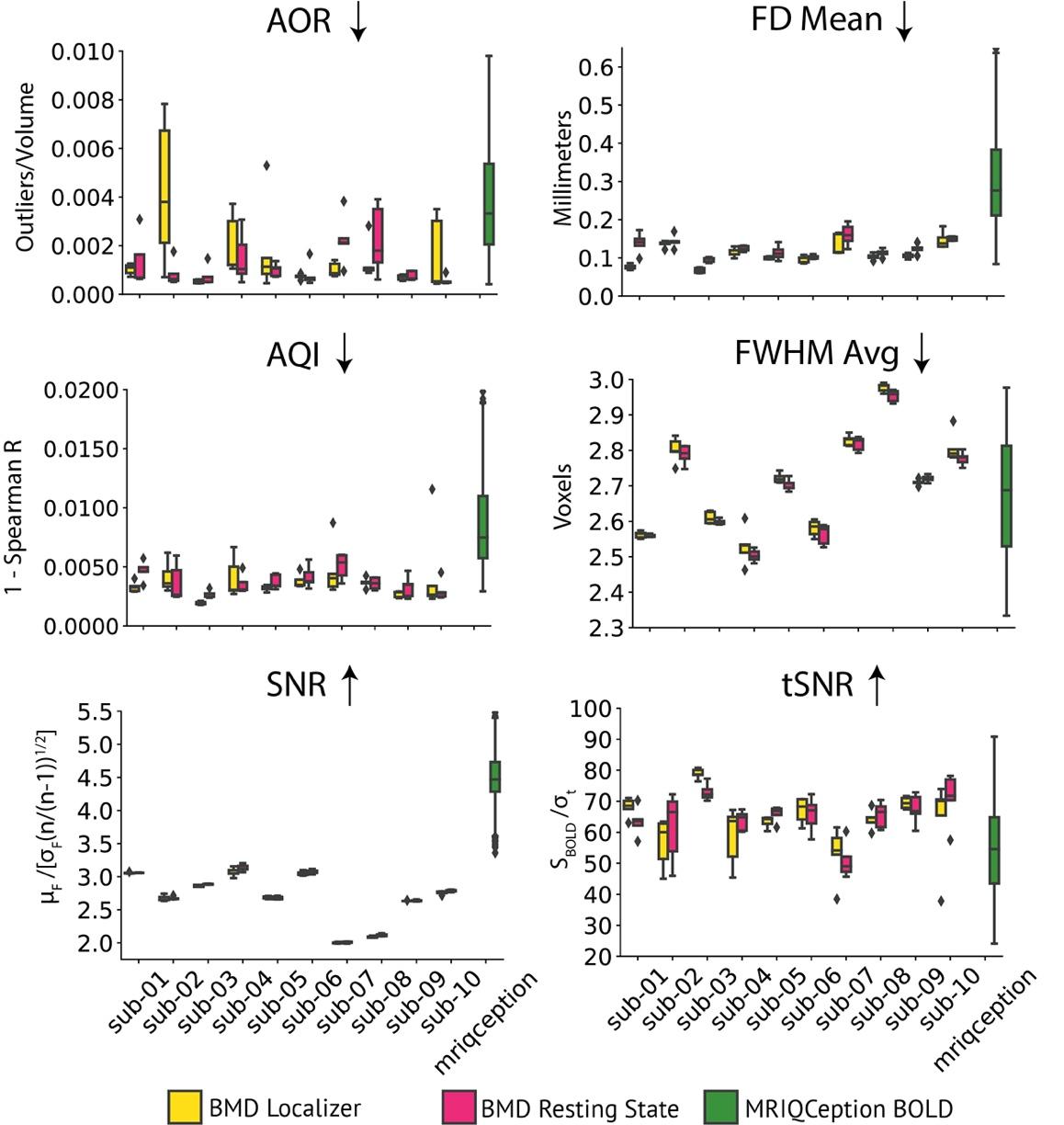
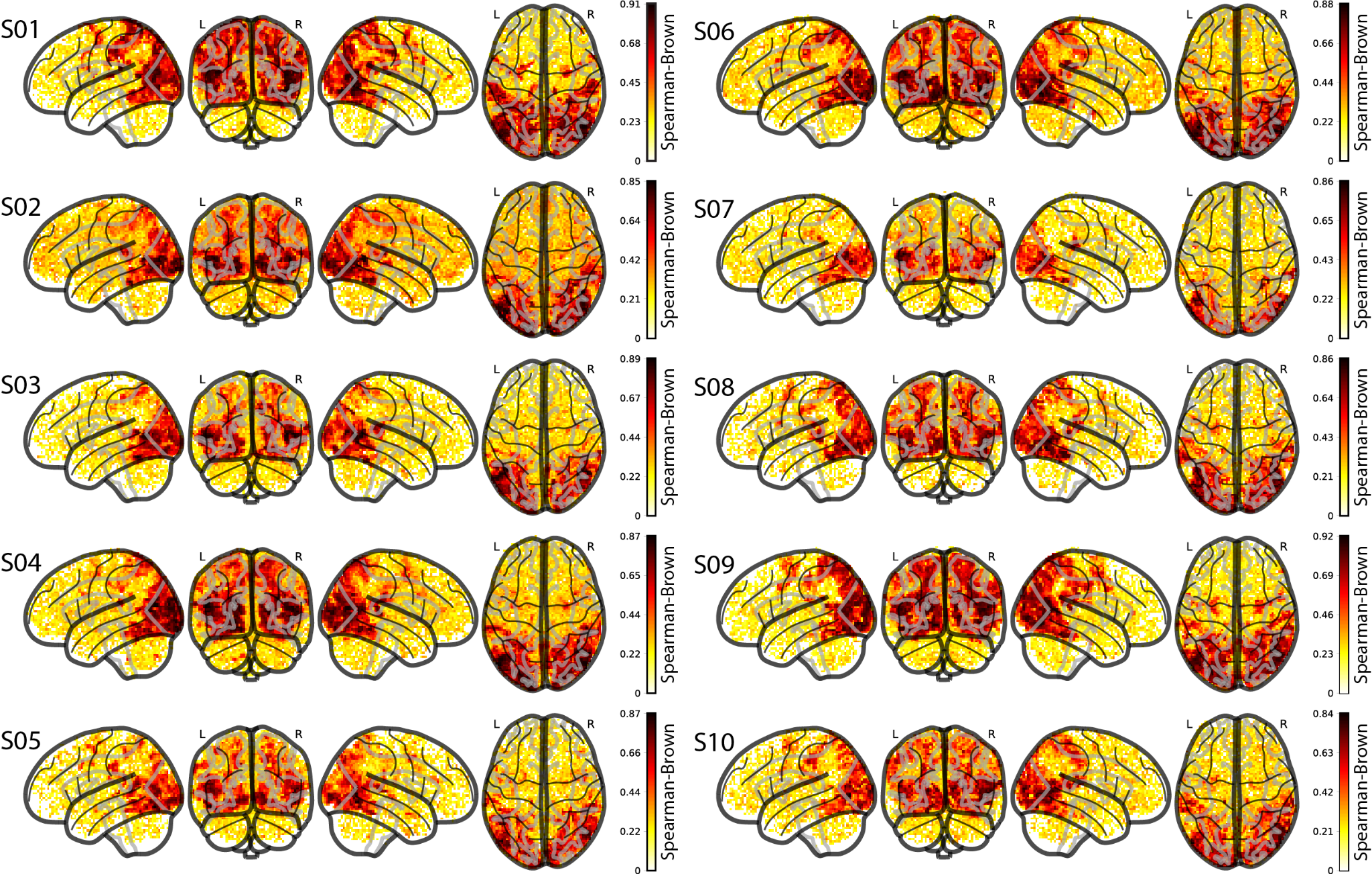
# Supplementary #1 Material



