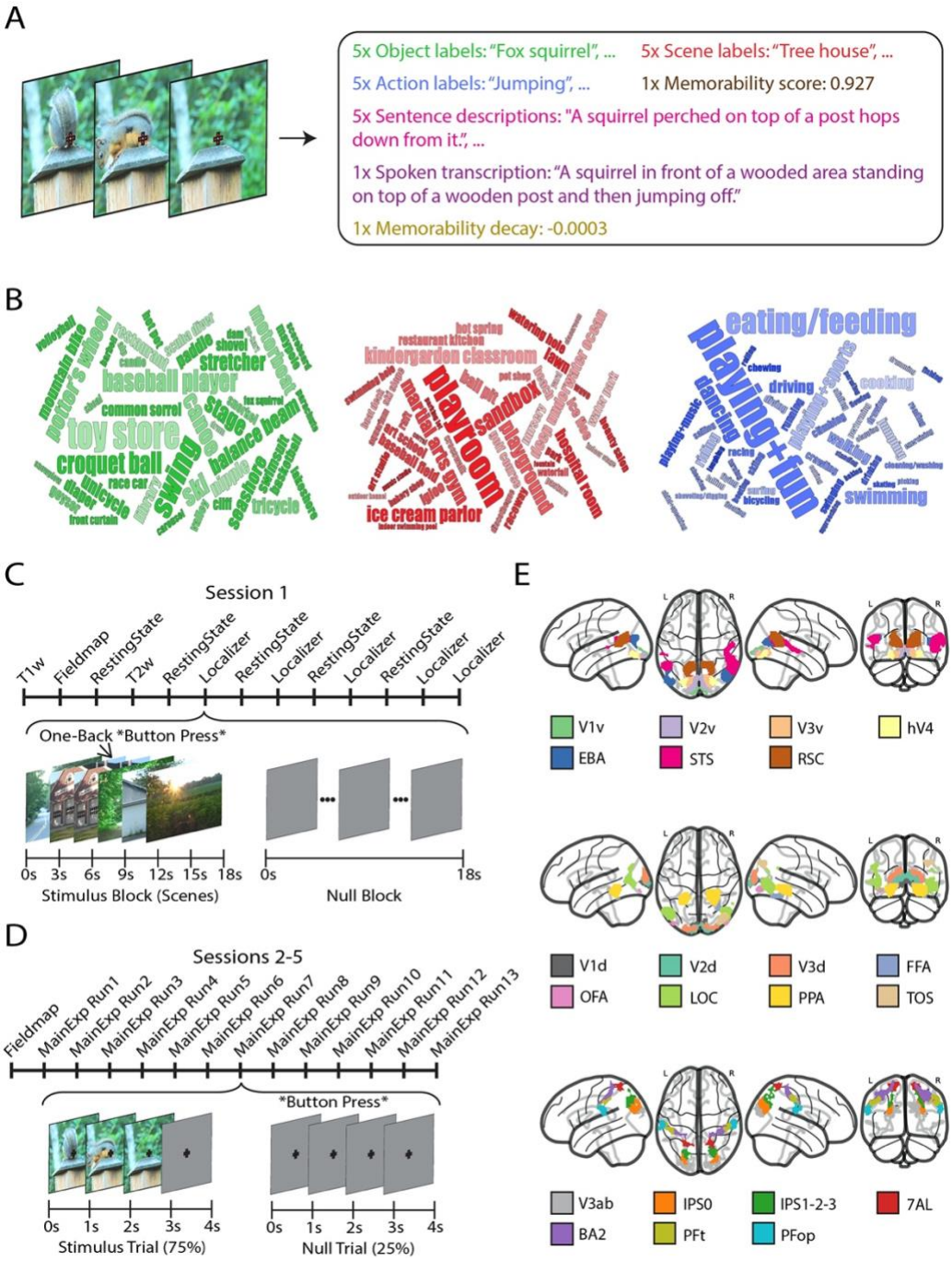

Figure 1: Experimental design and data acquisition.

(A) Stimuli and metadata: The 1,102 3-second video stimuli are sampled from the Moments in Time dataset and annotated with semantic and behavioral attributes: 5 object labels, 5 scene labels, 5 action labels, 5 sentence descriptions, 1 spoken transcription, 1 memorability score, and 1 memorability decay rate. **(B) Object, scene, and action label word clouds:** A qualitative visualization of the top 50 most frequent object (left, green), scene (middle, red), and action (right, blue) labels. A larger word corresponds to higher frequency of occurrence. **(C) fMRI session 1:** Subjects underwent T1- and T2-weighted structural runs interspersed with resting state and functional localizer functional runs to define category selective regions. **(D) fMRI sessions 2-5:** Sessions were identical in design and consisted of a fieldmap run followed by thirteen runs (ten training runs and three testing runs) in random order. The 3-second videos were presented on a gray background at a five degree visual angle overlaid with a red fixation cross. Stimuli presentation was followed by a 1-second intertrial interval composed of a red fixation cross on a gray background. Participants reported luminance changes of the fixation cross occurring irregularly between videos with a button press and did so reliably (hit rate 0.964 +/- 0.014 (mean ± SD)). **(E) Region of Interest Definitions.** Regions of interest for a representative subject (subject 1), functionally (from session 1) and anatomically defined.

We sampled 1,102 naturalistic 3-second videos depicting diverse dynamic visual events from the Moments in Time dataset (Monfort et al., 2020, 2022; Newman et al., 2020), a 1 million video action-recognition dataset frequently used in the computer vision community to benchmark video understanding models. The 3-second video duration, approximately the duration of human working memory (Baddeley, 1992; Barrouillet et al., 2004; Schneider, 2013), is an ideal length to capture brain responses to basic events (e.g. opening a door). Videos shorter than 3 seconds risk capturing incomplete actions (e.g. extending the arm) while videos longer than 3 seconds risk unwanted complexity by capturing ambiguous, composite actions (e.g. leaving the house). The naturalistic nature of the videos, as opposed to images, simple moving shapes, or synthetic cartoons, ensures that we drive and capture the full extent of neural activation responsible for visual event understanding (Buccino et al., 2004; Kret et al., 2011; Schultz & Pilz, 2009; Sonkusare et al., 2019).

The experimental design weighted benefits of both a large stimulus set and several stimulus repetitions to enable a variety of downstream analyses that depend more strongly on one or the other. We thus divided the 1,102 stimuli into two non-overlapping sets, one consisting of 1,000 stimuli with three repetitions per subject and the other consisting of 102 stimuli with ten repetitions per subject. We use terminology common to encoding studies, naming the former set

the training set and the latter set the testing set. This amounts to 4,020 unique fMRI trials per participant, and 40,200 unique fMRI trials across the entire dataset. The large number of trials and participants enable direct comparisons of results across participants and invite potential integration of data across participants (Hasson et al., 2004; Hasson, Furman, et al., 2008; Haxby et al., 2011, 2020).

Brain responses were recorded with whole-brain 3T fMRI at 2.5mm iso-voxel resolution to densely sample the widely-distributed cortical responses to video across the whole cortex (Buccino et al., 2004; Gazzola & Keysers, 2009; Kret et al., 2011; Rizzolatti & Sinigaglia, 2010; Yildirim et al., 2019). The fMRI experiment was split into 5 sessions. Session 1 (Figure 1C) contained crucial auxiliary brain measurements, interleaving high-resolution T1- and T2-weighted structural, video-based functional localizer (Lafer-Sousa et al., 2016), and resting state scans (Hutchison et al., 2013; Smith et al., 2013). Sessions 2 to 5 (Figure 1D) contained the main experiment and had identical structure. Each video trial was 4s long, consisting of a 3s silent video presentation followed by a 1s intertrial interval.

Semantic and behavioral metadata on visual events

Revealing how the brain mediates visual event understanding requires detailed descriptions of the different components in a visual event. We facilitate this research endeavor by providing an extensive set of stimuli metadata from crowd-sourced experiments and deep learning tools (Figure 1AB).

We provide the top-5 object and scene label predictions from deep neural networks trained on object and scene recognition (i.e., a ResNet50 architecture trained on ImageNet and Places respectively, Deng et al., 2009; He et al., 2016; Zhou et al., 2018) to annotate each event's basic spatial components. We next label each visual event's relational information with 5 action labels, 5 sentence text descriptions, and 1 spoken transcription through crowd-sourced human experiments (Monfort et al., 2021). Action labels (e.g., "opening") describe a core temporal aspect of the video. Sentence text descriptions (e.g. "A man opens a door") detail how only the most pertinent spatial parts of the event interact over time (13.06 mean +/- 2.800 std words per sentence). Spoken transcriptions tend to be more verbose with additional emotional and linguistic subtleties present in speech but often not in text (e.g. "The man opened the door pretty fast, he seemed anguished or angry") (26.72 mean +/- 17.55 std words per transcription).

Finally, to determine how visual event understanding interfaces with memory, we use a crowd-sourced memory game to behaviorally measure if participants recognize a visual event at a later time period (memorability: 0.8422 mean +/- 0.0888 std) and how this recognition performance fades over time (memorability decay rate: -0.0014 mean +/- 0.0011 std) (Newman et al., 2020).

This metadata allows researchers to immediately use any possible grouping, subdivision, or other transforms of these measures to investigate specific cognitive processes underlying visual event understanding. We provide first example analyses on this basis below.

(f)MRI data processing, response modeling, and ROI definition

We provide raw as well as preprocessed versions of the MRI data in the community-backed and standardized BIDS format (K. J. Gorgolewski et al., 2016) to ensure transparent quality assessment, reproducible preprocessing, and easy-to-share results. The raw data gives researchers complete control over preprocessing to pursue research questions at any stage of the analysis pipeline. The preprocessed data allows for immediate analyses into the spatial and temporal neural dynamics underlying visual event understanding, at both the group or single-participant level.

We processed the main fMRI experimental data in three main steps. We first processed the data using the standardized fMRIPrep tool (Esteban et al., 2019) to achieve reproducible and transparent results. We then temporally resampled the data from a TR of 1.75s to a TR of 1s, achieving a densely sampled time course time-locked to stimulus presentation (stimulus presentation happens every 4s, i.e. a multiple of 1s, but not 1.75s) (Kay et al., 2020). Finally, to accommodate variability in the hemodynamic responses at different locations in cortex, to resolve temporal structure of the stimuli, and to address response overlap from the rapid event-related design, we used Finite Impulse Response (FIR) basis functions. We modeled the hemodynamic response to visual events using data from 1-9s after stimulus onset (to account for the hemodynamic lag) in 1s steps (i.e. 9 bins of 1s length each) for each trial separately.

To guide analysis in a region-specific manner, we used the auxiliary functional localizer data from session 1 to define a set of 22 regions of interest (ROIs) previously reported to be involved in visual perception of natural images, natural video, or motion (Figure 1E) (Gazzola & Keysers, 2009; Le et al., 2017; Logothetis & Sheinberg, 1996; Rizzolatti & Sinigaglia, 2010; Silver &

Kastner, 2009; VanRullen & Thorpe, 2001). The set includes early visual, category-selective ventral visual, dorsal visual, and parietal regions.

All together, this preprocessing suite ensures a low threshold for researchers to interact with the brain data at their desired processing level.

(f)MRI image scans of high quality across subjects and task

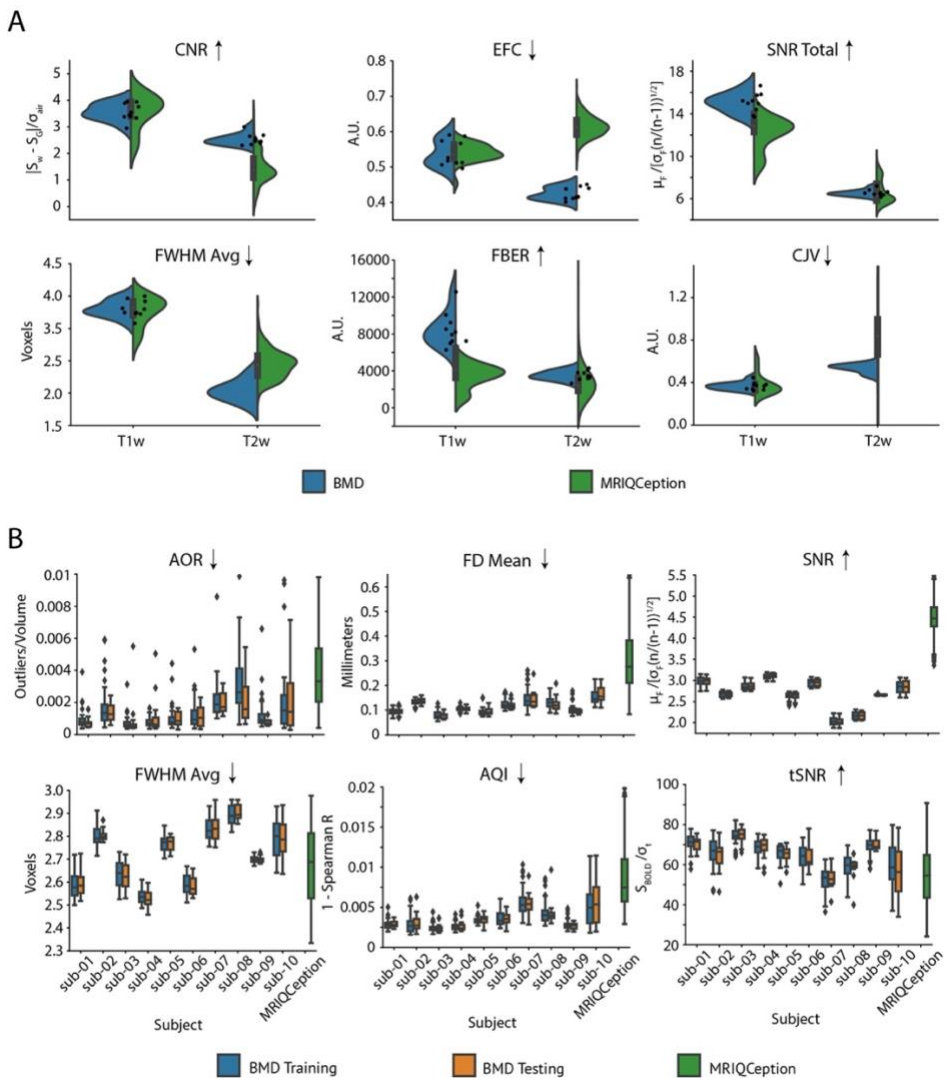

Figure 2: Preprocessed scan quality measures

A. T1- and T2-weighted structural scan data quality: Violin plots compare Image Quality Metrics (IQMs) of the T1- and T2-weighted structural scans from our BMD dataset (T1-weighted, n=10; T2-

weighted, n=10) with a collection of anonymous data pulled from the MRIQCeption API (T1-weighted, n=219; T2-weighted, n=695). We report the Signal to Noise Ratio (SNR Total), Contrast to Noise Ratio (CNR), Coefficient of Joint Variation (CJV), Entropy Focus Criterion (EFC), Average Full-Width Half Maximum Smoothness (FWHM Avg), and Foreground-Background Energy Ration (FBER) as a summary of the structural scan quality. Points correspond to the data value from an individual subject in the BMD dataset. An up arrow after the IQM title means a higher value corresponds to better data quality, while a down arrow after the IQM title means a lower value corresponds to better data quality. A.U. stands for "arbitrary units." **B. Training and testing functional scans data quality:** Panels compare boxplots of IQM values for each subject for the training and testing functional tasks in the BMD dataset (per subject: training, n=40; testing, n=12) with anonymous BOLD data pulled from the MRIQCeption API (BOLD, n=687). We report the Signal to Noise Ratio (SNR), Temporal Signal to Noise Ratio (tSNR), Mean Framewise Displacement (FD Mean), Average Full-Width Half Maximum Smoothness (FWHM Avg), AFNI Outlier Ratio (AOR), and AFNI Quality Index (AQI) to summarize the functional scan quality. The boxplot extends 1.5 times the high and low quartiles, with outliers defined as a scan with a value outside that range and denoted by diamonds. An up arrow after the IQM title means a higher value corresponds to better data quality, while a down arrow after the IQM title means a lower value corresponds to better data quality.

To assess the quality of the raw and preprocessed MRI data, we used the open-source and community-based MRIQC analysis package (Esteban et al., 2017). This yielded a comprehensive set of 44 functional and 68 structural quality metrics (full report available), of which we present a representative set of six structural (Figure 2A) and six functional MRI metrics (Figure 2B) (see Supplementary Figure S1 for the resting state and functional localizer scans). Since most quality metrics do not have a ground-truth reference value to compare against, we contextualize our values against the values of hundreds of anonymized studies with similar scanner parameters (1 < Tesla < 3, 1 <= TR < 3; aggregated with the MRIQCeption API).

Taken together, the IQMs indicate BMD's structural and functional scans are of excellent quality. The distribution of all but one IQM for each subject falls within or is noticeably better than the distribution seen in similar fMRI studies (Figure 2, green). The strong results of the IQMs tSNR (a measure of SNR over time), aor (an indicator of the number of outliers per fMRI volume), and aqi (a correlational measure of quality per volume) assure satisfactory functional SNR quality in light of the below-typical per volume SNR IQM. We highlight that within each participant, the range of quality metric values was especially consistent between the training and testing sets (Figure 2, blue and orange boxplots). This shows that the training and testing

sets are of comparable data quality, facilitating analyses that depend on this split. Further, none of the 10 participants were outliers as indicated by consistently lower within- than between-participant variability, encouraging group result inference.