## **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. X-axis labels are shared vertically.



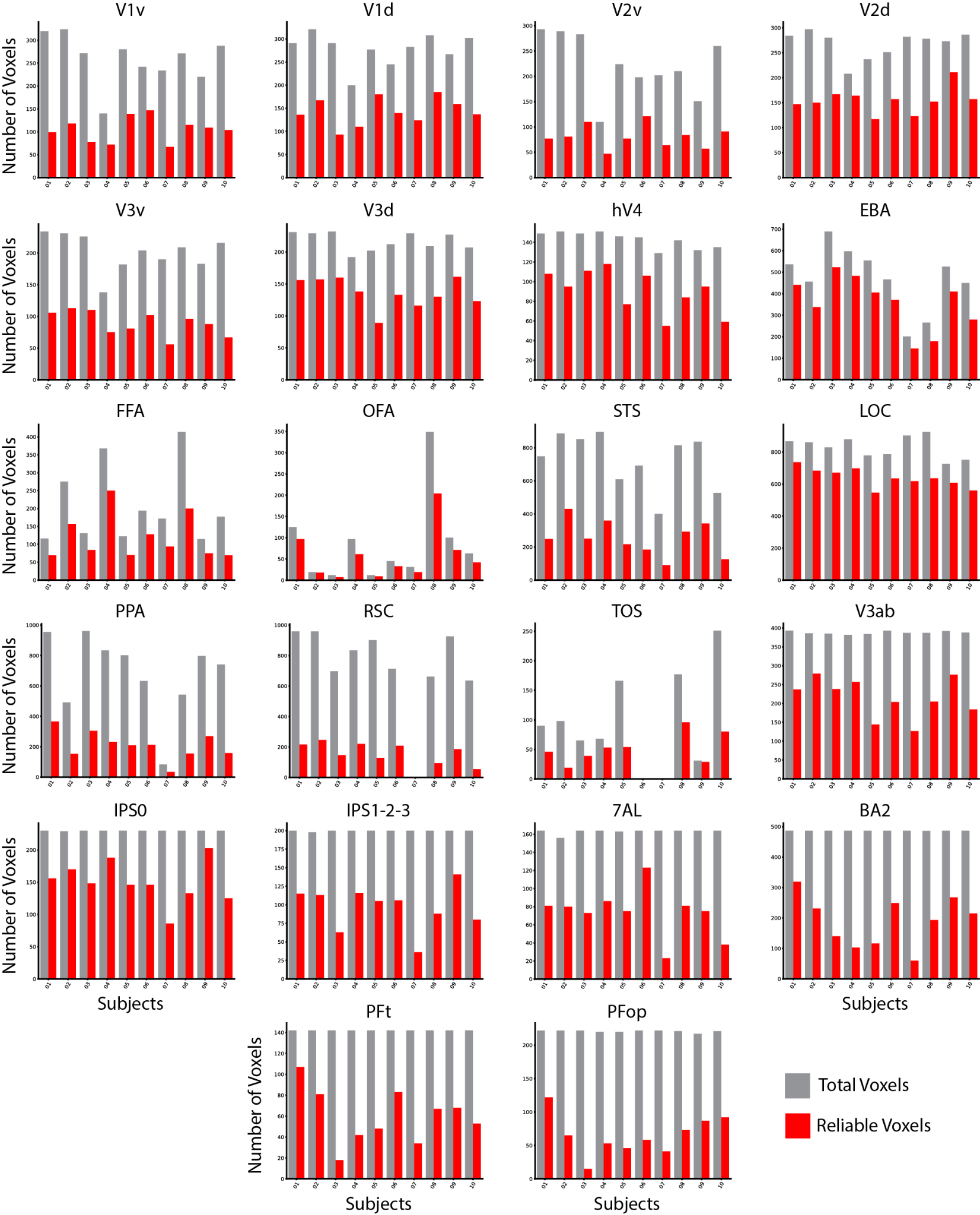
## **Figure S2: Whole-brain split-half reliability for all subjects**

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 () value obtained above in the formula: Spearman Brown = ().