Reliable univariate and multivariate fMRI response profiles

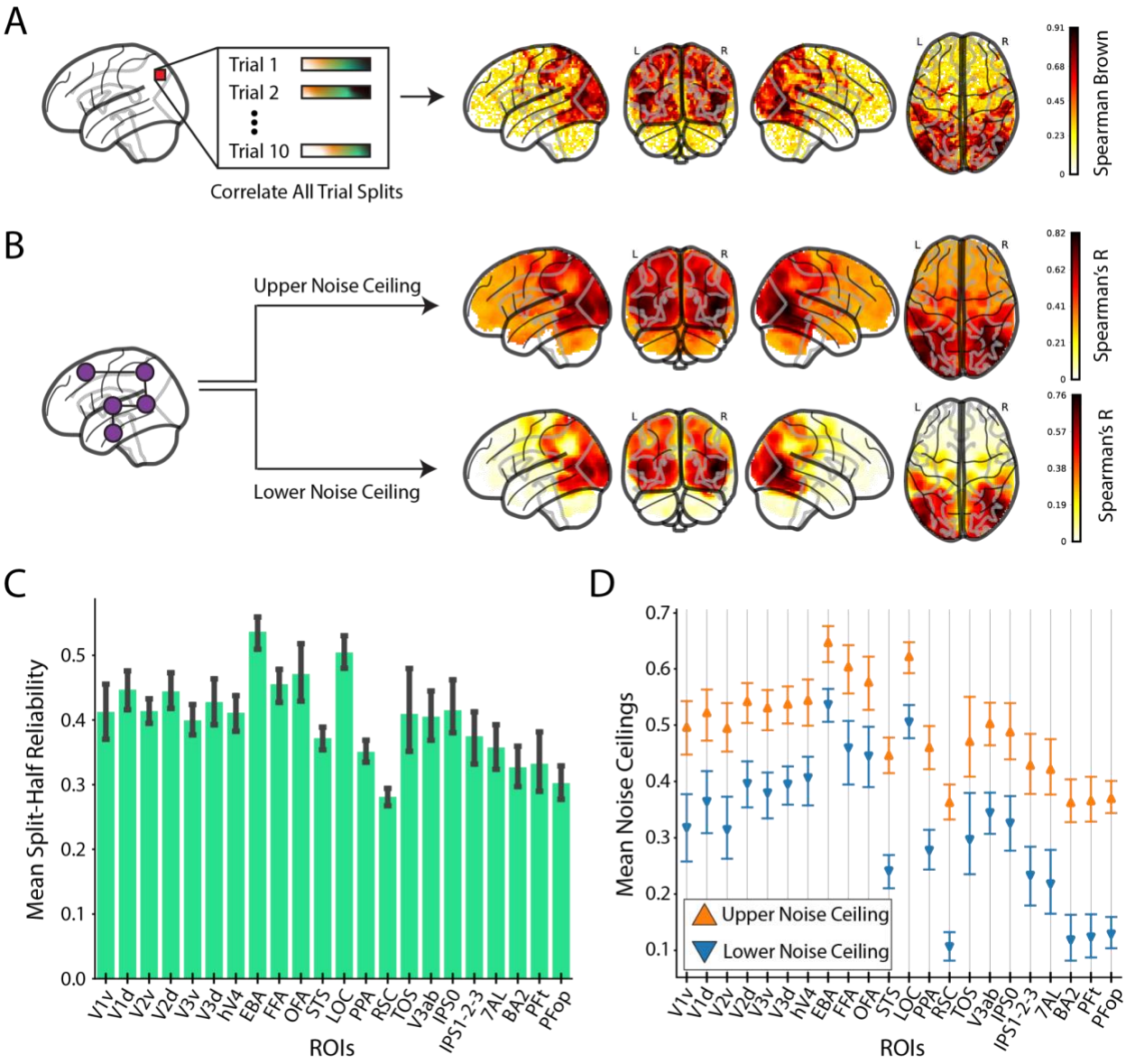

Figure 3: Whole-brain and ROI reliability after response modeling

(A) Whole-brain single-subject split-half reliability analysis: We perform a voxelwise split-half reliability analysis and present the voxels that pass the reliability criteria ($p < 0.05$, Spearman-Brown

$(2\rho/(1+\rho)))$ for a representative subject (subject 1). **(B) Whole brain searchlight noise-ceiling analysis:** We estimate the upper and lower noise-ceilings across the whole brain for a representative subject (subject 1) from a searchlight representational dissimilarity analysis (RSA). **C) ROI-based group split-half reliability:** For each of the 22 ROIs, we present the mean split-half reliability across participants. **(D) ROI-based group searchlight noise ceilings:** For each of the 22 ROIs, we calculate the mean upper (orange) and lower (blue) noise ceilings across participants. All error bars indicate the 95% confidence interval (n=10 for all ROIs except TOS (n=8) RSC (n=9)). The brain responses used for the reliability analyses are the beta values averaged over TRs 5-9 (the peak of the BOLD signal) from the testing set.

We provide both univariate and multivariate reliability measures to evaluate the reliability in the context of the two main fMRI analysis traditions today: the univariate framework that focuses on local information at a single voxel scale (Friston et al., 1994; Khosla et al., 2022; Naselaris et al., 2011; Ratan Murty et al., 2021; Schrimpf, Kubilius, Hong, et al., 2020), and the multivariate analysis framework that emphasizes the distributed nature of information in population codes (Haxby, 2012; Haynes, 2015; Kriegeskorte, 2008).

To assess univariate reliability, we identify voxels with a Spearman-Brown split-half reliability value (split across the 10 testing set stimuli trials) that satisfy a reliability criterion of $p<0.05$ (assessed by stimulus label permutation).

To assess multivariate reliability, we perform representational similarity analysis (RSA) in a searchlight approach (Kriegeskorte et al., 2006) to determine the upper (subject-to-group RDM correlation per voxel) and lower (leave-one-out RDM correlation per voxel) estimate of the noise ceilings. We report the univariate and multivariate reliability results across the whole brain (Figure 3AB for a representative subject, see Supplementary Figures S2 and S3 for all subjects) and in ROIs (Figure 3CD).

In both the whole-brain univariate and multivariate reliability analyses, we observe statistically significant reliability values across the occipital and parietal cortex, even extending into the frontal lobe. The ROI analyses yield equivalent results, showing high explainable variance in a functionally diverse set of ROIs responsible for visual event understanding. This demonstrates that BMD is well suited for comprehensive and advanced analysis at both the single- and multi-voxel spatial scale.

Modeling visual event understanding with a video-computable deep neural network for action recognition

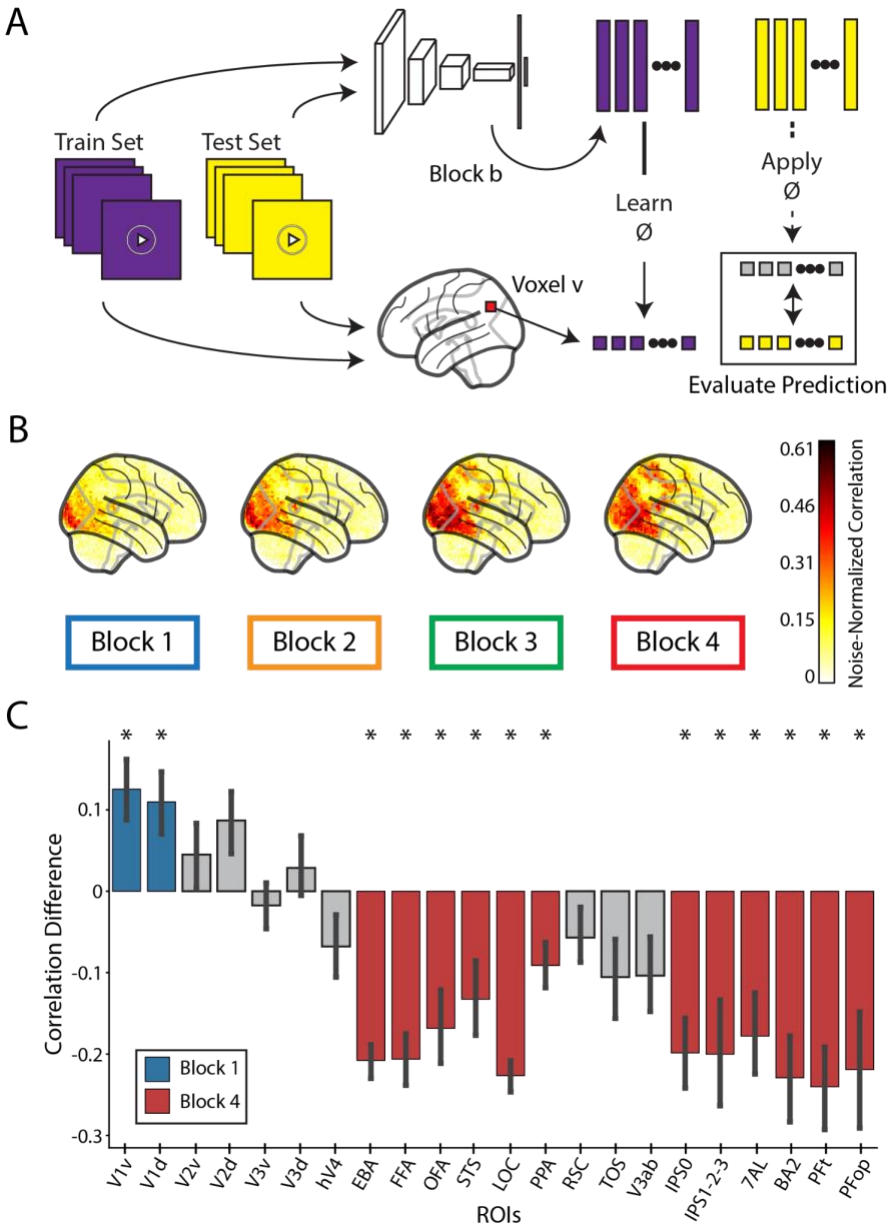

Figure 4: Evaluation of biologically-similar video-based encoding model

(A) Voxelwise encoding model procedure: All videos are shown to both a DNN and a human. Training set video embeddings are extracted from a block *b* of the DNN and used to learn a voxelwise mapping function to the human responses. This mapping is then applied to the testing set video embeddings to predict the brain response at each voxel. **(B) Whole-brain encoding accuracy across blocks:** We use the encoding model procedure with each of the four blocks of a TSM ResNet50 model trained to

recognize actions in videos to predict the neural response at each voxel in the whole brain. The brain figures show the subject-average noise-normalized predictive correlation (divided by the voxel's upper noise ceiling) at each voxel. **(C) ROI-based encoding accuracy difference:** Mean difference in predictive performance between block 1 and block 4 at each of the 22 ROIs. Predictive performance at each voxel is measured as the noise-normalized correlation between the brain responses and the predicted responses, averaged over all reliable voxels in each ROI. Error bars show the 95% confidence interval with data from n=10 subjects except for TOS (n=8, no responses from subjects 6 and 7) and RSC (n=9, no responses from subject 7). Significant ROIs are denoted with an asterisk and a color (blue for Block 1, red for Block 4, gray is not significant) corresponding to the significant layer (one sample two-sided t-test against a population mean of 0, Bonferroni corrected across 22 ROIs, $p < 0.05$).

A major goal in explaining cognition is providing explicit quantitative models that predict the underlying computations and their cortical organizations (Dayan & Abbott, 2001). Deep Neural Network (DNN)-based modeling has emerged as the currently dominant form of scientific modeling in visual neuroscience, due to their image-computable design, biologically-inspired architecture, and high neural prediction performance (Allen et al., 2022; Chang et al., 2019; Rajalingham et al., 2018; Schrimpf, Kubilius, Hong, et al., 2020; Schrimpf, Kubilius, Lee, et al., 2020; D. L. Yamins et al., 2013).

However, modeling visual events has been limited by the lack of a suitable dataset that accounts for complex distributed processes across the whole brain (Yildirim et al., 2019), drastic differences to image understanding (Buccino et al., 2004; Krekelberg et al., 2003; Kret et al., 2011; Schultz & Pilz, 2009; Senior et al., 2000; Shirai & Imura, 2014), and the temporal boundaries of a visual event (Aliko et al., 2020; Hasson, Yang, et al., 2008; Nishimoto et al., 2011; Seeliger et al., 2019). BMD breaks this impasse with its large number of short video stimuli and whole-brain responses.

Towards the goal of modeling visual events, we used a video-computable DNN following biological constraints. The DNN uses a recurrent ResNet50 backbone (He et al., 2016) that mimics the biological recurrent computations essential for human motion perception and categorization (Kietzmann et al., 2019; Koivisto et al., 2011; Pascual-Leone & Walsh, 2001; Silvanto, Cowey, et al., 2005; Silvanto, Lavie, et al., 2005) and builds on the ResNet family's strong neural predictivity performance seen for still images (Schrimpf, Kubilius, Lee, et al., 2020). The ResNet50's four recurrent blocks are connected using a Temporal Shift Network (Lin et al., 2019) designed to process video input in the natural, uni-directional temporal order. We

train the model on an action recognition task using the same dataset from which the BMD stimuli were sampled, the Moments in Time dataset (Monfort et al., 2020, 2022) (BMD stimuli were excluded from model training). We release this model to aid investigations of visual event understanding.

We queried the relationship between each of the model's blocks and BMD. Using a voxelwise encoding model approach (Naselaris et al., 2011) (Figure 6A), we observe a correspondence between DNN block depth and predictivity performance along the visual processing hierarchy and beyond. The predictivity of each of the DNN's four layers progressively spreads across the brain from posterior to anterior (Figure 6B). Focusing on specific ROIs and DNN Block 1 vs. 4, we observe that predictivity of DNN Block 4 becomes significantly greater than DNN Block 1 beginning in the early visual cortex and extending notably into dorsal visual cortex and parietal cortex (Figure 6C) (see Supplementary Figure S7 for results on all layers and ROIs).

BMD, with its whole-brain human fMRI responses to short events, breaks a video-based computational modeling impasse by allowing rigorous analyses in all of cortex, especially in the dorsal and parietal regions largely driven by dynamic stimuli. We are thus able to clarify previously conflicting results (Bakhtiari et al., 2021; Güçlü & van Gerven, 2017; Mineault et al., 2021), showing that a DNN trained on an action recognition task can accurately predict responses in the dorsal visual stream and even into the parietal cortex. This extends previous research demonstrating a hierarchical correspondence between DNNs and brains from still image stimuli (Cichy et al., 2016; Kriegeskorte, 2015; Kubilius et al., 2019; D. L. Yamins et al., 2013; D. L. K. Yamins et al., 2014) to dynamic video stimuli.

fMRI responses capture natural temporal event structure

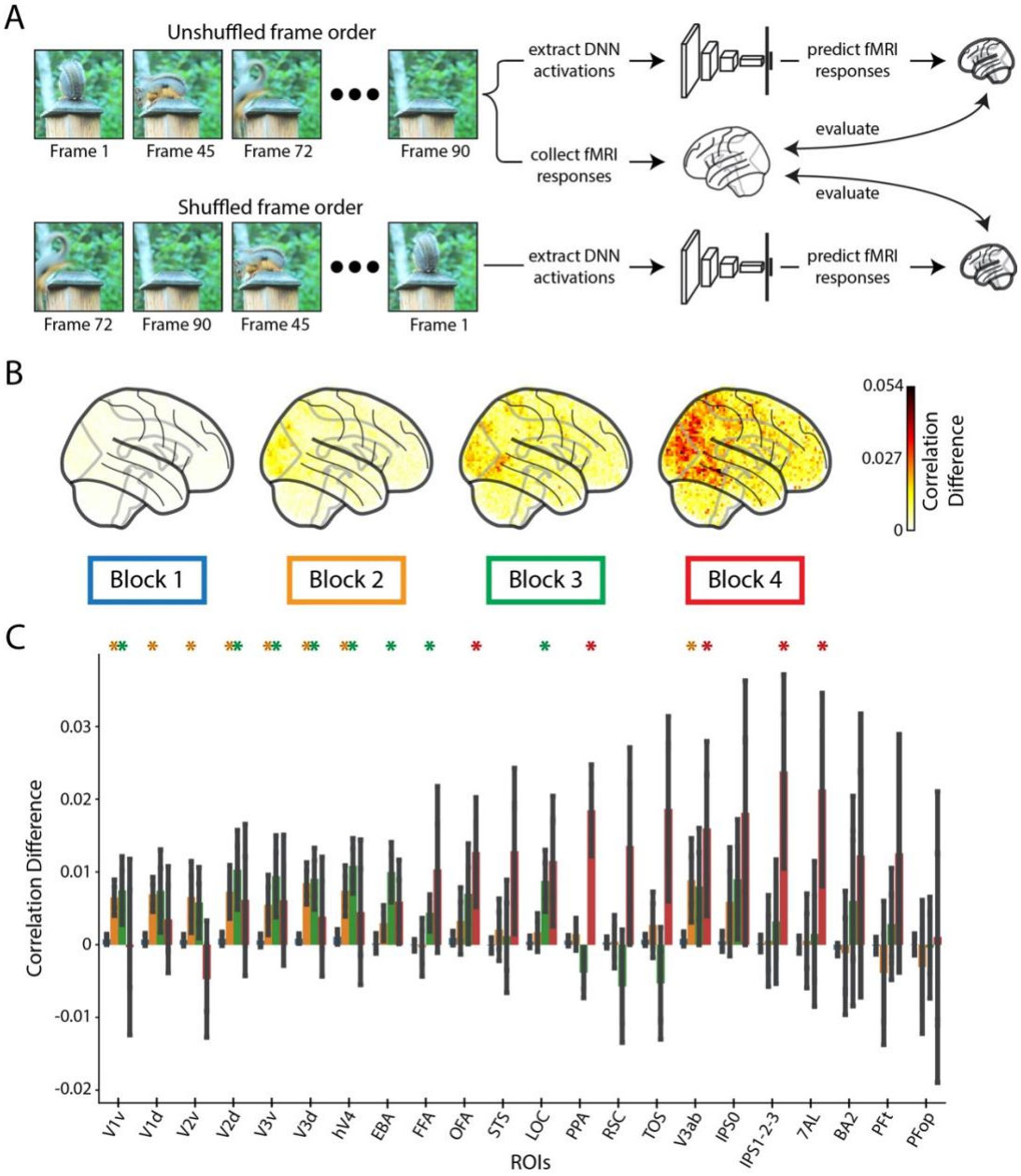

Figure 5: Importance of temporal order in predicting fMRI responses.

A) **Frame shuffling procedure:** We predict the real fMRI responses using both the DNN activations of the original (unshuffled) frame order and the DNN activations of the randomly shuffled frame order. A difference in prediction accuracy between the DNN activations of the unshuffled and shuffled frames indicates the preservation of correct temporal order in the fMRI response. **B**) **Whole-brain prediction difference:** Difference in the correlation averaged over participants between the shuffled frame prediction

accuracy and unshuffled frame prediction accuracy across the whole brain at different DNN layers (TSM model). **C**) **ROI-based prediction difference:** Difference in the correlation between the shuffled frame prediction accuracy and unshuffled frame prediction accuracy at different ROIs and DNN layers (TSM model). A colored asterisk above a bar indicates significant difference between the unshuffled and shuffled prediction accuracy at that DNN block (one sample two-sided t-test against a population mean of 0, Bonferroni corrected across 22 ROIs, $p < 0.05$). Error bars depict the 95% confidence interval.

A given visual event unfolds in a systematic, spatiotemporal sequence. Is the temporal structure of an event important for its cortical representation? Shuffling the frames of a video effectively destroys any meaningful temporal structure. We reasoned that if a voxel captured an event's spatiotemporal relationships, unshuffled (meaningfully ordered) video input would correspond better to the brain responses than shuffled video input. The shuffled and unshuffled video inputs are both temporally dynamic and contain identical frame-averaged spatial content, thereby isolating the effects of the encoding of ordered temporal content.

To investigate, we measured the neural prediction performance of the DNN model introduced above when given unshuffled video input and shuffled video input (Figure 5A). Since this DNN is engineered to process video input in a unidirectional temporal order, we know that shuffling the video frames will affect the DNN activation in at least one of the four blocks. By using the shuffled and unshuffled activations from these blocks to predict the brain responses, we can assess the effect of correct temporal ordering on our BMD brain responses.

Our results show the frame-shuffled input decreased DNN prediction accuracy across most of the visual system. The whole-brain analysis (Figure 5B) shows that this decrease in DNN prediction accuracy is most pronounced in both magnitude and coverage with increasing DNN blocks. The ROI analysis (Figure 5C) reveals that early visual ROIs were primarily sensitive to differences in earlier DNN blocks 2 and 3, while ventral and dorsal ROIs were most affected by differences in DNN block 4. These results suggest functionally-specific sensitivities to shuffled input along the visual system in line with known feature preferences in the visual system and DNN computations (Honey et al., 2012; Kubilius et al., 2019; Logothetis & Sheinberg, 1996).

Together this demonstrates that most visual regions captured meaningful temporal structure of the video stimuli, providing the necessary background for investigations into the processing of high-level dynamic visual concepts, such as event categories and actions. The use of our video-

computable DNN model allows the extraction of feature spaces at specific stages of video processing, inviting precise inquiry into an ROI's function in visual event understanding.

BMD tracks the temporal dynamics unfolding within events

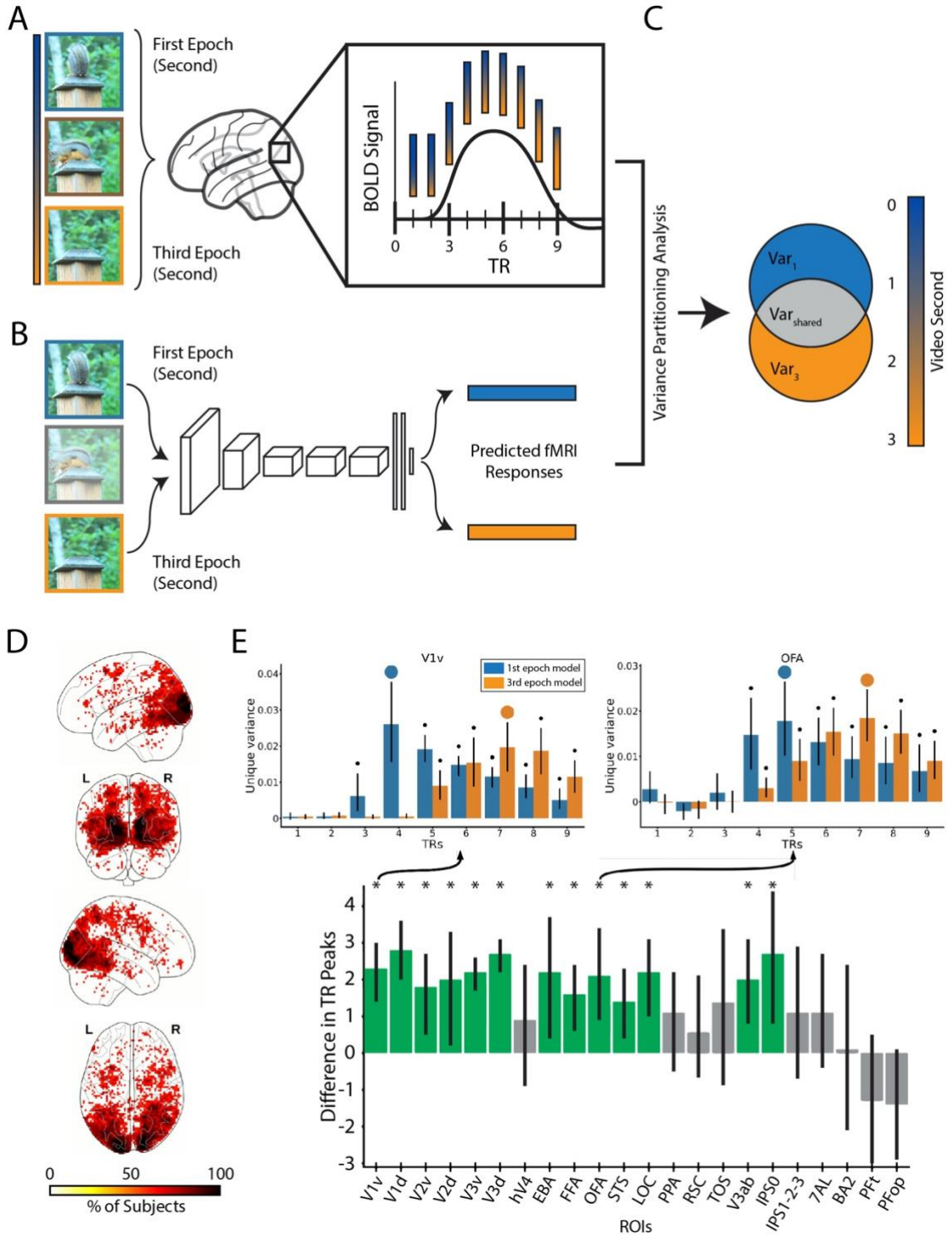

Figure 6: Encoding the temporal dynamics of the BOLD signal.

(A) TR estimated fMRI responses: We estimate the video-evoked brain response (beta values) of the first 9 TRs of the BOLD signal. **(B) DNN predicted fMRI responses:** We use an encoding model procedure to extract two sets of DNN activations, one from the first video second (first epoch) and the other from the third video second (third epoch). We then predict fMRI responses using the two sets. **(C) Variance partitioning analysis:** We calculate the unique variance explained by the first and third video epochs' predicted fMRI responses at each TR. **(D) Whole brain analysis:** Each voxel shows the percentage of subjects with a TR peak difference of 1 to 3 TRs at that specific voxel. Only significant voxels are plotted ($p < 0.05$, binomial test, FDR corrected). **(E) ROI analysis:** Upper ROI panels: We show the unique variance explained by the predicted fMRI responses of the first (blue) and third (orange) video epoch at TRs 1-9 in representative ROIs V1v and OFA. Black asterisks indicate significant unique variance greater than 0. Blue/orange circles indicate the significant peak TR of the first/third video epoch. Main ROI Panel: We depict the difference in first and third video epoch TR peaks at each ROI, averaged across subjects. Green bars with asterisks indicate TR peak differences significantly greater than 0. Error bars in both the upper and main plots reflect 95% confidence intervals.

Does the temporally sluggish BOLD signal capture any temporal information within events? One possibility is that the BOLD signal only captures a global representation of the event that has no temporal structure itself, akin to a time-less semantic label of "a person opens a door". Another possibility is that instead the BOLD signal captures delayed but temporally-resolved information, where different time points of the BOLD signal capture local snapshots of the changing event (Figure 6A).

To test the latter hypothesis, we extracted two sets of activations from a DNN, one set using only the first second (first epoch) of the videos and the other set using only the last second (third epoch) of the videos (Figure 6B). We utilized an encoding model procedure to measure the two sets' predictive performance at each of the first 9 seconds (corresponding to TRs in acquisition) (Figure 6A). This was done with a feed-forward DNN trained on object categorization from still images to avoid any confounds from temporal integration. We expected that if a voxel captures snapshots of the changing event, the best prediction accuracy of the first video epoch encoding model is at least one TR (one TR = 1s) earlier than the best prediction accuracy of the third video epoch encoding model. We used variance partitioning to identify the unique contribution of the first and third video epochs' predictions to the real fMRI responses (Figure 6C).

A whole-brain voxel-wise analysis (Figure 6D) revealed the percentage of subjects at each voxel that show a significant 1-3s delay between the best predicted time point (TR) using a video's first epoch and using a video's third epoch (only significant voxels are plotted). Results highlight significant temporal delays throughout the ventral and dorsal cortex, but most pronounced in the early visual cortex. An equivalent ROI-based analysis (Figure 6E) yielded a similar result pattern. ROIs in the visual brain (14 of the 22 total), mostly in the early visual and ventral stream, showed a significant timing difference (black asterisks) between the time points at which fMRI responses are most related to the contents of the first and the third epoch of video (Figure 6E, main bottom panel).

Together, these results support the hypothesis that early and late TRs of the BOLD signal better code early and late video snapshots, respectively. BMD captures the temporal dynamics of events at the level of seconds, with the most pronounced effects focused in the early visual cortex (Fairhall et al., 2014; Kiebel et al., 2008). This invites future research to use BMD's temporally well-defined stimuli to explore how visual event information is integrated over shorter time periods, bridging an important gap to temporal integration studies of longform movies.

Semantic metadata reveal a preference for sentence-level descriptions in the visual system

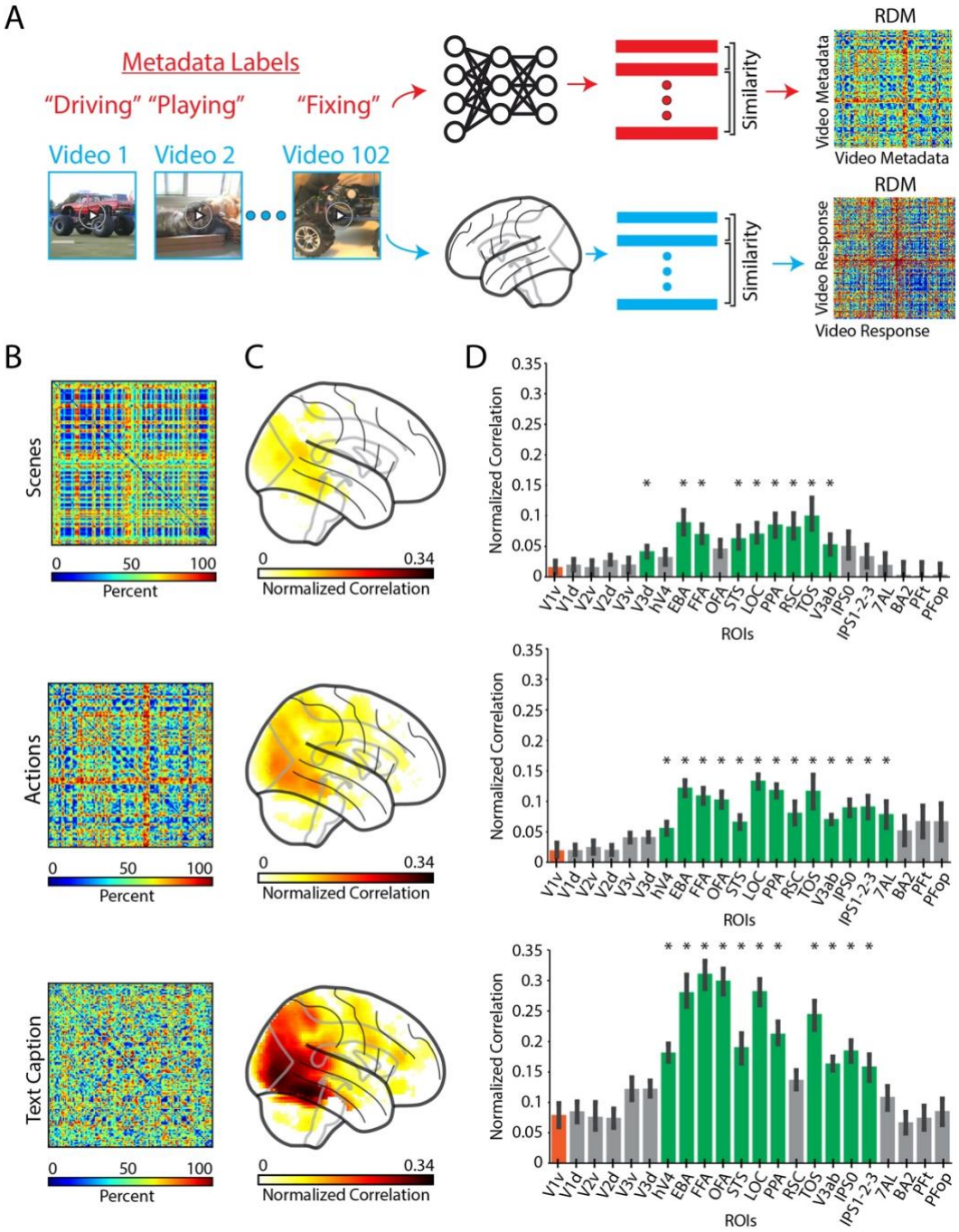

Figure 7: Metadata-driven analysis of the semantics of visual event encoding.

(A) RSA with metadata methodology: Video metadata are fed into a language model to produce a vector embedding. Similarly, the video-evoked brain responses are extracted within a spherical searchlight to produce a vector embedding at each voxel. We separately calculate the pairwise similarity of the metadata embeddings and video-evoked brain embeddings to create a representational dissimilarity matrix (RDM) of size n_stimuli x n_stimuli. The metadata RDM is then correlated with the searchlight-based voxelwise RDMs. **(B) Metadata RDMs:** The "scene" and "actions" metadata were fed into a FastText language model, and the "text caption" metadata was fed into a Sentence-BERT model to produce vector embeddings. The RDMs were computed using cosine distance, and the RDMs are visualized here with rank-normalized values (0-100% of maximum value rank) **(C) Whole-brain correlation of metadata RDMs with searchlight-based RDMs:** We correlate (Spearman's R) the metadata RDM with each searchlight-based RDM at each voxel in each subject, performed statistical analysis (one-sample two-sided t-test against a null correlation of zero, FDR correction with q=0.05), divided each voxel by the subject's upper noise ceiling, and averaged across subjects. Only significant voxels are shown. **(D) ROI-based correlation:** We show the mean of the noise-normalized correlations within each subject's ROI. Significance of the ROIs was determined by comparing each ROI's mean noise-normalized correlation against "V1v"'s (orange) mean noise-normalized correlation (paired t-test, two-sided, Bonferroni corrected across 21 ROIs, $p < 0.05$). Asterisks and green bars denote ROI significance. Error bars represent the 95% confidence interval with data from n=10 subjects except for TOS (n=8, no responses from subjects 6 and 7) and RSC (n=9, no responses from subject 7).

Words are often used to operationalize the aspects of visual stimulus typically most useful to humans. For example, one might label objects as "dog" and "ball", describe their relationship as "playing fetch," yet refrain from identifying contextually unimportant features like a "park bench." Here we leverage our semantic metadata of scene, action, and sentence labels to explore how spatial, temporal, and relational stimuli features, respectively, are encoded in the brain.

To quantify text-based data for numerical analyses, we run the text through AI language models (FastText, Bojanowski et al., 2017, for the single-word scene and action labels, and Sentence-BERT, Reimers & Gurevych, 2019, for sentence-level text descriptions), obtaining model embeddings (Figure 7A). We then use representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to relate the model embeddings (Figure 7B) to brain activity within a common representational space.

Whole-brain (Figure 7C) and ROI results (Figure 7D) show that the correlations of the three metadata types are primarily significant in the ventral and dorsal visual regions (green bars are significant, Figure 7D), but not in most of the early visual or parietal regions (Gray bars are not

significant, Figure 7D). We note that even the correlational significance of the action metadata failed to extend deep into the parietal cortex (with the exception of 7AL), despite the parietal cortex's heavy involvement in action observation (for a review, see Hardwick et al., 2018). Additionally, one might suspect that scene-selective regions (PPA, RSC, and TOS) would best correlate with the scene metadata, reasoning that sentence labels, while including scene information, contain mostly irrelevant and distracting non-scene content. Yet, the sentence labels exhibit correlation values nearly twice that of the scene metadata in PPA and TOS. This trend continues with action labels in the ventral and dorsal visual stream, consistent with the idea that complex scene analysis, rather than simpler tasks such as object recognition, is the objective of the visual brain (Doerig et al., 2022).

BMD's potential to reveal visuo-semantic representations throughout cortex promises to better inform theories of the visual system's objective and action encoding. This methodology can easily be extended to externally collected metadata to interrogate any desired facet of visual event understanding.