## 

## **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.



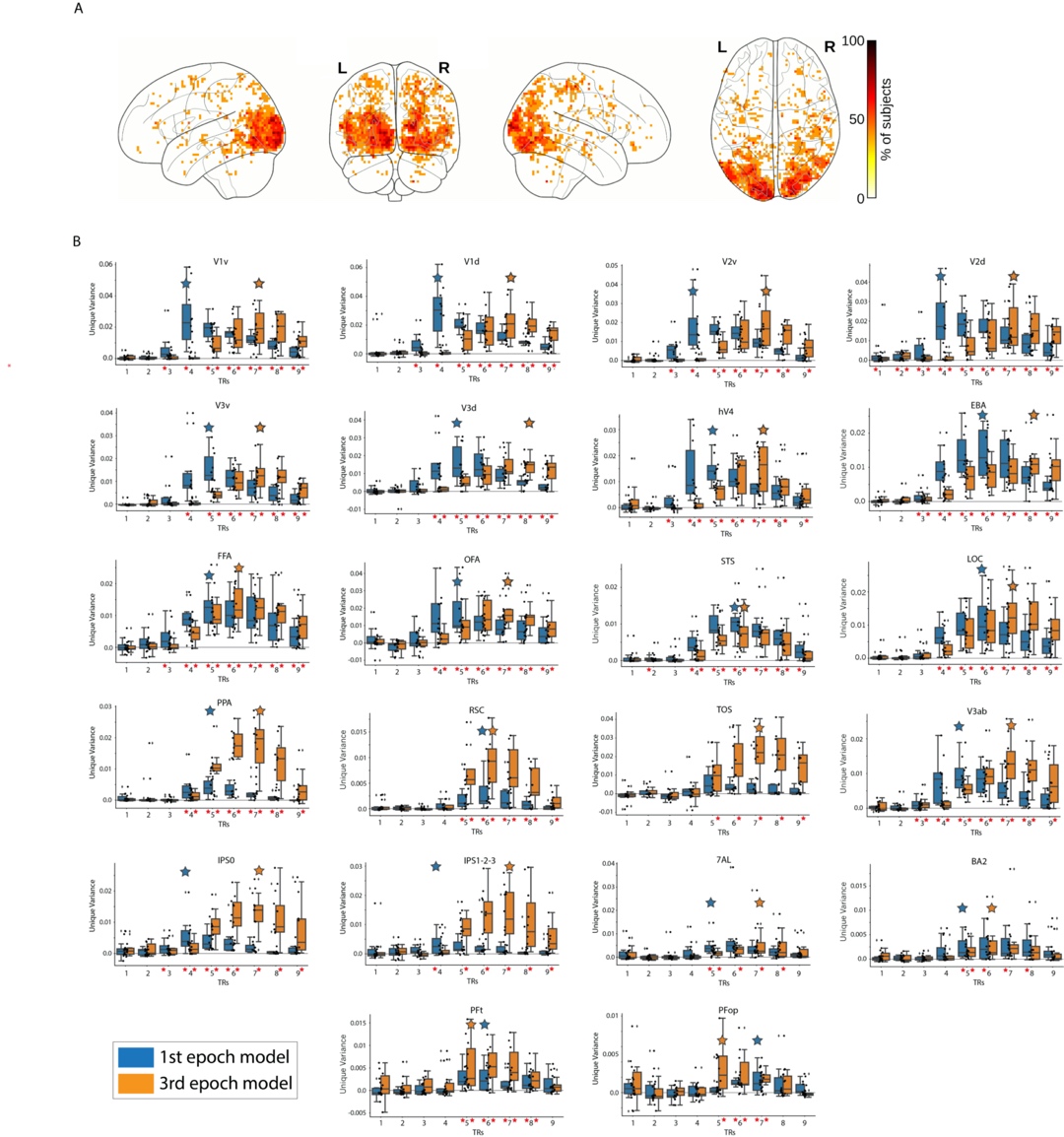
## **Figure S4: ROI reliability and size separated for each subject**

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. Y-axis and X-axis labels are shared horizontally and vertically, respectively.

## 

## **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. Y-axis and X-axis labels are shared horizontally and vertically, respectively.



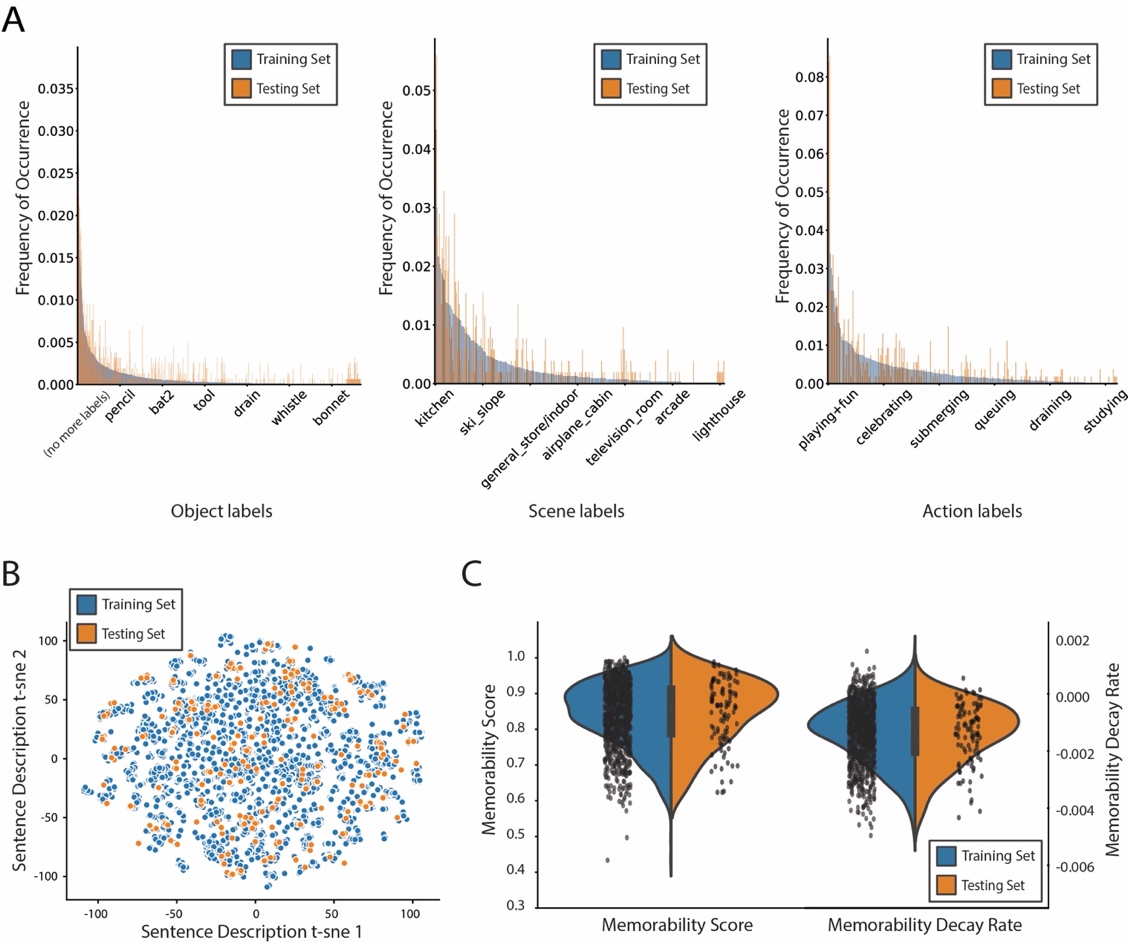
## **Figure S6: Encoding the temporal dynamics of the BOLD signal.**

**(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. Red asterisks along the x-axis indicate unique variance scores significantly greater than 0 (p<0.05, one-sample one-side t-test, FDR corrected across 9 TRs x 2 video epochs = 18 comparisons). Large blue/orange stars indicate the TR with the highest subject averaged unique variance for the first/third video epochs, respectively. The box plot encompasses the first and third data quartiles and the median (horizontal line). The whiskers extend to the minimum and maximum values within 1.5 times the interquartile range, and values falling outside that range are considered outliers (denoted by a diamond). The overlaid points show the value at each observation (n=10 for all ROIs except TOS (n=8) RSC (n=9)). Y-axis and X-axis labels are shared horizontally and vertically, respectively.

## 

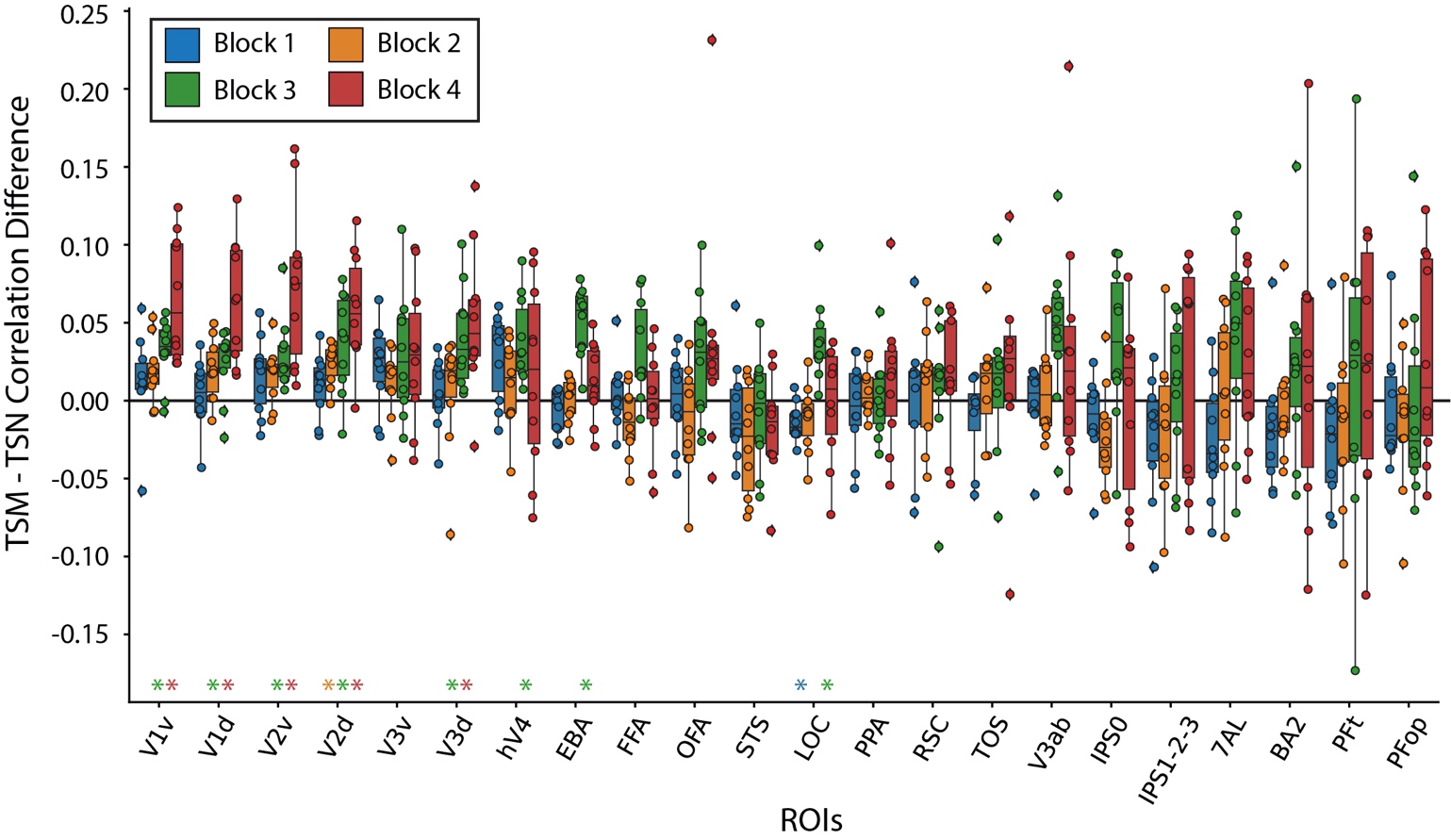
## **Figure S7: Encoding model performance on BMD**