Video memorability is reflected in high-level visual and parietal cortices

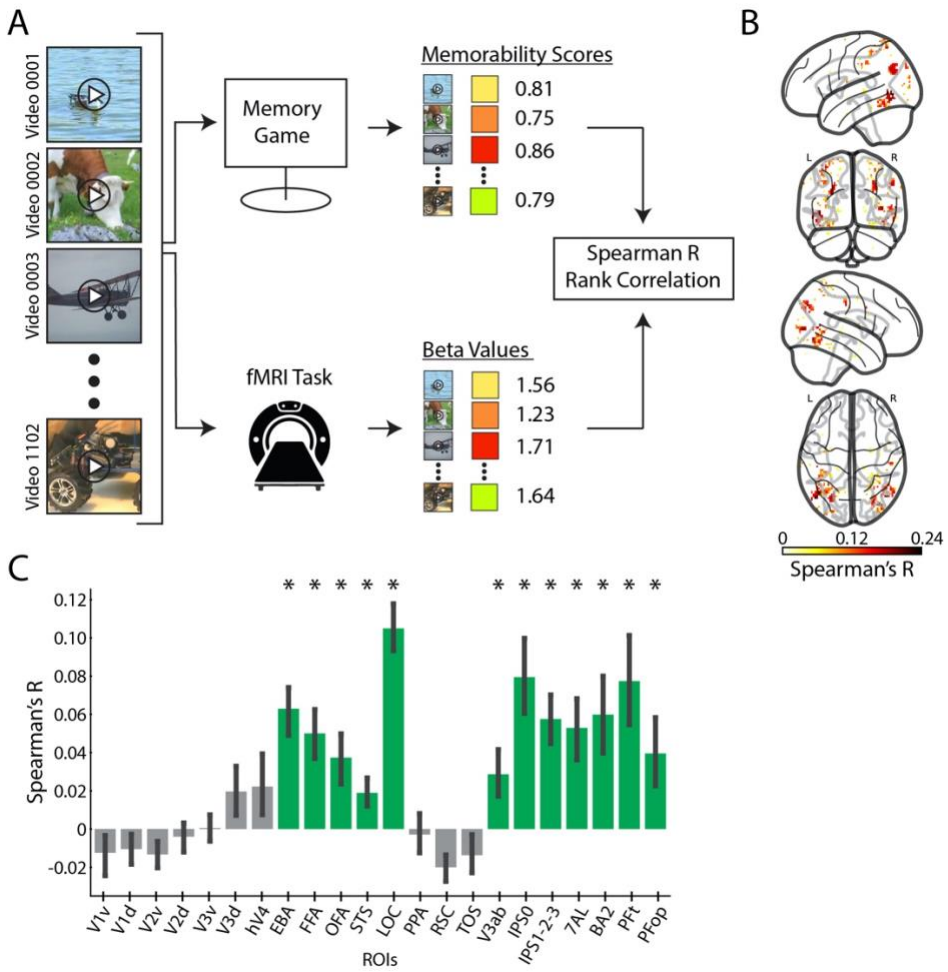

Figure 8: Correlates of video memorability in cortex

(A) Memorability analysis procedure: For each of the 1,102 video stimuli, we collect a memorability score (values between 0-1, collected via a behavioral memory game in an independent study by Newman et al., 2020) and brain response (beta value averaged across repetitions). We rank and correlate (Spearman's R) a vector of the stimulus memorability scores with a vector of the stimulus-evoked beta values at each voxel to obtain a correlation coefficient and corresponding p-value. **(B) Whole-brain memorability correlation results:** We show significant whole brain voxel-wise correlations (Spearman's R) between brain activity and memorability scores. Significant voxels are determined by a t-test across subjects (one-sample, one-sided) and FDR correction (q=0.05, positive correlation assumption). **(C) ROI memorability correlation results:** We show the mean Spearman correlation value between brain activity and memorability scores for each ROI. Significance is determined by a t-test against a null population

mean correlation of 0 (one sample, one-sided, Bonferroni corrected across 22 ROIs, $p < 0.05$). Significant ROIs are denoted with a green bar and black asterisk. Error bars depict a 95% confidence interval.

Of the many video advertisements, social media clips, and ordinary visual events humans perceive each day, some are predisposed to be recognized better than others (Goetschalckx et al., 2018; J. Han et al., 2015; Isola et al., 2014, p. 20023; Newman et al., 2020). Understanding the neural correlates of this stimulus attribute, termed memorability, has vast neuroscience and commercial applications. However, fMRI studies have largely focused on the memorability of still images, neglecting information-rich and dynamic videos (but see Han et al., 2015). Thus, we test where video memorability is reflected in the brain to give insights into the link between stimuli perception and memory (Bainbridge, 2019; Bylinskii et al., 2022).

Under the hypothesis that stimuli with higher memorability scores elicit a greater magnitude of brain response (Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Jaegle et al., 2019), we correlate a vector of video memorability scores with a vector of each voxel's corresponding brain responses (beta values) (Figure 8A). Whole-brain (Figure 8B) and ROI-based (Figure 8C) analyses converged in revealing an effect in ventral visual, dorsal visual, and parietal cortex, all regions involved in video perception (green colored bars in Figure 8C). This result contrasts with previous work using images, where the effects were largely relegated to the ventral visual cortex associated with image perception (Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Jaegle et al., 2019; Mohsenzadeh, Mullin, Oliva, et al., 2019).

BMD's use of short videos propels the study of the neural correlates of memorability into the video domain, inviting further work into how visual perception and memory formation share computational resources (Cohen et al., 1997; Martin & Chao, 2001; Riou et al., 2011; Slotnick et al., 2012; Vermeulen et al., 2008; Weinberger, 2004) and bridging the study of static image memory with longform movie memory (Bainbridge & Baker, 2022; Furman et al., 2007; Hasson, Furman, et al., 2008). Understanding how other video features support other cognitive faculties, such as empathy, theory of mind, or attention, is ripe for investigation with the BMD.

Discussion

Humans evolve in a dynamic visual world where critical information about the environment is conveyed through visual events. Progress in understanding how humans interpret visual

information has been achieved through work at the extremes of static images (Allen et al., 2022; Chang et al., 2019; Hebart et al., 2019) or longform movies (Aliko et al., 2020; Hanke et al., 2016; Hasson, Furman, et al., 2008; Lee et al., 2023; Nishimoto et al., 2011; Seeliger et al., 2019). But, these extremes have limitations and are not specifically optimized for visual event understanding. In this paper, we have introduced the BOLD Moments Dataset (BMD), which provides whole-brain fMRI responses to a large number of comprehensively annotated visual events. This dataset is poised to bridge the research communities of visual, cognitive, and computational neuroscience and lead to breakthroughs in understanding how perceptual and cognitive brain systems extract information from complex visual inputs.

As a case example, our results demonstrating correspondence in processing stages between video-computable DNN models and cortical regions extend previous studies that analyzed responses to still images. This suggests a continuation of, rather than break with, previous work and theory, and indicates the accumulation of new insights without the loss of established ones. Furthermore, we demonstrate that despite the sluggishness of the BOLD response, BMD nonetheless allows tracking of visual information processing at the level of seconds (Hasson et al., 2004; Hasson, Furman, et al., 2008; Kiebel et al., 2008; Murray et al., 2014; Orlov & Zohary, 2018; Piasini et al., 2021; Rust & DiCarlo, 2010). This reflects our careful acquisition and preparation of the fMRI data, and unlocks the potential for the application of analyses that focus on dynamics in visual processing. Thus, BMD presents a unique opportunity to probe brain function by leveraging existing artificial intelligence tools that provide methods for capturing facets of visual event, such as social expressions (Hu et al., 2022; Kahou et al., 2016; Tzirakis et al., 2018), action recognition (Carreira & Zisserman, 2017; Feichtenhofer et al., 2016; Monfort et al., 2022), integration of temporal features (Bertasius et al., 2021; Ji et al., 2013; Y. Wang et al., 2023), object detection (Fan et al., 2021; Shafiee et al., 2017), and video reconstruction (Han et al., 2019; Kupershmidt et al., 2022).

Much of the power of BMD consists in the comprehensive set of behavioral and semantic metadata that we have collected and provided. We have shown how these metadata can be combined with AI language tools to successfully delineate visual representations at different complexity levels. This yields theoretically relevant insights – for example, the superior predictivity of complex sentence descriptions in the ventral visual stream (see Figure 7) suggests that the function of this part of the brain extends beyond object recognition (Doerig et al., 2022). As another example, our crowd-sourced behavioral data on memorability (see Figure

8) demonstrates how BMD can be used to bridge visual processing and higher-level cognitive functions.

Though we have demonstrated that temporal information is meaningfully present in BMD (see Figures 5 and 6), BMD is fundamentally limited by the temporal resolution of fMRI signals. Alternative data acquisition methods are necessary to capture fine-grained neural dynamics at the level of milliseconds. A promising direction for future research is to measure M/EEG responses to the BMD stimuli and to use sophisticated methods to perform spatio-temporal integration across modalities (Cichy & Oliva, 2020; Mohsenzadeh, Mullin, Lahner, et al., 2019). The organization of BMD's brain data and metadata in BIDS format should ease the incorporation of BMD into future research endeavors.

In pursuit of interdisciplinary and transparent research, we used portions of BMD in the *The Algonauts Project 2021: How the Human Brain Makes Sense of a World in Motion*. This open challenge, in partnership with the Computational Cognitive Neuroscience (CCN) conference (Cichy et al., 2021; Naselaris et al., 2018), invites participants to predict held-out brain data using their computational models. The top three entries in *The Algonauts Project 2021* challenge each took drastically different modeling approaches (see reports in Supplementary), highlighting the creative space opened by BMD lying at the intersection of natural and artificial intelligence research.

For a full account of visual event understanding, research needs to look beyond the classical visual brain and into the whole brain, now possible with BMD. Combined with the advent of deep neural network models and advanced analysis methods, revolutionary opportunities for understanding human cognition have appeared. BMD provides the required data - brain imaging and metadata - to harvest those opportunities.

Methods

Participants

Ten healthy volunteers (6 female, mean age ± SD = 27.01 ± 3.96 years) with normal or corrected-to-normal vision participated in the experiment. All participants gave informed

consent, were screened for MRI safety, and were compensated for their time. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee (Institutional Review Board of Massachusetts Institute of Technology, approval code: 1510287948).

Stimuli

The stimulus set consisted of 1,102 videos in total. The videos were sampled from the Memento10k dataset (Newman et al., 2020), which is a subset of the Moments in Time dataset (Monfort et al., 2020) and Multi-Moments in Time dataset (Monfort et al., 2022). Each video was square-cropped and resized to 268x268 pixels. Videos had a duration of 3 seconds and frame rates ranging from 15-30 frames per second (mean = 28.3). The 1,102 videos were manually selected from the Memento10k dataset by two human observers to encompass videos that contained movement (i.e. no static content), were filmed in a natural context, and represented a wide selection of possible events a human might witness. Additional criteria were to be free of post-processing effects, textual overlays, excessive camera movements, blur, and objectionable or inappropriate content.

The 1,102 videos selected for the main experiment were split into “training” and “testing” sets; 102 videos were chosen for the testing set, and the remaining 1,000 videos formed the training set. Specifically, the testing set videos were chosen randomly from the 1,102 videos, and then checked manually to ensure no semantic overlap, in terms of objects plus actions, occurred between any pair of testing set videos. If semantic overlap was found between a pair of videos, one of these videos was swapped with a video randomly selected from the pool of remaining videos and incorporated into the testing set. This was repeated until a semantically-diverse testing set was formed. The training and testing sets are only intended to differ in the number of repetitions shown to the participant. In this way, the BMD dataset contains a low repetition (3 repetitions) training set of 1,000 videos and a high repetition (10 repetitions) testing set of 102 videos to facilitate analyses dependent on either large number of stimuli or large number of repetitions. The training and testing sets are additionally intended to mirror training and testing sets common in machine learning applications, reflecting its potential use for model building and evaluation.

(f)MRI experimental design

The fMRI data collection procedure was as follows: subjects completed a total of 5 separate fMRI sessions on separate days. Session 1 consisted of structural scans, functional localizer runs, and functional resting state scans all interspersed (Figure 1C). Sessions 2-5 consisted of the main functional experimental runs where the subjects viewed the training and testing set videos (Figure 1D). Throughout the whole experiment for a given subject, each training set video was shown a total of 3 times and each testing set video was shown a total of 10 times, resulting in 3,000 and 1,020 trials in training and testing sets, respectively. We organized trials into experimental runs such that they either only contained testing set videos (test runs) or only contained training set videos (training runs).

Session 1

Functional localizers:

Subjects completed five functional localizer runs (Figure 1C). Subjects viewed videos corresponding to one of five categories (faces, bodies, scenes, objects, and scrambled objects) in order to functionally localize each subject's category selective regions (Lafer-Sousa et al., 2016), and performed a one-back vigilance task to ensure attention to the task. Each stimulus category included 48 individual stimuli. Each run consisted of 5 blocks of fixation baseline (null) and 20 blocks of stimulus presentation for a total of 25 blocks. Baseline blocks occurred every sixth block, with 5 stimulus blocks of each category presented in between baseline blocks in a randomized order. The duration of each video was 3 seconds, with each block lasting 18 seconds. Each category stimulus block included 5 unique videos chosen randomly from the 48 stimuli plus one one-back stimulus repetition. Subject accuracy on the one-back task was 0.941 +/- 0.011 (mean ± SD).

Resting state:

Resting state data was obtained across 5 runs (Figure 1C). Within each run, participants were instructed to keep their eyes closed, to not think of anything specific, but to remain awake.

Session 2-5

Each of the 4 sessions after session 1 had identical structure. Since fixation against a dynamic video background is difficult, each session began with a 2.5 minute fixation training outside the

scanner to provide the subjects with real-time feedback of any eye movements, voluntary or involuntary (Guzman-Martinez et al., 2009). In this fixation training, subjects viewed a black-and-white random dot display flickering in counter phase. With fixation, this display evoked the illusion of a uniform gray display. Breaking fixation with eye movements disrupts this illusion, and a vivid black-and-white random dot display is perceived. Subjects were instructed to keep their eyes fixated as to minimize the disruption of the illusion.

Main experiment:

Inside the scanner, videos were presented at the center of the screen subtending 5 degrees visual angle and overlaid with a central red fixation cross (0.52 degree visual angle). Subjects were instructed to focus on the fixation cross for the duration of the main experiment. The experiment consisted of both video presentation trials (occurring 75% of the time) and null trials (occurring 25% of the time). Both trial types were 4 seconds long. The video presentation trial consisted of a 3 sec video presentation followed by a 1 sec intertrial interval. The videos were presented in random order with the constraint of no consecutive repetition. The null trial consisted of the presentation of a gray screen. During the null trial the fixation cross turned darker for 1,000 ms and participants reported the change with a button press. Participant accuracy in this task was 0.964 +/- 0.014 (mean ± SD).

The testing and training set videos were presented within training and testing runs, where each test run consisted of 113 trials and each train run consisted of 100 trials. Test and training runs were randomly interspersed with the restriction of excluding successive repetitions of test runs within one session. Each session contained 3 test runs and 10 training runs and lasted approximately 100 minutes.

Metadata

Visual events consist of complex combinations of objects, locations, actions, sounds and more. To capture the many dimensions of visual events, we characterized each video with a set of seven annotations, including object labels, scene labels, action labels, text descriptions, a spoken transcription, a memorability score, and an index of memorability decay rate (Figure 1AB). Five annotations were collected for the object, scene, action, and text description labels to ensure comprehensive coverage and form a group consensus.

Object labels

For each video, we obtained 5 object labels by feeding a sequence of video frames sampled at regular intervals into a ResNet50 model (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) to classify objects. We averaged the model's predictions for each of the frames to get the top 5 object predictions. We do not enforce that all 5 object labels describe the same object; instead, the five object labels describe up to five objects in the video.

Scene labels

Similar to the object labels, we obtained 5 scene labels per video by feeding a sequence of video frames sampled at regular intervals into a ResNet50 model (He et al., 2016) pre-trained on Places (Zhou et al., 2018) to classify scenes. We average the model's predictions for each of the frames to get the top 5 scene predictions.

Action labels

The 5 action labels were human-generated by workers on the crowd-sourcing platform Prolific. We restricted the possible action labels to be from one of 292 possible action labels that broadly encompass meaningful human actions (as in Monfort et al., 2022). The participant viewed these 292 possible action labels, watched the video, and selected one action label that best described the video. Each of the 5 action labels per video were produced by different participants, and the same human may annotate multiple videos. Two authors manually reviewed the labels to ensure the labels assigned to the video were sensical (i.e. participants were not choosing labels at random).

Text descriptions

We obtained five high-level descriptions of the content of each video in the form of written descriptions. The five text descriptions were human-generated from participants on the crowd-sourcing platform Amazon Mechanical Turk (AMT). Their task was to watch the video and type a one sentence description. Each of the 5 text descriptions were produced by a different human participant, and each participant was allowed to annotate multiple videos. The authors manually checked the text descriptions to ensure the text descriptions pertained to the video (i.e. participants followed instructions) and to correct obvious typos.

Spoken transcriptions

We obtained one spoken description per video to capture emotional and descriptive nuances typically conveyed in speech but not present in typing. The spoken description was collected via human participants on AMT, as in Monfort et al., 2021. Participants were instructed to watch the video and verbally describe it. They were given no instructions pertaining to the length of the description. We record the audio file and use Google's speech-to-text transcription to generate a text transcription. The transcription was manually checked to ensure it pertained to the video (i.e. participants followed instructions) and to correct obvious typos. We release the text transcription but not the original audio file for privacy purposes.

Memorability score and decay rate

The memorability score and decay rate were measured by Newman et al., 2020, where AMT human participants played a video memory game. The game consisted of a continuous video stream where the participant pressed the spacebar upon seeing a repeated video. Repeated videos were presented at various delays, from 30 seconds to ten minutes. The participant's responses were then used to calculate a video's memorability score from 0 (no recall) to 1 (perfect recall) and memorability decay rate from 0 (no decay) to -inf (instantaneous decay).

fMRI data preprocessing and analysis

fMRI data acquisition

The MRI data were acquired with a 3T Trio Siemens scanner using a 32-channel head coil. During the experimental runs, T2*-weighted gradient-echo echo-planar images (EPI) were collected (TR = 1750 ms, TE = 30 ms, flip angle = 71°, FOV read = 190 mm, FOV phase = 100%, bandwidth = 2268 Hz/Px, resolution = 2.5 × 2.5 × 2.5 mm, slice gap = 10%, slices = 54, multi-band acceleration factor = 2, ascending interleaved acquisition). Additionally, a T1-weighted image (TR = 1900 ms, TE = 2.52 ms, flip angle = 9°, FOV read = 256mm, FOV phase = 100%, bandwidth = 170 Hz/px, resolution = 1.0 x 1.0 x 1.0 mm, slices = 176 sagittal slices, multi-slice mode= single shot, ascending) and T2-weighted image (TR = 7970ms, TE = 120 ms, flip angle = 90°, FOV read = 256 mm, FOV phase = 100%, bandwidth = 362 Hz/Px, resolution = 1.0 × 1.0 × 1.1 mm, slice gap = 10%, slices = 128, multi-slice mode = interleaved, ascending)

were obtained as high-resolution anatomical references. We acquired resting state and functional localizer data using acquisition parameters identical to the main experimental runs.