A) TSM ResNet50 trained on M4: Features were extracted after block’s 1 (blue), 2 (orange), 3 (green), and 4 (red) in the ResNet 50 architecture. B) TSM MobileNetV2 trained on Kinetics-400: Features were extracted after the first bottleneck layer (blue), third bottleneck layer (orange), sixth bottleneck layer (green), and last 2D convolutional layer before the average pool (red) in the MobileNetV2 architecture. C) TimeSformer S+T trained on HowTo100M: Of the model’s twelve layers, features were extracted after the first (blue), fourth (orange), eighth (green), and twelfth (red) layers. The box plot on the left side in each panel shows the noise-normalized predictivity of four of each architecture’s features at each of the 22 ROIs. The features were extracted at early (blue), intermediate (orange and green), and late (red) processing stages in each architecture to capture increasingly high-level degrees of transformations. The box plot on the right side in each panel shows the brain prediction difference between each architecture’s latest and earliest layers for each subject and ROI. For the box plots on the right, a blue or red colored box plot denotes a significant difference in correlations from 0 (p<0.05, two-sided one-sample t-test, Bonferroni corrected for n=22 comparisons), and gray denotes no significance.The box plots encompass the first and third data quartiles and the median (horizontal line). The whiskers extend to the minimum and maximum values within 1.5 times the interquartile range, and values falling outside that range are considered outliers (denoted by a diamond). The overlaid points show the value at each observation (n=10 for all ROIs except TOS (n=8) RSC (n=9)).



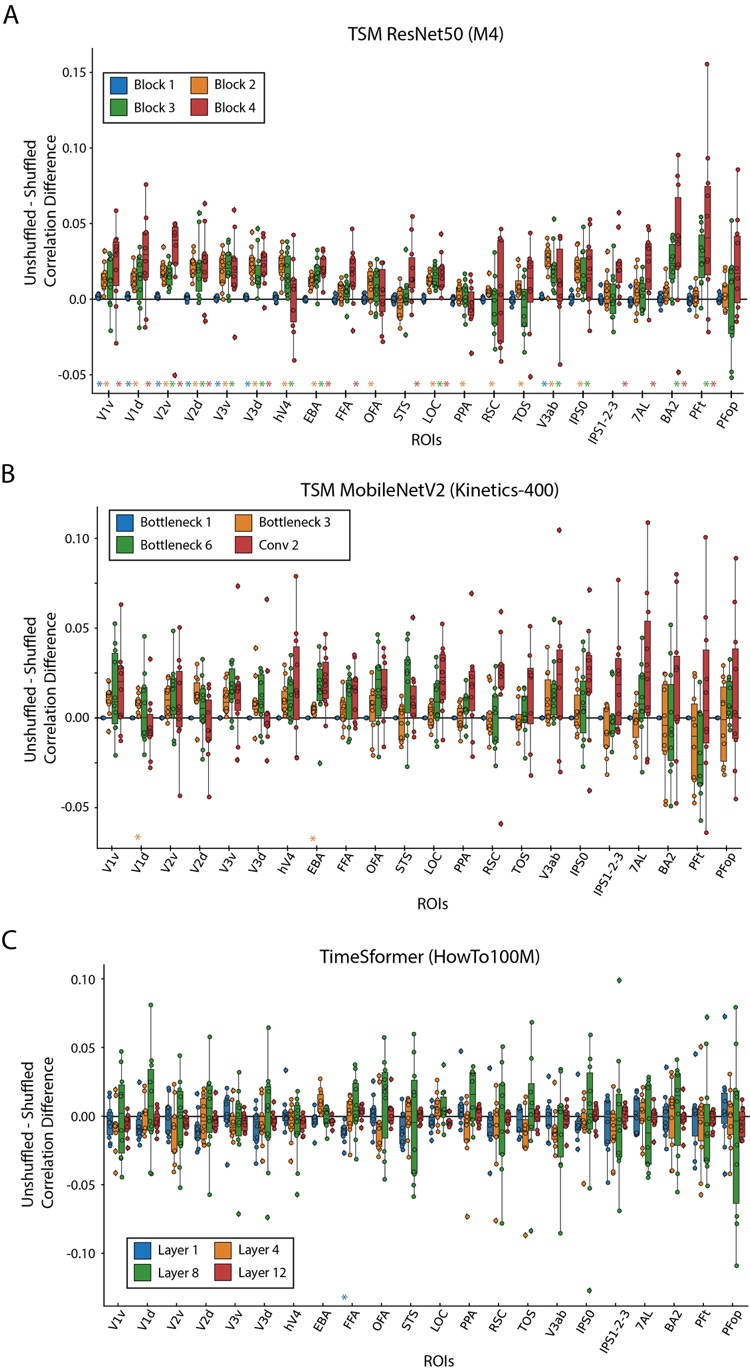
## **Figure S8: Distributions of stimuli metadata between training and testing sets**

The training and testing sets consist of 1,000 and 102 different videos, respectively. An author manually inspected the pairs of testing set videos to ensure no high-level semantic overlap, in terms of objects and actions. **A) Object, scene, and action label frequency of occurrence:** The barplot depicts the frequency of occurrence, (between 0 and 1) of, from left to right, the single-word object, scene, and action labels of the 1,102 video stimuli used in the BOLD Moments Dataset. The frequency bars for each label are separated by training (blue) and testing (orange) splits to show their similar frequency of distributions. **B) Text description and spoken transcription t-sne distances.** The scatterplot shows the t-sne components (n=2 components, perplexity=10, number of iterations=1000) of each text description or spoken transcription embedding. The 6 sentence descriptions per video (5 text descriptions and 1 spoken transcription) serve as a useful proxy for the video’s content. The t-sne plot shows the training and testing set stimuli cover similar spaces of video content. **C) Memorability distribution.** The distribution of the memorability scores and memorability decay rates (1 per video) between the training and testing splits are highly similar and approximately normal. Note that the positive memorability decay rates, while theoretically implausible, reflect the true experimental results detailed in the Memento10k dataset. Users may want to set positive values to 0 depending on the analysis.



**Figure S9: The effect of Temporal Shift Module (TSM) on brain prediction performance**

**(A) TSM vs TSN prediction performance:** The difference in subject brain prediction performance of a TSM ResNet50 and Temporal Segment Network (TSN) ResNet50 each trained on a 10,000-video subset of the M4 dataset (Multi-moments Minus Memento) was computed at each of the four Blocks for each ROI. TSM results in increased brain prediction performance most prominently in early visual ROIs. In both panels, colored asterisks along the x-axis plot 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, FDR correction across 22 ROIs x 4 blocks = 88 comparisons, p < 0.05). The box plot encompasses the first and third data quartiles and the median (horizontal line). The whiskers extend to the minimum and maximum values within 1.5 times the interquartile range, and values falling outside that range are considered outliers (denoted by a diamond). The overlaid points show the value at each observation (n=10 for all ROIs except TOS (n=8) RSC (n=9)).



**Figure S10: The effect of frame shuffling on brain prediction performance across different architectures**

We compute the difference in the correlation between the shuffled frame prediction accuracy and unshuffled frame prediction accuracy at all 22 ROIs and four layers of a **(A)** TSM ResNet50, **(B)** TSM MobileNetV2, and **(C)** TimeSformer model. Features were extracted at increasing levels of depth in each model (blue, orange, green, red) that reflect higher levels of model processing stages. Only the TSM ResNet50 architecture trained on the M4 dataset (Multi-moments Minus Memento10k) showed evidence of robust differences across cortex between shuffled and unshuffled input. Colored asterisks along the x-axis plot 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, FDR correction across 22 ROIs x 4 blocks = 88 comparisons, p < 0.05). The box plot encompasses the first and third data quartiles and the median (horizontal line). The whiskers extend to the minimum and maximum values within 1.5 times the interquartile range, and values falling outside that range are considered outliers (denoted by a diamond). The overlaid points show the value at each observation (n=10 for all ROIs except TOS (n=8) RSC (n=9)).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Original Citation** | **Dataset Name** | **Number of Subjects** | **Number of Unique stimuli**  **(Shared across all subjects)** | **Stimulus Repetitions Within Subject**  **(Stimuli x repetitions)** | **Provided Stimulus Metadata?** | **Stimulus Superset** | **MRI Scanner Strength of Main Task** | **Auxiliary Measurements (in addition to structural)?** | **Other Neuroimaging Modalities** | **Experiment Superset** |
| Ours | BOLD Moments | 10 | 1,102 3 second videos (1102) | 1,000 videos x 3  102 videos x 10 | Yes | Moments in Time  Multi-Moments in Time  Memento10k | 3T | Yes | EEG (in progress) | None |
| (Chang et al., 2019) | BOLD5000 | 4 | 4,916 images (113)a | 4,916 images x 1 | Yes | SUN  COCO  ImageNet | 3T | Yes | None | None |
| (Hanke et al., 2016) | Forrest Gump | 15 | 1 2-hour movie (1) | 1 movie x 1 | Yes | None | 3T | Yes | fMRI (audio-only)c  fMRI (music listening) | StudyForrest |
| (Allen et al., 2022) | NSD | 8 | 70,566 images (515) | 10,000d images x 3 | Yes | COCO | 7T | Yes | None | None |
| (Seeliger et al., 2019) | Doctor Who | 1 | 30 45-minute TV episodes (n/a)  7 1-3 minute video clips (n/a) | 30 episodes x 1  7 video clips x 22 | No | None | 3T | Yes | Nonee | None |
| (Aliko et al., 2020) | Naturalistic Neuroimaging Database (NNDb) | 86 | 10 movies (10)f | 1 movie x 1 | No | None | 1.5T | Yes | Noneg | None |
| (Hebart et al., 2023) | THINGS-data | 3 | 8,740 images (8,740) | 8640 x 1  100 x 12 | Yes | THINGS | 3T | Yes | MEG  EEG | THINGS initiative |
| (Nishimoto et al., 2011) | Vim-2 | 3 | 2 continuous video streams (2) | 1 stream (7200 seconds) x 1  1 stream (540 seconds) x 10 | No | None | 4T | No | None | None |
| (Wen et al., 2018) | n/a | 3 | 2 continuous video streams (2) | 1 374-clip stream (2.4 hours) x 2  1 598-clip stream (40 minute) x 10 | No | None | 3T | No | None | None |
| (Zhou et al., 2023) | Human Action Dataset (HAD) | 30 | 21,600 2 second video clips (0) | 720 videos x 1 | Yes | Human Action Clips and Segments (HACS) Clips | 3T | No | MEG (in progress) | None |
| (Boyle et al., 2020) | Friends s01 - s06 | 6 | 146 22-minute TV episodes (146)h | 1 episode x 1 | No | None | 3T | Yes | None | Courtois NeuroMod |
| (Boyle et al., 2020) | movie10 | 6 | 4 ~1-3 hour movies (4) | 2 movies x 1  2 movies x 2 | No | None | 3T | Yes | None | Courtois NeuroMod |
| (Horikawa & Kamitani, 2017) | Generic Object Decoding (GOD) | 5 | 1,250 images (1,250) | 1200 images x 1  50 images x 35 | Yes | ImageNet | 3T | Yes | fMRI (mental imagery) | None |

a Chang et al., 2019: 4 subjects saw 112 images, 3 subjects saw 1 image

b Chang et al., 2019: Images were sampled from SUN (1000), COCO (2000), and ImageNet (1916)

c Hanke et al., 2016: Audio-visual fMRI was originally acquired

d Allen et al., 2022: Subjects each saw between 9-10,000 images

e Seeliger et al., 2019: Audio-visual fMRI was originally acquired

f Aliko et al., 2020: 8 movies were seen by 6 subjects, 1 movie was seen by 18 subjects, and 1 movie was seen by 20 subjects

g Aliko et al., 2020: Audio-visual fMRI was originally acquired

h Boyle et al., 2020: Subject 04 completed seasons 1-4 and part of season 5

## Table S1: Comparison of large visual fMRI datasets

We compare large, naturalistic, task-based, visual fMRI datasets across features of interest to computational neuroscientists for modeling purposes. The features highlight a dataset’s stimuli (type, number, overlap with existing stimuli sets, and annotations), fMRI acquisition (scanner strength, auxiliary measurements, complementary neuroimaging modalities, and subset in greater neuroimaging efforts), and experimental design (number of subjects and stimulus repetitions). This table summarizes a niche each dataset fills and should not be mistaken for a comparison of dataset quality. Values may differ slightly from the original publication in order to facilitate comparisons across datasets and necessarily summarize information. Note that while some datasets are not officially part of a larger experiment superset, many have been used in independent studies and thus may have additional stimuli metadata and neuroimaging data. Such cases are not noted in this table to maintain clarity. Please see the original publication for the most accurate information.