Preprocessing

All MRI data was first converted to BIDS format (K. J. Gorgolewski et al., 2016). All data from all sessions was then preprocessed using the standardized fMRIPrep preprocessing pipeline. As recommended by fMRIPrep to increase transparency and reproducibility in MRI preprocessing, we copy their generated preprocessing text in its entirety below:

> “”
>
> Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.2.1 (Esteban et al., 2019, 2022; RRID:SCR_016216), which is based on *Nipype* 1.5.1 (Esteban, Oscar et al., 2022; K. Gorgolewski et al., 2011; RRID:SCR_002502).
>
> ### Anatomical data preprocessing
>
> A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, RRID:SCR_001847, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: *ICBM 152 Nonlinear*

Asymmetrical template version 2009c [Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym].

Functional data preprocessing

For each of the 62 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. A B0-nonuniformity map (or *fieldmap*) was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of *SDCFlows* inspired by the epidewarp.fsl script and further improvements in HCP Pipelines (Glasser et al., 2013). The *fieldmap* was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other *SDCflows* tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox & Hyde, 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD

was computed using two formulations following Power (absolute sum of relative motions, Power et al., 2014) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al., 2002). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces).

> Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).
>
> Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.
>
> “”

Next, the fMRI data from the main experimental runs (sessions 2-5) underwent temporal resampling. In detail, we resampled each voxel's time series using cubic interpolation to change the acquisition TR of 1.75 second to a new TR of 1 second. The 1.75 second acquisition TR combined with a 4 second trial length allowed for a dense sampling of the BOLD response relative to stimulus onset and thereby a good estimate of the BOLD response shape. However, as on every trial the timing of MR acquisition with respect to the trial was different, analysis of the BOLD response time-locked to the onset of each trial was cumbersome. The interpolation to a 1 second TR achieved a time series with regular sampling of the BOLD response relative to stimulus onset, enabling analysis time-locked to the onset of the videos.

Functional localizer scans were subsequently smoothed with a 9 mm full width half maximum of the Gaussian kernel. The main experimental functional runs use unsmoothed data.

General linear model

Functional localizer

To model the hemodynamic response to the localizer videos, the preprocessed fMRI data, video, and fixation baseline onsets and durations were included in a general linear model (GLM). Motion and run regressors were included as regressors of no interest. All regressors were convolved with a hemodynamic response function (canonical HRF) to calculate beta estimates.

Main experiment

We modeled the BOLD-signal of each voxel in the preprocessed fMRI data of each participant as a weighted combination of simple Finite Impulse Response (FIR) basis functions. We modeled the BOLD response with respect to each video onset from 1 to 9 seconds in 1 second steps (corresponding to the resolution of the resampled time series). Within this time interval the voxel-wise time course of activation was high-pass filtered (removing signal with $f<1/128$ Hz) and serial correlations due to aliased biorhythms or unmodelled neuronal activity were accounted for using an autoregressive AR(1) model.

Using FIR in the way described above, we modeled every trial in the experimental run of each session, simultaneously capturing the spatial variability and the temporal evolution of brain responses underlying visual event understanding. For every session we generated separate FIR models for training and testing sets. Overall, for each video condition in the testing set we extracted 10 (repetitions) x 9 (seconds) estimated beta values, and for each video condition in the training set we extracted 3 (repetitions) x 9 (seconds) beta estimates.

ROI definitions of early visual cortex and ventral visual stream

We computed five t-contrasts (FWE corrected at $p=0.05$) per subject based on the beta values from the localizer experiment to quantify category-specific voxel activations. The t-contrasts were used to localize voxels in early visual regions (objects & scrambled objects > baseline; V1v, V1d, V2v, V2d, V3v, V3d, hV4), a body-selective region (bodies > objects; EBA), an object-selective region (objects > scrambled objects; LOC), face-selective regions (faces > objects; FFA, OFA, STS), and scene-selective regions (scenes > objects; PPA, RSC, TOS). Together, these 15 ROIs cover brain regions along the ventral visual pathway thought to transform low-level visual features into complex, semantic representations useful for object recognition.

The ROI masks for each subject were defined using the subject’s t-contrast maps. Voxels outside the anatomical region of interest were manually set to 0, and ROIs were limited to 1,000 voxels in size. In the event that two or more ROIs overlapped, we used a probability map computed from all ten subjects’ t-contrasts to assign each overlapping voxel to the ROI it most likely belonged to. In detail:

We first created a probabilistic map for each of the 8 category-selective and 7 early visual ROIs, using data from all 10 subjects. Each subject's five t-contrast maps from the localizer experiment were family-wise error (FWE) corrected at $p = 0.05$. We then added each subject's binarized version of the FWE corrected t-contrast maps (1 if the voxel passed FWE correction, 0 if not) and divided by the total number of subjects. This resulted in 5 probabilistic maps corresponding to each of the 5 t-contrasts, where each voxel had a value from 0 to 1 in steps of 0.1 (0, 0.1, 0.2, …, 1.0) representing the decimal percentage of the number of subjects where that voxel was significant. We used the "objects & scrambled objects > baseline" t-contrast probability map to define a probabilistic map for each of the 7 early visual ROIs by masking the t-contrast probability map with the early visual masks defined by L. Wang et al., 2015. We used the remaining 4 category-selective t-contrast probabilistic maps to define a probabilistic map for each of the 8 category-selective ROIs by visually inspecting the appropriate t-contrast probability map (e.g. inspecting the scenes > objects t-contrast probability map to define the scene-selective ROIs) and manually setting the voxels clearly outside the region of interest to 0. This way we defined a total of 15 separate probabilistic maps.

Next, we defined 15 ROI masks for each subject. For each subject, we masked their own appropriate FWE corrected t-contrast map with the corresponding binarized t-contrast probability map. The ROIs were limited to the top 1,000 voxels. This process resulted in fifteen subject-specific ROI masks.

Lastly, we assigned any overlapping voxels to a single ROI. In the event two or more voxels overlapped, we assigned the overlapping voxels to a single ROI based on the ROI probability map in step 1. Exemplified for two ROIs A and B, if ROI A and ROI B overlapped on voxel x, we indexed into ROI A's probability map at voxel x and ROI B's probability map at voxel x. We assigned voxel x to the ROI with the higher probability. If ROI A and B have equal probabilities at voxel x, we grew a patch by one voxel along each dimension and compared the ROIs' mean probabilities within that patch. This process was repeated until there was no longer a tie between ROIs and each ROI was non-overlapping.

Subject 6 did not show any responses in TOS and RSC, and subject 7 did not show any responses in RSC. The ROI masks were used in subsequent analyses to extract active voxels within a subject's ROI during the main experiment.

ROI definitions of dorsal visual stream

We additionally defined dorsal regions of interest using anatomical landmarks, since our functional localizer was designed to define early visual and ventral category-selective regions only. All dorsal regions were defined the same way for all subjects. Specifically, we use the maximum probability map in Wang et al. (2015) to anatomically define area V3ab (grouping areas V3a and V3b), IPS0, and IPS1-2-3 (grouping areas IPS1, IPS2, and IPS3). We additionally define 7AL, BA2, PFt, and PFop in more superior regions of the dorsal stream using the atlas described in Glasser et al., 2016.

Univariate split-half reliability analysis

To select voxels with high signal-to-noise ratio, we defined a selection criteria based on split-half trial reliability. Our assumption behind the criteria was that the voxel responses on different trials corresponding to the same video should be more correlated with each other than voxel responses to different videos. As the voxel response to a video, we used the beta estimates from the FIR model averaged over time points TR 5-9 (representing the peak of a typical BOLD signal). We then z-scored the voxel responses across the videos.

We divided the voxel responses of the testing set stimuli trials (n=10 trials) into two equal splits and calculated the Pearson correlation (ρ) between the splits. The split-half reliability was calculated using the Spearman-Brown formula ($2\rho/(1+\rho)$), where the maximum reliability is 1. We calculated the split-half reliability for all possible combinations of splits and used the mean reliability as the reliability for that voxel.

To assess if reliability is better than chance, we first estimated chance-level reliability. For each voxel, we calculated the split-half reliability for all possible combinations of splits while randomly permuting the video indices for one of the two splits. This process was repeated 100 times with a different video index permutation each time. This procedure resulted in 100 random reliability values for each voxel, which was used to calculate a p-value. The voxels that satisfy our reliability criteria ($p < 0.05$) are referred to as “reliable voxels”. Due to the testing set’s high number of repetitions, the reliable voxels were defined using data from the testing set runs.

Multivariate searchlight-based reliability analysis

We computed the upper and lower noise ceilings at each voxel in the whole brain using the testing set and present subject-specific and subject-averaged results (Supplementary Figure S3). For each subject separately, the raw beta values at each voxel were z-scored across stimuli, averaged across TRs 5-9, and averaged across stimuli repetitions to result in a (n_stimuli x 1) vector of beta values. A spherical searchlight with radius 4 voxels was defined and centered on a voxel *v*. The (n_stimuli x 1) vector of beta values over all voxels contained within the searchlight sphere compose a (n_stimuli x n_voxel) matrix. The 1-Pearson correlation of the matrix resulted in a single Representational Dissimilarity Matrix (RDM) of size (n_stimuli x n_stimuli) for the voxel *v*. The searchlight then centered on the next voxel, and the procedure was repeated until an RDM was calculated for each voxel. This procedure was repeated for each of the 10 subjects.

To compute the upper noise ceiling at voxel *v*, the voxel's searchlight-computed RDM from one subject was correlated (Spearman's R) with the 10-subject group averaged RDM. The average of each subject's correlation to the 10-subject group averaged RDM is the upper noise ceiling for voxel *v*. This process was repeated over all voxels to result in the upper noise ceiling values throughout the whole brain displayed in Figure 3B. The upper noise ceiling estimates the highest correlation that a model can be expected to obtain given the noise in the data.

To compute the lower noise ceiling at voxel *v*, the voxel's searchlight-computed RDM from one left-out subject was correlated (Spearman's R) with the remaining 9-subject group averaged RDM. The average of each left-out subject's correlation to the corresponding 9-subject group averaged RDM is the lower noise ceiling for voxel *v*. This process was repeated over all voxels to estimate the lower noise ceiling values throughout the whole brain.

Action recognition TSM ResNet50 model training

The model adopts the architecture of a Temporal Shift Module (TSM) (Lin et al., 2019), with ResNet50 as the backbone network. We trained our model on the M4 (Multi-Moments minus Memento) training dataset for 120 epochs by using LSEP loss (Monfort et al., 2022). The M4 training dataset consists of 1,012,169 videos which are in the Multi-Moments in Time dataset but not in the Memento dataset to ensure no overlap with the 1,102 BMD stimuli. Our model

was initialized with the weights of the ResNet50 trained on ImageNet-1k dataset. During the training phase, our model split the input video into 8 segments and sampled 1 frame from each segment. We used SGD optimizer to optimize our model. The learning rate followed the cosine learning rate schedule and was initialized as 0.02. The weight decay was set to be 0.0001. The model achieved a precision-at-one score 0.593, a precision-at-five score of 0.829, and a mAP score of 0.636 (loss of 2.75054).

DNN block to cortex correspondence procedure

We used an encoding model procedure to quantify the correspondence between DNN Blocks and regions of cortex. For the DNN, we train a Temporal Shift Module (TSM) network (Lin et al., 2019) with a ResNet50 backbone on the M4 dataset (Multi-Moments in Time Minus Memento10k). In this way, we achieved a model that consecutively processed video frames (as opposed to frame averages), incorporated biologically necessary recurrent computations, and learned to perform a video-based task (i.e. action recognition) from the same set of short, natural videos of which we sample the BMD stimuli.

We ran inference on the TSM ResNet50 model using the 1,102 videos used in the fMRI experiment and extracted the activations for each video. The activations for a given block were extracted after the nonlinearity. We then used an encoding model procedure. In detail, we fit a voxel-wise linear model that predicted each individual voxel response (beta values z-scored across stimuli and averaged over TRs 5-9) from the DNN activations of the training set, and then evaluated the model on its ability to predict voxel responses to the videos of the testing set. We only predicted the values of the voxels that met the split-half reliability criteria, as described in the “Univariate Split-Half Reliability Analysis” methods section above, in order to model meaningful signal. The DNN activations underwent PCA (n=500) to ensure fair comparison of activations of different embedding sizes. For each voxel *v*, we fit a linear model from the training set DNN activations (size (n_training_videos x n_PCA_components)) to the training set fMRI responses averaged over the 3 trial repetitions (size (n_training_videos x n_voxels)). We then predicted testing set voxel responses (size (n_testing_videos x n_voxels)) by applying the linear fit on the testing set DNN activations (size (n_testing_videos x n_PCA_components)). We evaluated the performance of the prediction by correlating (Pearson) the predicted testing set voxel responses with the true testing set fMRI responses of each of the 10 testing set repetitions (size (n_testing x n_voxels)). The final performance of the prediction is the average correlation of the 10 repetitions. The noise-normalized correlation is this 10-repetition average Pearson

correlation divided by the voxel's split-half correlation value (the pearson correlation value before Spearman-Brown).

In this way, we obtained an encoding model accuracy (correlation) at each voxel in the whole brain for each of the four ResNet50 Blocks and for each subject. We averaged the noise normalized correlation at each voxel across subjects for each of the four Blocks and displayed the results in a whole-brain volume (Figure 4B).

We then computed the difference in encoding accuracy between Block 1 and Block 4 at each of the 22 ROIs to determine if a region's brain responses were predicted significantly better by activations of early or late DNN Blocks. For each subject, we computed both Block 1 and Block 4's average noise-normalized correlation (encoding accuracy) within each ROI and took the difference (Block 1 - Block 4). We then performed a t-test (one-sample, two-sided) against a null hypothesis of zero correlation and corrected for multiple comparisons across ROIs (Bonferroni, $p < 0.05$ with $n=22$). We plotted the subject-averaged Block 1 - Block 4 noise-normalized correlation differences and denoted the significant ROIs with an asterisk and a color corresponding to significance with Block 1 (blue) and BLock 4 (red) (Figure 4C). See supplementary Figure S7 to see each Block's encoding accuracy at each ROI.

Shuffling analysis to determine importance of temporal order

To determine whether the temporal order of visual information is preserved in fMRI responses of the human visual system, we compared the encoding performance of the TSM model trained on the M4 dataset (Multi-Moments in Time Minus Memento10k) with preserved order of visual information with the model with a randomly shuffled order of visual information. TSM requires eight frames as the input to the model. These frames were sampled uniformly. In the original (unshuffled) order, the order of frames was preserved for all the videos. In the shuffled case, the indices of the frames were shuffled randomly, and then for all the videos, the same order of shuffled indices was used to create the input to the TSM model. We used ten such random shuffles of indices to introduce more randomness.

We first extracted the activations from the four blocks of the TSM model for the unshuffled case and each shuffled case. Then we performed PCA to extract the top 100 components for each block. Then, we performed an encoding model procedure to predict the fMRI responses of the

testing videos. We repeated the shuffling ten times and then took the mean encoding correlation across ten shuffles to compare with the encoding results using unshuffled order of frames. As brain responses, we used beta values z-scored across stimuli and averaged over TRs 5-9. We only predicted the “reliable” voxels, as defined in the “Univariate Split-Half Reliability Analysis” methods section above.

At each of the four blocks, we computed the difference in encoding accuracy (correlation) between activations that used the unshuffled and shuffled video input (unshuffled minus shuffled). We visualized the difference in correlation at each block in a whole-brain volume (Figure 5B). Within each Block, we then computed the difference in encoding accuracy (unshuffled minus shuffled) within each ROI. We performed a t-test (one-sample, two-sided) between the subject-averaged difference in correlation and a null hypothesis correlation of zero. We corrected for multiple comparisons across ROIs (Bonferroni, $p < 0.05$ with $n=22$). The bar plot in Figure 5C displays the subject-averaged unshuffled minus shuffled correlation difference at each TSM ResNet50 Block and ROI. Significant blocks were marked with an appropriately colored asterisk.

Encoding and variance partitioning analysis procedure

The encoding algorithm involved two steps. In the first step we non-linearly transformed the stimuli videos from pixel space to the feature space of a computer vision model, using a deep neural network (DNN). We fed the first and third video second frames of the 1000 training and 102 testing videos to an AlexNet architecture (Krizhevsky et al., 2017) pre-trained on the ILSVRC-2012 image classification challenge (Russakovsky et al., 2015), and we extracted the corresponding feature maps at each layer. We then applied the following operations to the feature maps of both video seconds (first and third), independently: we appended the feature maps of all layers, averaged them across frames, standardized them (using the mean and standard deviation of the training videos feature maps) and downsampled them to 100 components through principal components analysis (PCA) (computed on the training videos feature maps). This resulted in the training feature maps of shape (1000 training videos × 100 features × 2 video seconds), and test feature maps of shape (102 testing videos × 100 features × 2 video seconds). In the second step, we linearly mapped the stimuli videos feature space onto voxel space, thus predicting the fMRI responses to videos. For each combination of (10 subjects × N fMRI voxels × 9 fMRI TRs), we trained the weights of a linear regression to predict

the fMRI training data (averaged over the three repeats) using the training feature maps of both video seconds independently as predictors, and then multiplied the learned weights with the test feature maps. This resulted in two synthetic fMRI test data instances of shape (10 subjects × 102 test videos × 9 fMRI TRs × N fMRI voxels), one for each video second.

To test our hypothesis we ran a variance partitioning analysis between the biological fMRI test data and the two instances of synthetic fMRI test data. At each subject, TR, and voxel we ran a searchlight (Kriegeskorte et al., 2006) to calculate the portion of the biological fMRI test data (averaged over the ten repeats) uniquely explained by, respectively, the synthetic fMRI test data of the first or third video seconds. We then observed at which TRs the unique variance explained by the two versions of synthetic test data peaked, and subtracted the peak TR of the first video second synthetic data from the peak TR of the third video second synthetic data. Next, we created subject wise binary whole brain masks with ones in voxels that show TR peak differences in the range 1 to 3 and zeros elsewhere, summed the binary masks across subjects, and performed a binomial test with FDR correction to remove the non-significant voxels.