The added value of brain responses to a short video dataset versus a static image dataset

We emphasize that a short video (e.g., 3 second duration, as in BMD) fMRI dataset is not better or worse than a static image fMRI dataset; rather, they are different in terms of stimulus features and corresponding brain responses that may make one better suited to answer specific research questions. Most obvious, short videos contain a temporal dimension that static images do not, allowing the video to communicate crucial contextual information about how spatial components in our environment move (or not) and spatially relate to each other over time. The benefit of this temporal dimension is clear in our everyday lives – we can interpret transitions between states (a door is being opened, not closed), direction (a steering wheel is being turned to left, not right or still), reactions (the child laughed when shown the picture), motion (the baby is crawling slowly, not fast), and more.

The contextual value of a video’s temporal dimension is reflected in BMD’s own action and sentence text description metadata. Concerning action labels, images can only be labelled with a limited subset of actions or else be highly constrained in order to capture a specific action. For example, the action of a baseball player “hitting” the ball can only be captured with an image if the photo were taken at very specific instant in time, otherwise the action may be ”standing” or “swinging”. Even a short video like in BMD easily captures these actions without heavily constraining the space of possible videos that correspond to “hitting”. Concerning text descriptions, short videos can capture temporal sequences of events that an image cannot. Examples of such captions sampled from some of BMD’s first videos include (emphasis our own):

* Video 0001: "A mallard is in the water alone *swimming around* and *putting its beak in.*"
* Video 002: "A man *is showing another man* how to *move feet back and forth*."
* Video 005: "A woman guides a little boy's arms *up and down as other kids stretch* around him."
* Video 006: "a chess tournament is going on this is focused on two players one *is moving their queen and taking something* to put the king in checkmate"

Static frames of these videos cannot capture the temporal facts that the mallard is “putting its beak in”, the man “is showing another man how”, “a woman guides…as other kids stretch”, and a chess player “is moving their queen and taking something to put the king in checkmate.” This temporal information adds valuable context that often makes one’s understanding of the 3s video vastly different compared to any one of its single static frames.

But do these differences in short videos and static images translate to differences in fMRI brain responses? Yes, previous work has found that videos evoke a greater extent (Bartels & Zeki, 2004; Konen & Kastner, 2008; Press et al., 2001; Schultz & Pilz, 2009; Yildirim et al., 2019) and pattern (Buccino et al., 2004; Kret et al., 2011; Lingnau & Downing, 2015; Wurm & Caramazza, 2022) of cortex responding to videos than images throughout occipitotemporal, dorsal visual, and parietal cortex. In this manuscript we describe our highly reliable activations throughout cortex (Figure 3) with notably high reliability in parietal cortex, a region of the brain that weakly responds to static images. These highly reliable brain responses are not just a result of increased participant engagement or stimulus saliency; we even show that BMD brain responses capture temporal information from the videos (Figure 5, Figure 6, Supplementary 1 Figure 9, Supplementary 2 Figure 2) despite the BOLD response’s temporal sluggishness and fMRI’s low sampling rate.

In the neighboring field of computer vision, researchers have long recognized that videos and images demand different modeling approaches (Ahn et al., 2023; Bertasius et al., 2021; Lin et al., 2019; Tong et al., 2022; Wang et al., 2016) and training datasets (Goyal et al., 2017; Kay et al., 2017; Miech et al., 2019; Monfort et al., 2020; Soomro et al., 2012) for strong task performance. Videos continue to be at the forefront of ground breaking computer vision research due to their creative, cross-domain, and practical applications in text-to-video generation (Ho et al., 2022; Singer et al., 2022; Wu et al., 2023), video understanding with large language models (Ju et al., 2022; Maaz et al., 2023; Zhang et al., 2023), and efficient action recognition and pose estimation (Liu et al., 2023; Qing et al., 2024; Zheng et al., 2023).

Taken together, short video fMRI datasets offer unique opportunities to advance the field of computational neuroscience where static image fMRI datasets cannot. They can advance methodologies around estimating BOLD signals in response to rapid stimulus presentations (Misaki et al., 2013; Prince et al., 2022; Wittkuhn & Schuck, 2021), elucidate cognitive functions concerning temporal integration (Fairhall et al., 2014; Hasson et al., 2008; Orlov & Zohary, 2018), test temporally specific cognitive objective functions (Doerig et al., 2022; Kanwisher et al., 2023), and detail how multiple visual pathways interact to achieve an understanding of an event (Lingnau & Downing, 2015; Mineault et al., 2021; Pitcher & Ungerleider, 2021; Wurm & Caramazza, 2022). As neuroscience and computer science research become increasingly intertwined (Allen et al., 2022; Chen et al., 2023; Cichy et al., 2019, 2021), BMD is well-suited to integrate with state-of-the-art video modeling work from the computer vision community. Importantly, a short video dataset like BMD can make these scientific advancements while staying connected to the vast body of still image work by sharing event-related paradigms, multivariate and univariate methodologies, representational similarity analyses, and/or encoding and decoding techniques. Short video datasets offer more ecological validity than static images while retaining experimental control and offer tremendous potential to advance our understanding of the human visual system.

## 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 (see MRIQC documentation for details on all 112 IQMs:  [https://mriqc.readthedocs.io/en/latest/measures.html](https://www.google.com/url?q=https://mriqc.readthedocs.io/en/latest/measures.html&sa=D&source=docs&ust=1700961676876197&usg=AOvVaw2tB9D6uITvdnuysrYQtTlw)). 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:

Where is the mean intensity of the foreground, is the standard deviation of the foreground intensity, and 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 (|SW - SG|) and divides them by the standard deviation of the values in the surrounding air (σair). 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:

Where is the mean intensity of the foreground, is the standard deviation of the foreground intensity, and 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 summarize the top three challenge entries, highlighting their different modeling approaches and insights at the intersection of natural and artificial intelligence research.