The variance partitioning analysis for the ROIs was similar but performed on the reliable (split-half reliability $p<0.05$) voxels within each ROI. Again, this results in time courses that reveal how well the synthetic fMRI test data from either the first or third video second explains the real fMRI data at each of the nine TRs. To quantify this difference, we again subtracted the peak TRs of the first and third video second synthetic data.

RSA-based decoding analysis procedure

The decoding analysis is based on Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and broadly consists of correlating a Representational Dissimilarity Matrix (RDM) defined by the metadata with a RDM at each voxel in the brain defined by the brain responses (Figure 7A).

To define the metadata RDM, we feed the 5 scene, 5 action, and 5 text caption metadata annotations from each of the 102 testing set videos into a language model to generate vector embeddings for each label. The scene and action labels were fed into the FastText model (Bojanowski et al., 2017) to compute single-word embeddings and the text descriptions were fed into the Sentence-BERT (Reimers & Gurevych, 2019) model to compute sentence-level

embeddings. To minimize the effect of noise in the annotations, we average the 3 most similar vector embeddings together to result in a single vector embedding that represents the scene, action, or text caption for that video. We then compute the pairwise cosine distance between each video's vector embedding to produce a single 102 x 102 Representational Dissimilarity Matrix (RDM) for the scene, action, and text caption metadata. Figure 7B shows the rank-normalized (rank each distance value and divide by the maximum rank) RDM for the scene, action, and text description RDMs, respectively. We did not perform this analysis on the object labels since the 5 object labels often describe up to 5 different objects within the video, thus interjecting noise. The 5 scene, action, and text description labels, on the other hand, mostly described the same scene, action, or overall description.

To define the RDMs at each voxel in the brain for each subject, we perform a searchlight analysis in the way described in the "multivariate searchlight-based reliability analysis" methods section. To summarize, we center a sphere (radius of 4 voxels) around voxel *v* and extract the beta values (TRs 5-9 averaged over repetitions and z-scored across conditions) for all testing set conditions at all voxels encompassed in the sphere. Each stimuli thus has a corresponding vector of beta values, one from each voxel within the searchlight sphere. We compute the 1 - Pearson R correlation between all pairs of stimuli vectors to obtain an RDM at the centered voxel *v*. We repeat this process for all voxels in the whole brain for each subject.

We then correlate (Spearman's R) the metadata RDM (cosine-distance, not rank-normalized) with the searchlight-based RDMs at each voxel for each of the 10 subjects separately. For the whole-brain analysis (Figure 7C), we compute a t-test (1-sample, 2-sided) against a null hypothesis of a correlation of 0 at each voxel then perform FDR correction (q=0.05, assuming positive correlation) on all p-values in the whole brain to obtain a set of significant voxels. We compute the noise-normalized correlation by dividing the correlation with the voxel's upper noise ceiling and plot the 10-subject average noise-normalized correlation at each significant voxel. For the ROI-based analysis (Figure 7D), after we correlate (Spearman's R) the metadata RDM with the searchlight-based RDM at each voxel, we compute the average noise-normalized correlation within each ROI. We compute a t-test (paired, 2-sided) between each ROI's average noise-normalized correlation and "V1v"'s average noise-normalized correlation and correct for multiple comparisons (Bonferroni, $p < 0.05$ with n=21). The bar plots in Figure 7D depict the 10-subject averaged noise-normalized correlations within each ROI.

Memorability analysis procedure

For each subject, we averaged the beta values (z-scored across conditions) over TRs 5-9 and over repetitions (3 repetitions per training set stimuli and 10 repetitions per testing set stimuli) to obtain one beta value per video. From the memory game implemented by Newman et al., 2020, we had one memorability score per video. Under the hypothesis that the magnitude of brain response positively correlates with stimuli memorability (Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Jaegle et al., 2019), we performed a ranked correlation (Spearman's R) between the vector of memorability scores (size (1102 x 1)) and the vector of beta values (size (1102 x 1)) at each voxel for each subject. In this way, we obtained a correlation value at each voxel in the brain for each subject.

For the whole-brain analysis, we first performed a t-test (one-sample, one-sided) at each voxel against a null hypothesis of zero correlation. We then performed FDR correction on the p-values (q=0.05, assuming positive correlation). We visualized the subject-averaged correlations of the significant voxels that passed FDR correction in the whole-brain volume (Figure 8B).

For the ROI analysis, we computed the average correlation within each ROI for each subject. We then computed a t-test (one-sample, one-sided) for each ROI against a null hypothesis of zero average correlation and corrected for multiple comparisons (Bonferroni, $p<0.05$ with $n=22$). We plotted the subject-average correlation at each ROI (Figure 8C) and denoted significance with an asterisk and a green colored bar.

Data and Code Availability

Upon peer-reviewed publication, the raw (dicom) and pre-processed (f)MRI data, ROI masks, stimulus set, stimulus metadata, ResNet50 TSM model weights (and GitHub link to training code), The Algonauts Project 2021 supplementary material, and analysis starter scripts (in Matlab and Python) will be made available for download through the Open Science Framework (OSF).

Acknowledgements

This research was funded by DFG (CI-241/1-1, CI241/1-3,CI-241/1-7) and ERC grant (ERC-2018-StG) to R.M.C.; the Vannevar Bush Faculty Fellowship program funded by the ONR (N00014-16-1-3116) to A.O.; the Alfons and Gertrud Kassel foundation to G.R.; the EECS MathWorks Fellowship to B.L.. We also thank the MIT-IBM Watson AI Lab for support. The experiments were conducted at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, Massachusetts Institute of Technology, on a Siemens PrismaFit 3T scanner (Erlangen, Germany) supported with funding from a NIH Shared Instrumentation Grant (1S10OD021569). We would like to thank Santani Teng and Emilie Josephs for their valuable writing feedback.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://www.frontiersin.org/articles/10.3389/fninf.2014.00014

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, *7*(1), Article 1. https://doi.org/10.1038/s41597-020-00680-2

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), Article 1. https://doi.org/10.1038/s41593-021-00962-x

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Baddeley, A. (1992). Working Memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Bainbridge, W. A. (2019). Chapter One - Memorability: How what we see influences what we remember. In K. D. Federmeier & D. M. Beck (Eds.), *Psychology of Learning and Motivation* (Vol. 70, pp. 1–27). Academic Press. https://doi.org/10.1016/bs.plm.2019.02.001

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-34075-1

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152. https://doi.org/10.1016/j.neuroimage.2017.01.063

Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-26467-5

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178. https://proceedings.neurips.cc/paper/2021/hash/d384dec9f5f7a64a36b5c8f03b8a6d92-Abstract.html

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Bartels, A., & Zeki, S. (2004). Functional brain mapping during free viewing of natural scenes. *Human Brain Mapping*, *21*(2), 75–85. https://doi.org/10.1002/hbm.10153

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), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, *331*(6013), 83–87. https://doi.org/10.1126/science.1195870

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *Proceedings of the 38th International Conference on Machine Learning*, 813–824. https://proceedings.mlr.press/v139/bertasius21a.html

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2004). Action Observation Activates Premotor and Parietal Areas in a Somatotopic Manner: An fMRI Study. In *Social Neuroscience*. Psychology Press.

Bylinskii, Z., Goetschalckx, L., Newman, A., & Oliva, A. (2022). Memorability: An Image-Computable Measure of Information Utility. In B. Ionescu, W. A. Bainbridge, & N. Murray (Eds.), *Human Perception of Visual Information* (pp. 207–239). Springer International Publishing. https://doi.org/10.1007/978-3-030-81465-6_8

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action Observation and Acquired Motor Skills: An fMRI Study with Expert Dancers. *Cerebral Cortex*, *15*(8), 1243–1249. https://doi.org/10.1093/cercor/bhi007

Carandini, M. (2005). Do We Know What the Early Visual System Does? *Journal of Neuroscience*, *25*(46), 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005

Carreira, J., & Zisserman, A. (2017). *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 6299–6308. https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), Article 1. https://doi.org/10.1038/s41597-019-0052-3

Cichy, R. M., Dwivedi, K., Lahner, B., Lascelles, A., Iamshchinina, P., Graumann, M., Andonian, A., Murty, N. A. R., Kay, K., Roig, G., & Oliva, A. (2021). *The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion*. https://doi.org/10.48550/ARXIV.2104.13714

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Cichy, R. M., & Oliva, A. (2020). A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron*, *107*(5), 772–781. https://doi.org/10.1016/j.neuron.2020.07.001

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), Article 6625. https://doi.org/10.1038/386604a0

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5), 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Massachusetts Institute of Technology Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

DeYoe, E. A., & Van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. *Trends in Neurosciences*, *11*(5), 219–226. https://doi.org/10.1016/0166-2236(88)90130-0

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). *Semantic scene descriptions as an objective of human vision*. https://doi.org/10.48550/ARXIV.2209.11737

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*(2), 647–660. https://doi.org/10.1016/j.neuroimage.2007.09.034

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, *12*(9), e0184661. https://doi.org/10.1371/journal.pone.0184661

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

Esteban, O., Markiewicz, C. J., Goncalves, M., Provins, C., Kent, J. D., DuPre, E., Salo, T., Ciric, R., Pinsard, B., Blair, R. W., Poldrack, R. A., & Gorgolewski, K. J. (2022). *fMRIPrep: A robust preprocessing pipeline for functional MRI*. Zenodo. https://doi.org/10.5281/zenodo.7430291

Esteban, Oscar, Markiewicz, Christopher J., Burns, Christopher, Goncalves, Mathias, Jarecka, Dorota, Ziegler, Erik, Berleant, Shoshana, Ellis, David Gage, Pinsard, Basile, Madison, Cindee, Waskom, Michael, Notter, Michael Philipp, Clark, Daniel, Manhães-Savio, Alexandre, Clark, Dav, Jordan, Kesshi, Dayan, Michael, Halchenko, Yaroslav O., Loney, Fred, … Ghosh, Satrajit. (2022). *nipy/nipype: 1.8.3* (1.8.3). Zenodo. https://doi.org/10.5281/ZENODO.596855

Fairhall, S. L., Albi, A., & Melcher, D. (2014). Temporal Integration Windows for Naturalistic Visual Sequences. *PLoS ONE*, *9*(7), e102248. https://doi.org/10.1371/journal.pone.0102248

Fan, L., Zhang, T., & Du, W. (2021). Optical-flow-based framework to boost video object detection performance with object enhancement. *Expert Systems with Applications*, *170*, 114544. https://doi.org/10.1016/j.eswa.2020.114544

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). *Convolutional Two-Stream Network Fusion for Video Action Recognition*. 1933–1941. https://openaccess.thecvf.com/content_cvpr_2016/html/Feichtenhofer_Convolutional_Two-Stream_Network_CVPR_2016_paper.html

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Furman, O., Dorfman, N., Hasson, U., Davachi, L., & Dudai, Y. (2007). They saw a movie: Long-term memory for an extended audiovisual narrative. *Learning & Memory*, *14*(6), 457–467. https://doi.org/10.1101/lm.550407

Gazzola, V., & Keysers, C. (2009). The Observation and Execution of Actions Share Motor and Somatosensory Voxels in all Tested Subjects: Single-Subject Analyses of Unsmoothed fMRI Data. *Cerebral Cortex*, *19*(6), 1239–1255. https://doi.org/10.1093/cercor/bhn181

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Goetschalckx, L., Moors, P., & Wagemans, J. (2018). Image memorability across longer time intervals. *Memory*, *26*(5), 581–588. https://doi.org/10.1080/09658211.2017.1383435

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

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Güçlü, U., & van Gerven, M. A. J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336. https://doi.org/10.1016/j.neuroimage.2015.12.036

Guzman-Martinez, E., Leung, P., Franconeri, S., Grabowecky, M., & Suzuki, S. (2009). Rapid eye-fixation training without eyetracking. *Psychonomic Bulletin & Review*, *16*(3), 491–496. https://doi.org/10.3758/PBR.16.3.491

Han, J., Chen, C., Shao, L., Hu, X., Han, J., & Liu, T. (2015). Learning Computational Models of Video Memorability from fMRI Brain Imaging. *IEEE Transactions on Cybernetics*, *45*(8), 1692–1703. https://doi.org/10.1109/TCYB.2014.2358647

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, *198*, 125–136. https://doi.org/10.1016/j.neuroimage.2019.05.039

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.92

Hardwick, R. M., Caspers, S., Eickhoff, S. B., & Swinnen, S. P. (2018). Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience & Biobehavioral Reviews*, *94*, 31–44. https://doi.org/10.1016/j.neubiorev.2018.08.003

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced Intersubject Correlations during Movie Viewing Correlate with Successful Episodic Encoding. *Neuron*, *57*(3), 452–462. https://doi.org/10.1016/j.neuron.2007.12.009

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The Neuroscience of Film. *Projections*, *2*(1), 1–26. https://doi.org/10.3167/proj.2008.020102

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, *72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *ELife*, *9*, e56601. https://doi.org/10.7554/eLife.56601

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, *76*(2), 423–434. https://doi.org/10.1016/j.neuron.2012.08.011

Hu, M., Ge, P., Wang, X., Lin, H., & Ren, F. (2022). A spatio-temporal integrated model based on local and global features for video expression recognition. *The Visual Computer*, *38*(8), 2617–2634. https://doi.org/10.1007/s00371-021-02136-z

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., Handwerker, D. A., Keilholz, S., Kiviniemi, V., Leopold, D. A., de Pasquale, F., Sporns, O., Walter, M., & Chang, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage*, *80*, 360–378. https://doi.org/10.1016/j.neuroimage.2013.05.079

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the Intentions of Others with One's Own Mirror Neuron System. *PLOS Biology*, *3*(3), e79. https://doi.org/10.1371/journal.pbio.0030079

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *ELife*, *8*, e47596. https://doi.org/10.7554/eLife.47596

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

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Chandias Ferrari, R., Mirza, M., Warde-Farley, D., Courville, A., Vincent, P., Memisevic, R., Pal, C., & Bengio, Y. (2016). EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, *10*(2), 99–111. https://doi.org/10.1007/s12193-015-0195-2

Kanske, P., Böckler, A., Trautwein, F.-M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain–behavior relations for empathy and Theory of Mind. *NeuroImage*, *122*, 6–19. https://doi.org/10.1016/j.neuroimage.2015.07.082

Kay, K., Jamison, K. W., Zhang, R.-Y., & Uğurbil, K. (2020). A temporal decomposition method for identifying venous effects in task-based fMRI. *Nature Methods*, *17*(10), Article 10. https://doi.org/10.1038/s41592-020-0941-6

Khosla, M., Ratan Murty, N. A., & Kanwisher, N. (2022). A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, *32*(19), 4159-4171.e9. https://doi.org/10.1016/j.cub.2022.08.009

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent Processing in V1/V2 Contributes to Categorization of Natural Scenes. *Journal of Neuroscience*, *31*(7), 2488–2492. https://doi.org/10.1523/JNEUROSCI.3074-10.2011

Konen, C. S., & Kastner, S. (2008). Representation of Eye Movements and Stimulus Motion in Topographically Organized Areas of Human Posterior Parietal Cortex. *Journal of Neuroscience*, *28*(33), 8361–8375. https://doi.org/10.1523/JNEUROSCI.1930-08.2008

Krekelberg, B., Dannenberg, S., Hoffmann, K.-P., Bremmer, F., & Ross, J. (2003). Neural correlates of implied motion. *Nature*, *424*(6949), 674–677. https://doi.org/10.1038/nature01852

Kret, M. E., Pichon, S., Grèzes, J., & de Gelder, B. (2011). Similarities and differences in perceiving threat from dynamic faces and bodies. An fMRI study. *NeuroImage*, *54*(2), 1755–1762. https://doi.org/10.1016/j.neuroimage.2010.08.012

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

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

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://www.frontiersin.org/article/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/7813d1590d28a7dd372ad54b5d29d033-Abstract.html

Kupershmidt, G., Beliy, R., Gaziv, G., & Irani, M. (2022). *A Penny for Your (visual) Thoughts: Self-Supervised Reconstruction of Natural Movies from Brain Activity*. https://doi.org/10.48550/ARXIV.2206.03544

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *Journal of Neuroscience*, *36*(5), 1682–1697. https://doi.org/10.1523/JNEUROSCI.3164-15.2016

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, *1*(1), 76–85.

Le, A., Vesia, M., Yan, X., Crawford, J. D., & Niemeier, M. (2017). Parietal area BA7 integrates motor programs for reaching, grasping, and bimanual coordination. *Journal of Neurophysiology*, *117*(2), 624–636. https://doi.org/10.1152/jn.00299.2016

Lee, H., Chen, J., & Hasson, U. (2023). A functional neuroimaging dataset acquired during naturalistic movie watching and narrated recall of a series of short cinematic films. *Data in Brief*, *46*, 108788. https://doi.org/10.1016/j.dib.2022.108788

Lin, J., Gan, C., & Han, S. (2019). *TSM: Temporal Shift Module for Efficient Video Understanding*. 7083–7093. https://openaccess.thecvf.com/content_ICCV_2019/html/Lin_TSM_Temporal_Shift_Module_for_Efficient_Video_Understanding_ICCV_2019_paper.html

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621. https://doi.org/10.1146/annurev.ne.19.030196.003045

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology*, *11*(2), 194–201. https://doi.org/10.1016/S0959-4388(00)00196-3

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, *34*, 28757–28771. https://proceedings.neurips.cc/paper/2021/hash/f1676935f9304b97d59b0738289d2e22-Abstract.html

Mohsenzadeh, Y., Mullin, C., Lahner, B., Cichy, R. M., & Oliva, A. (2019). Reliability and Generalizability of Similarity-Based Fusion of MEG and fMRI Data in Human Ventral and Dorsal Visual Streams. *Vision*, *3*(1), Article 1. https://doi.org/10.3390/vision3010008

Mohsenzadeh, Y., Mullin, C., Oliva, A., & Pantazis, D. (2019). The perceptual neural trace of memorable unseen scenes. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-42429-x

Monfort, M., Jin, S., Liu, A., Harwath, D., Feris, R., Glass, J., & Oliva, A. (2021). *Spoken Moments: Learning Joint Audio-Visual Representations From Video Descriptions*. 14871–14881. https://openaccess.thecvf.com/content/CVPR2021/html/Monfort_Spoken_Moments_Learning_Joint_Audio-Visual_Representations_From_Video_Descriptions_CVPR_2021_paper.html

Monfort, M., Pan, B., Ramakrishnan, K., Andonian, A., McNamara, B. A., Lascelles, A., Fan, Q., Gutfreund, D., Feris, R. S., & Oliva, A. (2022). Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9434–9445. https://doi.org/10.1109/TPAMI.2021.3126682

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

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, *17*(12), 1661–1663. https://doi.org/10.1038/nn.3862

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., Poldrack, R. A., Shohamy, D., & Kay, K. (2018). Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. *Trends in Cognitive Sciences*, *22*(5), 365–367. https://doi.org/10.1016/j.tics.2018.02.008

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Newman, A., Fosco, C., Casser, V., Lee, A., McNamara, B., & Oliva, A. (2020). Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 223–240). Springer International Publishing. https://doi.org/10.1007/978-3-030-58517-4_14

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Olshausen, B. A., & Field, D. J. (1996a). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*(2), 333–339. https://doi.org/10.1088/0954-898X_7_2_014

Olshausen, B. A., & Field, D. J. (1996b). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), Article 6583. https://doi.org/10.1038/381607a0

Orlov, T., & Zohary, E. (2018). Object Representations in Human Visual Cortex Formed Through Temporal Integration of Dynamic Partial Shape Views. *Journal of Neuroscience*, *38*(3), 659–678. https://doi.org/10.1523/JNEUROSCI.1318-17.2017

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, *292*(5516), 510–512.

Piasini, E., Soltuzu, L., Muratore, P., Caramellino, R., Vinken, K., Op de Beeck, H., Balasubramanian, V., & Zoccolan, D. (2021). Temporal stability of stimulus representation increases along rodent visual cortical hierarchies. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-24456-3

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*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Press, W. A., Brewer, A. A., Dougherty, R. F., Wade, A. R., & Wandell, B. A. (2001). Visual areas and spatial summation in human visual cortex. *Vision Research*, *41*(10), 1321–1332. https://doi.org/10.1016/S0042-6989(01)00074-8

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, *38*(33), 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, *12*(1), 5540. https://doi.org/10.1038/s41467-021-25409-6

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://doi.org/10.48550/ARXIV.1908.10084

Ress, D., & Heeger, D. J. (2003). Neuronal correlates of perception in early visual cortex. *Nature Neuroscience*, *6*(4), Article 4. https://doi.org/10.1038/nn1024

Riou, B., Lesourd, M., Brunel, L., & Versace, R. (2011). Visual memory and visual perception: When memory improves visual search. *Memory & Cognition*, *39*(6), 1094–1102. https://doi.org/10.3758/s13421-011-0075-2

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, *11*(4), Article 4. https://doi.org/10.1038/nrn2805

Roberts, J., Wallis, G., & Breakspear, M. (2013). Fixational eye movements during viewing of dynamic natural scenes. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00797

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and Tolerance (“Invariance”) Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Schneider, W. X. (2013). Selective visual processing across competition episodes: A theory of task-driven visual attention and working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130060. https://doi.org/10.1098/rstb.2013.0060

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* (p. 407007). bioRxiv. https://doi.org/10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Schultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, *194*(3), 465–475. https://doi.org/10.1007/s00221-009-1721-9

Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E., & Gerven, M. A. J. van. (2019). *A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time* (p. 687681). bioRxiv. https://doi.org/10.1101/687681

Senior, C., Barnes, J., Giampietroc, V., Simmons, A., Bullmore, E. T., Brammer, M., & David, A. S. (2000). The functional neuroanatomy of implicit-motion perception or 'representational momentum.' *Current Biology*, *10*(1), 16–22. https://doi.org/10.1016/S0960-9822(99)00259-6

Shafiee, M. J., Chywl, B., Li, F., & Wong, A. (2017). *Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video*. https://doi.org/10.48550/ARXIV.1709.05943

Shirai, N., & Imura, T. (2014). Implied motion perception from a still image in infancy. *Experimental Brain Research*, *232*(10), 3079–3087. https://doi.org/10.1007/s00221-014-3996-8

Silvanto, J., Cowey, A., Lavie, N., & Walsh, V. (2005). Striate cortex (V1) activity gates awareness of motion. *Nature Neuroscience*, *8*(2), Article 2. https://doi.org/10.1038/nn1379

Silvanto, J., Lavie, N., & Walsh, V. (2005). Double Dissociation of V1 and V5/MT activity in Visual Awareness. *Cerebral Cortex*, *15*(11), 1736–1741. https://doi.org/10.1093/cercor/bhi050

Silver, M. A., & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences*, *13*(11), 488–495. https://doi.org/10.1016/j.tics.2009.08.005

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2012). Visual memory and visual mental imagery recruit common control and sensory regions of the brain. *Cognitive Neuroscience*, *3*(1), 14–20. https://doi.org/10.1080/17588928.2011.578210

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., & Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682. https://doi.org/10.1016/j.tics.2013.09.016

Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D., & Tolhurst, D. J. (2003). The Receptive-Field Organization of Simple Cells in Primary Visual Cortex of Ferrets under Natural Scene Stimulation. *Journal of Neuroscience*, *23*(11), 4746–4759. https://doi.org/10.1523/JNEUROSCI.23-11-04746.2003

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-End Speech Emotion Recognition Using Deep Neural Networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5089–5093. https://doi.org/10.1109/ICASSP.2018.8462677

VanRullen, R., & Thorpe, S. J. (2001). The Time Course of Visual Processing: From Early Perception to Decision-Making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461. https://doi.org/10.1162/08989290152001880

Vermeulen, N., Corneille, O., & Niedenthal, P. M. (2008). Sensory load incurs conceptual processing costs. *Cognition*, *109*(2), 287–294. https://doi.org/10.1016/j.cognition.2008.09.004

Wang, L., Mruczek, R. E. B., 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

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M., & Fei-Fei, L. (2023, January 23). *Eidetic 3D LSTM: A Model for Video Prediction and Beyond*. International Conference on Learning Representations. https://openreview.net/forum?id=B1lKS2AqtX

Watson, A. B., & Ahumada, A. J. (1985). Model of human visual-motion sensing. *JOSA A*, *2*(2), 322–342. https://doi.org/10.1364/JOSAA.2.000322

Weinberger, N. M. (2004). Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience*, *5*(4), Article 4. https://doi.org/10.1038/nrn1366

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/9a1756fd0c741126d7bbd4b692ccbd91-Abstract.html

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yildirim, I., Wu, J., Kanwisher, N., & Tenenbaum, J. (2019). An integrative computational architecture for object-driven cortex. *Current Opinion in Neurobiology*, *55*, 73–81. https://doi.org/10.1016/j.conb.2019.01.010

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. https://doi.org/10.1109/42.906424

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Supplementary Material

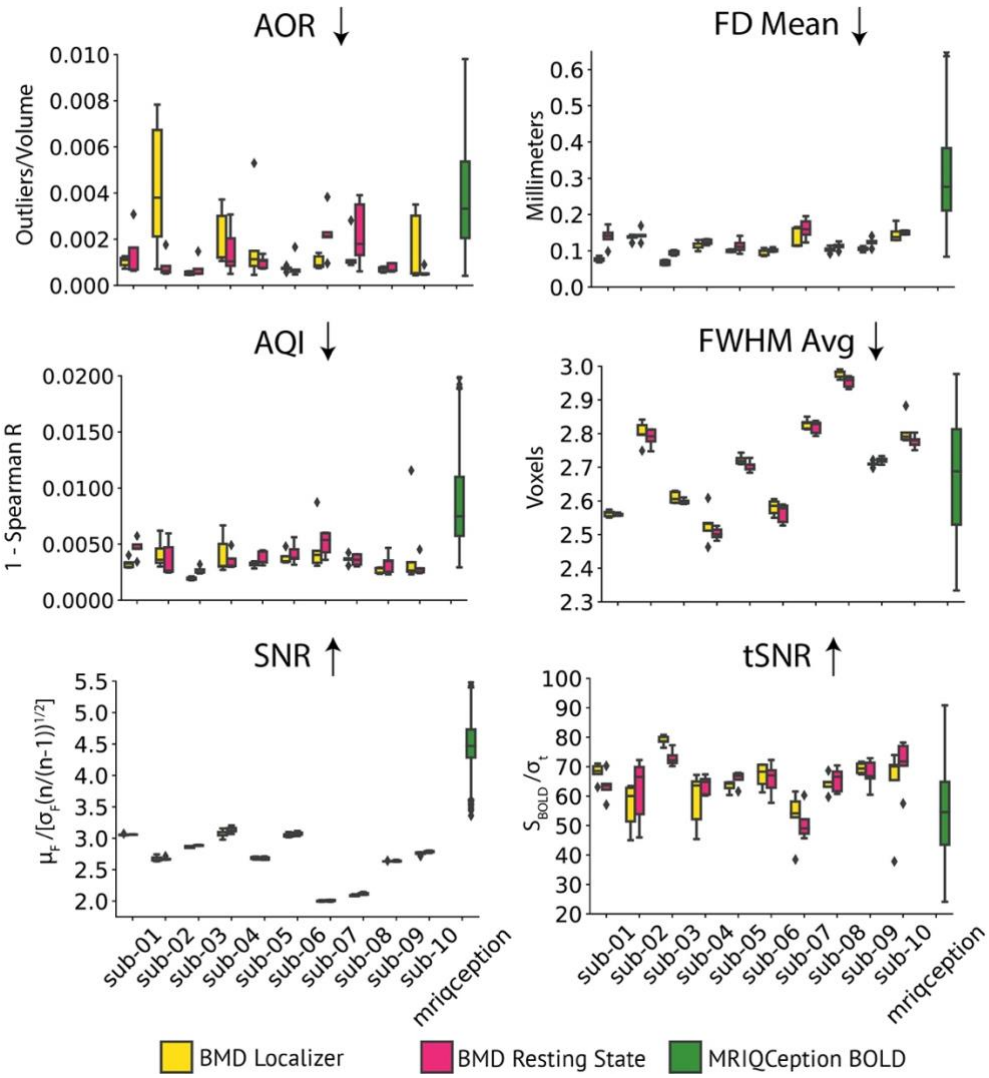

Figure S1: Localizer and resting state functional scan data quality

Panels compares boxplots of IQM values for each subject for the localizer and resting state functional tasks in the BOLD Moments experiment (per subject: localizer, n=5; resting state, n=5) with anonymous BOLD data pulled from the MRIQCeption API (BOLD, n=687). MRIQCeption does not distinguish between different tasks within BOLD scans. The boxplot extends 1.5 times the high and low quartiles, with outliers defined as a scan with a value outside that range and denoted by diamonds. The up or down arrows after the IQM title correspond to whether higher or lower IQM values denote higher data quality.

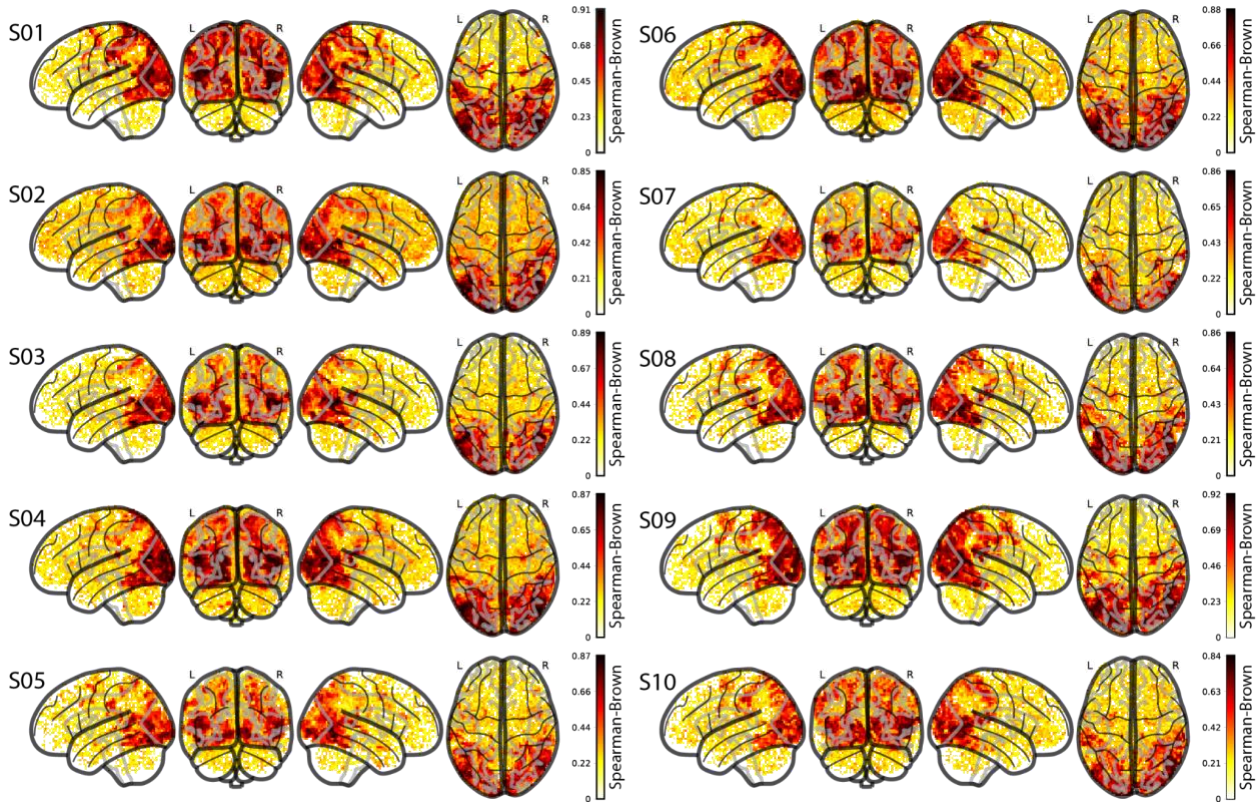

Figure S2: Whole-brain split-half reliability for all subjects

The glass brains show the split-half reliability (Spearman Brown) at every voxel for each of the ten subjects. A Pearson R correlation value was obtained by correlating random splits of the 10 repetitions from the 102 testing videos. The Spearman Brown split-half reliability was computed using the Pearson R (ρ) value obtained above in the formula: Spearman Brown = ($2\rho/(1+\rho)$).

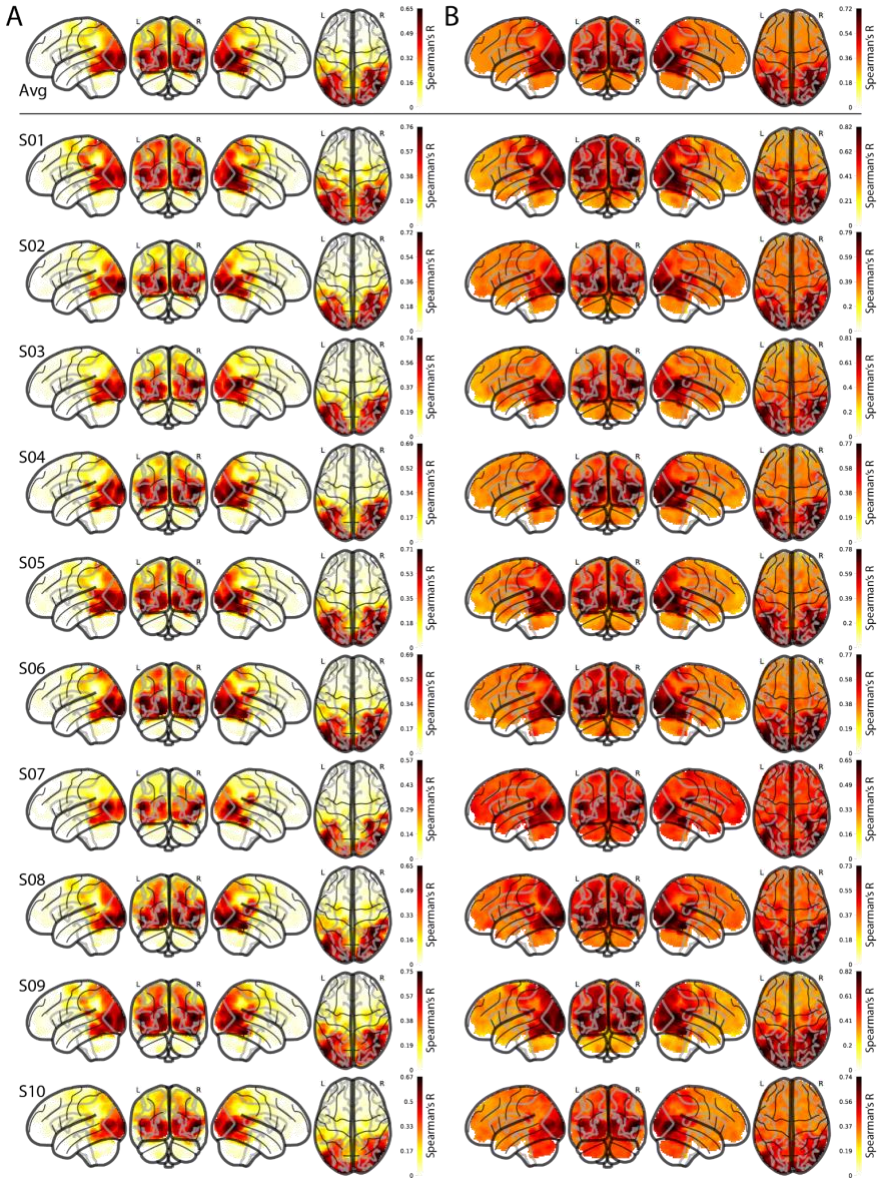

Figure S3: Whole-brain multivariate searchlight-based noise ceilings for all subjects

(A) Lower noise ceiling: The lower noise ceiling is computed using a leave-one-out correlation procedure, where a subject's RDM at a given voxel *v* is correlated (Spearman's R) with the remaining nine-subject average RDM at that voxel *v*, repeated over all voxels. **(B) Upper noise ceiling:** The upper noise ceiling is computed by correlating (Spearman's R) a subject's RDM at a given voxel *v* with the ten-subject group average RDM at that voxel *v*, repeated over all voxels. We show the whole-brain visualization for the upper and lower noise ceilings averaged over all subjects (top row) and each subject individually (bottom). The brain responses used to compute the RDMs are from the beta values of the testing set.