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 (Yang et al., 2021). 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) (Janik & Olesik, 2021). 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” (Nishimoto, 2021) 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 online or with the BMD dataset.

**References**

Ahn, D., Kim, S., Hong, H., & Ko, B. C. (2023). *STAR-Transformer: A Spatio-Temporal Cross Attention Transformer for Human Action Recognition*. 3330–3339. https://openaccess.thecvf.com/content/WACV2023/html/Ahn\_STAR-Transformer\_A\_Spatio-Temporal\_Cross\_Attention\_Transformer\_for\_Human\_Action\_Recognition\_WACV\_2023\_paper.html

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

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

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

Boyle, J. A., Pinsard, B., Boukhdhir, A., Belleville, S., Brambatti, S., Chen, J., & Bellec, P. (2020). The Courtois project on neuronal modelling-first data release. *26th Annual Meeting of the Organization for Human Brain Mapping*.

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.

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

Chen, Z., Qing, J., & Zhou, J. H. (2023). *Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity* (arXiv:2305.11675). arXiv. https://doi.org/10.48550/arXiv.2305.11675

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., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., & Oliva, A. (2019). *The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence*. https://doi.org/10.48550/ARXIV.1905.05675

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

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

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., & Memisevic, R. (2017). *The “Something Something” Video Database for Learning and Evaluating Visual Common Sense*. 5842–5850. https://openaccess.thecvf.com/content\_iccv\_2017/html/Goyal\_The\_Something\_Something\_ICCV\_2017\_paper.html

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

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

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, *12*, e82580. https://doi.org/10.7554/eLife.82580

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). *Imagen Video: High Definition Video Generation with Diffusion Models* (arXiv:2210.02303). arXiv. https://doi.org/10.48550/arXiv.2210.02303

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15037

Janik, R. A., & Olesik, M. A. (2021). *Bionn team solution preliminary write-up*.

Ju, C., Han, T., Zheng, K., Zhang, Y., & Xie, W. (2022). Prompting Visual-Language Models for Efficient Video Understanding. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 105–124). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19833-5\_7

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask ‘why’ questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. https://doi.org/10.1016/j.tins.2022.12.008

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset* (arXiv:1705.06950). arXiv. https://doi.org/10.48550/arXiv.1705.06950

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

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

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

Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in Cognitive Sciences*, *19*(5), 268–277. https://doi.org/10.1016/j.tics.2015.03.006

Liu, D., Li, Q., Dinh, A.-D., Jiang, T., Shah, M., & Xu, C. (2023). *Diffusion Action Segmentation*. 10139–10149. https://openaccess.thecvf.com/content/ICCV2023/html/Liu\_Diffusion\_Action\_Segmentation\_ICCV\_2023\_paper.html

Maaz, M., Rasheed, H., Khan, S., & Khan, F. S. (2023). *Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models* (arXiv:2306.05424). arXiv. https://doi.org/10.48550/arXiv.2306.05424

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*. 2630–2640. https://openaccess.thecvf.com/content\_ICCV\_2019/html/Miech\_HowTo100M\_Learning\_a\_Text-Video\_Embedding\_by\_Watching\_Hundred\_Million\_Narrated\_ICCV\_2019\_paper.html

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

Misaki, M., Luh, W.-M., & Bandettini, P. A. (2013). Accurate decoding of sub-TR timing differences in stimulations of sub-voxel regions from multi-voxel response patterns. *NeuroImage*, *66*, 623–633. https://doi.org/10.1016/j.neuroimage.2012.10.069

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

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

Nishimoto, S. (2021). *Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features* (p. 2021.08.22.457251). bioRxiv. https://doi.org/10.1101/2021.08.22.457251

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

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

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

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

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, *11*, e77599. https://doi.org/10.7554/eLife.77599

Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., & Sang, N. (2024). MAR: Masked Autoencoders for Efficient Action Recognition. *IEEE Transactions on Multimedia*, *26*, 218–233. https://doi.org/10.1109/TMM.2023.3263288

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

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2022). *Make-A-Video: Text-to-Video Generation without Text-Video Data* (arXiv:2209.14792). arXiv. https://doi.org/10.48550/arXiv.2209.14792

Soomro, K., Zamir, A. R., & Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild* (arXiv:1212.0402). arXiv. https://doi.org/10.48550/arXiv.1212.0402

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems*, *35*, 10078–10093.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 20–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-46484-8\_2

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

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, *28*(12), 4136–4160. https://doi.org/10.1093/cercor/bhx268

Wittkuhn, L., & Schuck, N. W. (2021). Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21970-2

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., & Shou, M. Z. (2023). *Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation*. 7623–7633. https://openaccess.thecvf.com/content/ICCV2023/html/Wu\_Tune-A-Video\_One-Shot\_Tuning\_of\_Image\_Diffusion\_Models\_for\_Text-to-Video\_Generation\_ICCV\_2023\_paper.html

Wurm, M. F., & Caramazza, A. (2022). Two ‘what’ pathways for action and object recognition. *Trends in Cognitive Sciences*, *26*(2), 103–116. https://doi.org/10.1016/j.tics.2021.10.003

Yang, H., Zhang, S., Wu, Y., Li, Y., & Gu, S. (2021). *Effective Ensemble of Deep Neural Networks Predicts Neural Responses to Naturalistic Videos* (p. 2021.08.24.457581). bioRxiv. https://doi.org/10.1101/2021.08.24.457581

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, H., Li, X., & Bing, L. (2023). *Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding* (arXiv:2306.02858). arXiv. https://doi.org/10.48550/arXiv.2306.02858

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M. (2023). Deep Learning-based Human Pose Estimation: A Survey. *ACM Computing Surveys*, *56*(1), 11:1-11:37. https://doi.org/10.1145/3603618

Zhou, M., Gong, Z., Dai, Y., Wen, Y., Liu, Y., & Zhen, Z. (2023). A large-scale fMRI dataset for human action recognition. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-023-02325-6