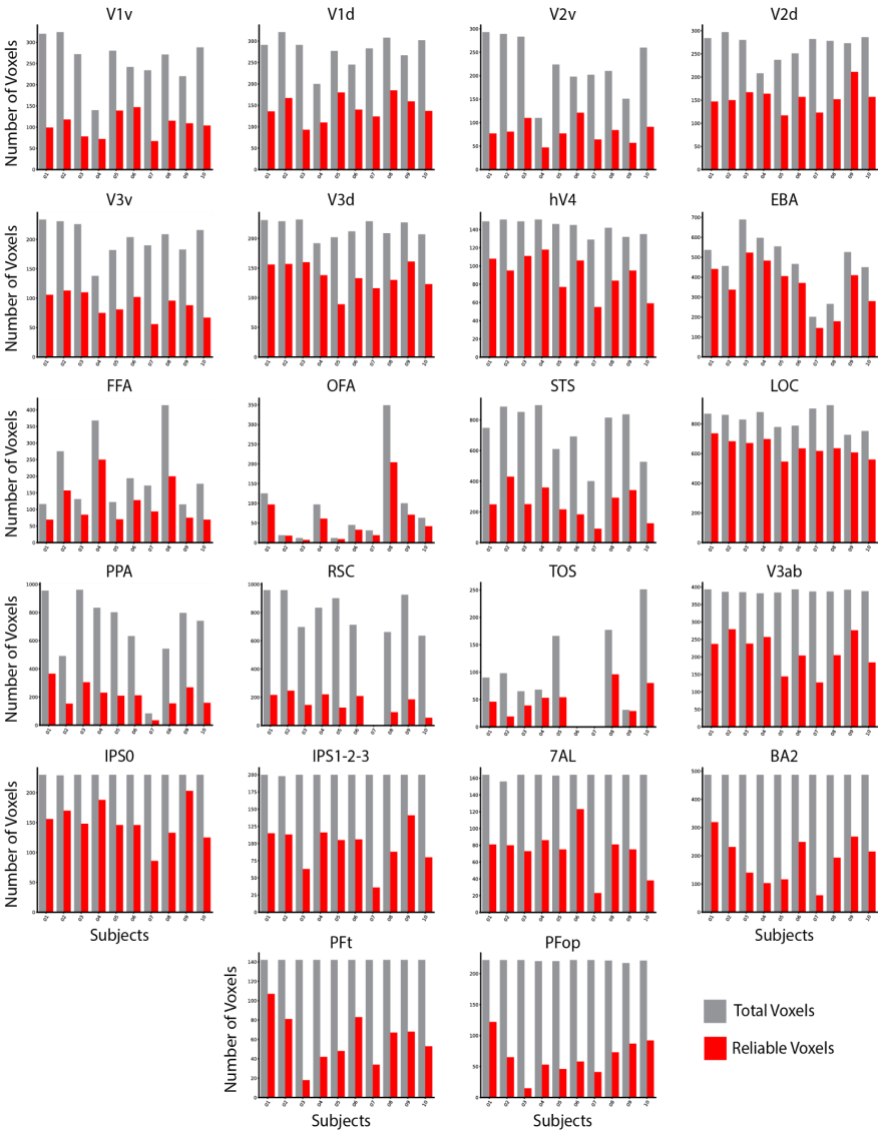

Figure S4: ROI reliability and size separated for each subject

Barplots of the total number of voxels in the ROI mask (gray bar) and the total number of reliable voxels ($p < 0.05$, Spearman-Brown) in the ROI mask (red bar) for each subject across the twenty-two ROIs. Subject 6 did not show any activation from the functional localizer task for ROI TOS, and subject 7 did not show any activation from the functional localizer task for ROIs RSC and TOS.

Figure S5: Average reliability in each ROI for each subject

Barplots of the average split-half reliability in each ROI of all reliable voxels. Separated for each subject.

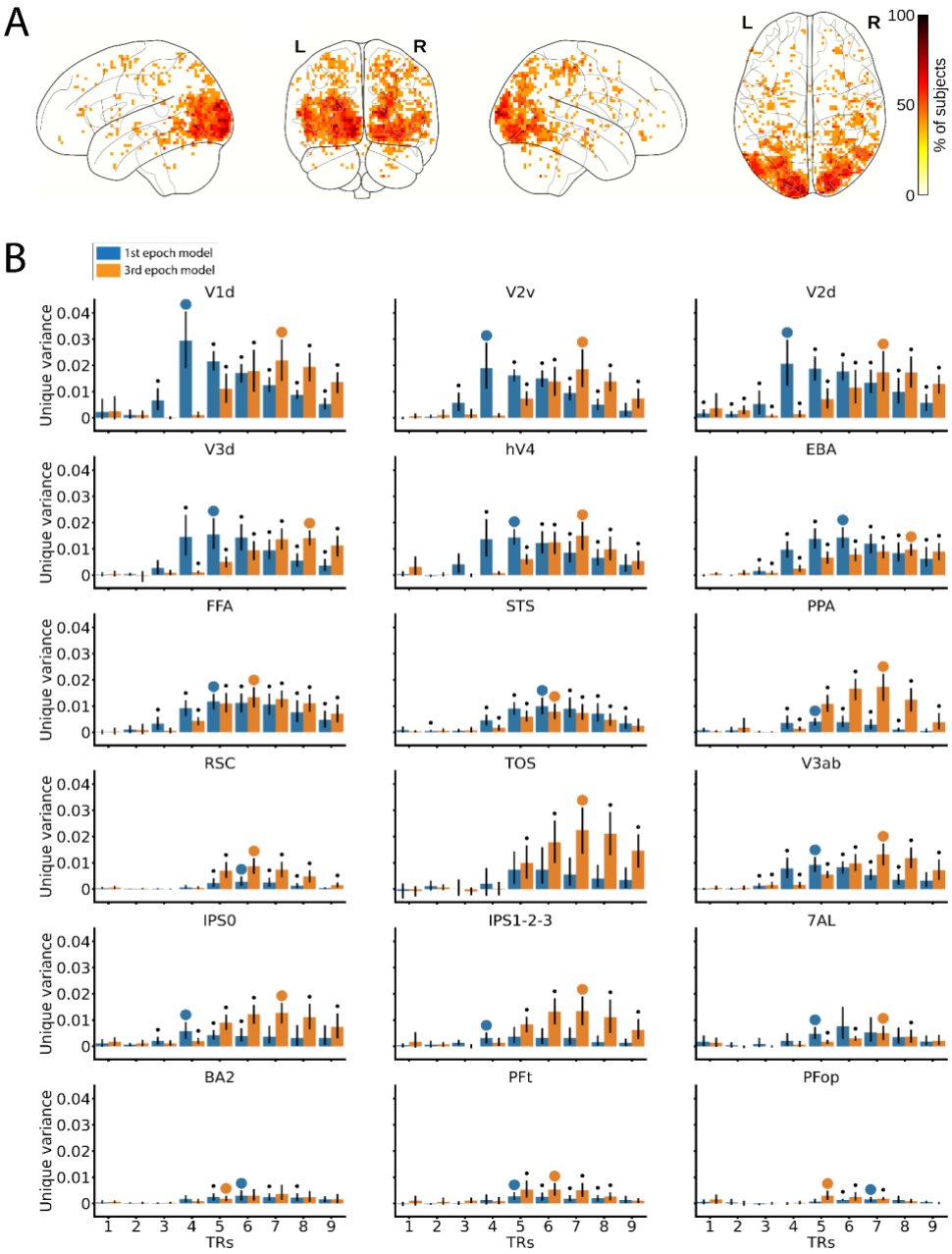

Figure S6: Encoding the temporal dynamics of the BOLD signal.

(A) Whole-brain analysis: Each voxel shows the percentage of subjects with a TR peak difference of 2 TRs at that specific voxel. Only significant voxels are plotted ($p < 0.05$, binomial test, FDR corrected). The effect of interest is showing predominantly in the visual cortex. **(B) ROI analysis:** Unique variance explained by the first and third video epoch (second) synthetic fMRI data, at each TR. Error margins reflect 95% confidence intervals. Asterisks indicate unique variance scores significantly greater than 0. Blue/orange asterisks indicate the significant unique variance peak TRs of the first/third video epoch synthetic fMRI data.

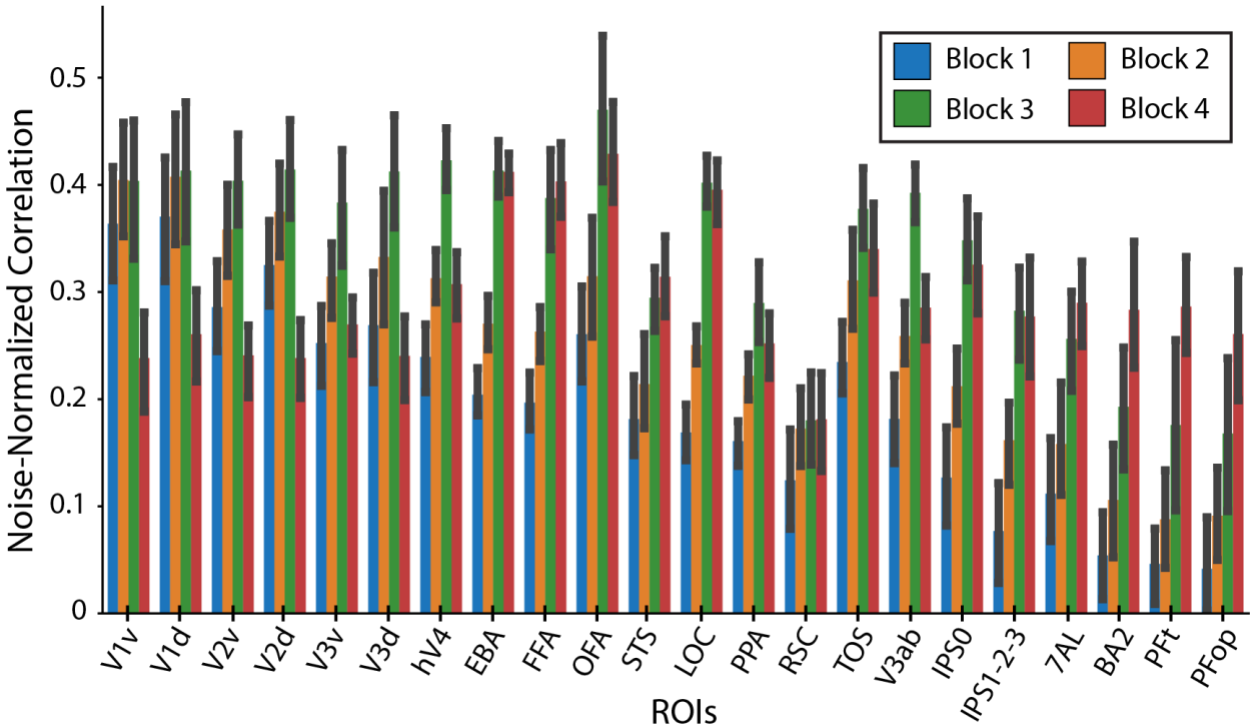

Figure S7: ROI-based Predictivity of the TSM ResNet50 Model

The barplot shows the noise-normalized predictivity of each TSM ResNet50 DNN block at each of the 22 ROIs. The values plotted are the mean ROI prediction over subjects. Error bars depict the 95% confidence interval. Data is from n=10 subjects except for TOS (n=8, no responses from subjects 6 and 7) and RSC (n=9, no responses from subject 7).

Structural and functional scan quality assessment

We use MRIQC (Esteban et al., 2017) to measure the quality of our study's original or minimally preprocessed structural and functional MRI scans. MRIQC is an open-source software that outputs a large and diverse set of image quality metrics (IQMs) to comprehensively quantify the quality of (f)MRI data in a standardized and reproducible manner. IQMs are calculated at the level of a single run, and group reports are generated for all T1w, T2w, and BOLD runs in the study. We present a representative subset of 6 IQMs to summarize the quality of our structural scans and another subset of 6 IQMs to summarize the quality of our functional scans. Note that no set of metrics can fully describe data quality in itself. Thus, when choosing IQMs to represent the structural and functional scan quality, we primarily considered the following three criteria:

First, the representative IQMs for the structural scans and for the functional scans should capture metrics especially relevant to the properties of structural and functional scans.

Second, IQMs that are useful for describing the quality of both structural and functional scans are preferred in order to create more cohesive and shared IQM subsets between the structural and functional scans.

Third, IQMs commonly reported in previous literature are preferred in order to increase comparisons across studies and be more familiar to readers.

We additionally use MRIQCeption to contextualize our study's group reported results within a large collection of anonymized group reports from studies of comparable scanner parameters (1 < Tesla < 3, 1 <= TR < 3).

For structural (T1w and T2w) scans, we present the results from the following IQMs:

SNR Total - Signal to Noise Ratio: SNR Total for structural scans is computed by averaging the SNR across the cerebrospinal fluid (snr_csf), gray matter (snr_gm), and white matter (snr_wm). SNR is calculated by the following formula:

$$\frac{\mu_F}{\sigma_F\sqrt{n(n-1)}}$$

Where μ_F is the mean intensity of the foreground, σ_F is the standard deviation of the foreground intensity, and n is the number of voxels in the foreground mask. Higher values correspond to higher quality.

CNR - Contrast to Noise Ratio: CNR, an extension of SNR, computes the absolute value difference of the gray and white matter image values ($|S_W - S_G|$) and divides them by the standard deviation of the values in the surrounding air (σ_{air}). Higher values correspond to higher quality.

CJV - Coefficient of Joint Variation: CJV is the ratio of the coefficient of variation in the gray matter to the coefficient of variation in the white matter. Lower values correspond to higher quality.

EFC - Entropy Focus Criterion: EFC is the shannon entropy of voxel intensities normalized by the maximum shannon entropy value. It measures ghosting and blurring due to head motion. Lower values correspond to higher quality.

FWHM Avg - Average Full-Width Half Maximum Smoothness: FWHM Avg is the average spatial distribution of voxel intensities in an image using a gaussian width estimator. Lower values correspond to higher quality.

FBER - Foreground-Background Energy Ratio: FBER is the ratio of the mean energy inside the head to the mean energy outside the head. Higher values correspond to higher quality.

For functional scans, we present the results from the following IQMs:

SNR - Signal to Noise Ratio: SNR for functional scans is calculated by the following formula:

$$\frac{\mu_F}{\sigma_F\sqrt{n(n-1)}}$$

Where μ_F is the mean intensity of the foreground, σ_F is the standard deviation of the foreground intensity, and n is the number of voxels in the foreground mask. Higher values correspond to higher quality.

tSNR - Temporal Signal to Noise Ratio: tSNR divides the mean BOLD signal across time by the temporal standard deviation map. Higher values correspond to higher quality.

FD Mean - Mean Framewise Displacement: FD Mean computes the average displacement of all six motion parameters. Lower values correspond to higher quality.

FWHM Avg - Average Full-Width Half Maximum Smoothness: FWHM Avg is the average spatial distribution of voxel intensities in an image using a gaussian width estimator. Lower values correspond to higher quality.

AOR - AFNI Outlier Ratio: AOR is the average fraction of outliers found in each fMRI volume as computed by AFNI's "3dToutcount" function. Lower values correspond to higher quality.
AQI - AFNI Quality Index: AQI computes the average distance between each volume and the median volume of a series, given by AFNI's "3dTqual" function. Lower values correspond to higher quality.

The Algonauts Project 2021 challenge approaches of the top three winners

The Algonauts Project 2021: How the Human Brain Makes Sense of a World in Motion is an open challenge that took place during the spring and summer of 2021 and culminated in an interactive workshop and speaking event at the Computational Cognitive Neuroscience (CCN) conference (Cichy et al., 2021; Naselaris et al., 2018). For the challenge, participants submit the predictions of their computational model on held-out brain data (see http://algonauts.csail.mit.edu/challenge.html for the final challenge leaderboard and details). We highlight the top three challenge entries, noting their different modeling approaches and insights.

The first-place team "huze" approached this challenge using an ensemble of 6 different models that together integrate meaningful features of video understanding: spatiotemporal, motion, edge, and audio features. They then weighted the outputs of each model representation and found that the predictivity for each ROI was highest when combining features from all models. They additionally optimized the receptive field size for each of the four I3D RGB model layers and ROI (Monfort et al., 2020). They showed that early ROIs benefited most from smaller receptive fields on low-level layers (layers 1 and 2) and later ROIs benefited most from larger receptive fields on high-level layers (layers 3 and 4), replicating neuroscience results (Dumoulin & Wandell, 2008).

The second-place team "bionn" was interested in evaluating a range of DNNs from the more classical supervised CNNs (AlexNet, VGG19, ResNet50, and ResNet152) to the more modern contrastive learning and visual transformer networks (simclr, pclv2, and visual transformer network ViT). They found the ResNet models, specifically ResNet152, outperformed the visual transformer and contrastive learning networks. Similar to "huze", "bionn" also took advantage of pooling the model features to simulate small receptive fields for early regions and large receptive fields for later regions.

The third-place team "shinji" experimented with state-of-the-art spatiotemporal vision features from TimeSformer (Bertasius et al., 2021) and classical, neurophysiology-based motion energy features (Nishimoto et al., 2011; Watson & Ahumada, 1985). Looking exclusively at the TimeSformer model, they first saw that earlier layers (layers 4-6 out of 12) best predicted early

visual regions (V1-V4) while later layers (layers 9-11 out of 12) best predicted later visual regions (EBA, LOC, STS, FFA, and PPA). In early visual regions (V1-V3), the motion-energy model outperformed the TimeSformer model, and in the later visual regions (V4, EBA, LOC, STS, FFA, and PPA), the TimeSformer model was better. However, the combination of both the TimeSformer and motion-energy features was best for all ROIs except for FFA, STS, and PPA.

For more details about the approaches of the top three challenge winners, see the PDFs of their full reports, available with the dataset.