1 **Video-evoked fMRI BOLD responses are highly consistent across different**

2 **data acquisition sites**

3

4 Running title: Cross-site consistency of video fMRI

5

6

7 Lisa Byrge[1], Dorit Kliemann[2], Ye He[3], Hu Cheng[4], J. Michael Tyszka[6], Ralph Adolphs[5-

8 7], Daniel P. Kennedy[4,8-9]

9

10 [1]Department of Psychology, University of North Florida

11 [2]Department of Psychological and Brain Sciences, The University of Iowa

12 [3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

13 [4]Department of Psychological and Brain Sciences, Indiana University

14 [5]Division of Biology and Biological Engineering, California Institute of Technology

15 [6]Division of the Humanities and Social Sciences, California Institute of Technology

16 [7]Chen Neuroscience Institute, California Institute of Technology

17 [8]Cognitive Science Program, Indiana University

18 [9]Program in Neuroscience, Indiana University

19

20 Correspondence: lbyrge@unf.edu

21

9

10   Conflict of Interest: The authors declare no competing financial interests.

11

12   Data availability: The primary data (Video1 and Video2) analyzed for this

13   manuscript is publicly in the National Database for Autism Research (NDAR; Hall et

14   al., 2012; https://nda.nih.gov/about.html).

15

1   **Abstract:** Naturalistic imaging paradigms, in which participants view complex

2   videos in the scanner, are increasingly used in human cognitive neuroscience.

3   Videos evoke temporally synchronized brain responses that are similar across

4   subjects as well as within subjects, but the reproducibility of these brain responses

5   across different data acquisition sites has not yet been quantified. Here we

6   characterize the consistency of brain responses across independent samples of

7   participants viewing the same videos in fMRI scanners at different sites (Indiana

8   University and Caltech). We compared brain responses collected at these different

9   sites for two carefully matched datasets with identical scanner models, acquisition,

10   and preprocessing details, along with a third unmatched dataset in which these

11   details varied. Our overall conclusion is that for matched and unmatched datasets

12   alike, video-evoked brain responses have high consistency across these different

13   sites, both when compared across groups and across pairs of individuals. As one

14   might expect, differences between sites were larger for unmatched datasets than

15   matched datasets. Residual differences between datasets could in part reflect

16   participant-level variability rather than scanner- or data- related effects. Altogether

17   our results indicate promise for the development and, critically, generalization of

18   video fMRI studies of individual differences in healthy and clinical populations alike.

19

20   **Keywords:** video fMRI, naturalistic viewing, reliability, reproducibility, inter-

21   subject correlations, synchrony, harmonization

22

1

Introduction

3   Problems with reproducibility and reliability of scientific findings have

4   arisen across numerous fields over the past two decades (Ioannidis, 2005).

5   Functional magnetic resonance imaging (fMRI) studies have been far from immune,

6   with inconsistent results found across numerous fMRI paradigms (Nickerson, 2018;

7   Poldrack et al., 2017; Zuo et al., 2019; He et al., 2020; Elliott et al., 2020).

8   Inconsistent results could indicate true and potentially relevant differences in study

9   populations. But different data processing and data analysis choices can yield

10 different conclusions from the same datasets (e.g., Eklund et al., 2016; Botvinik-

11 Nezer, et al., 2020), and datasets collected from different scanners at different sites

12 can contain non-biological variability due to differences in scanners and protocols

13 (Friedman et al, 2006; Yu et al., 2018). Altogether these considerations indicate the

14 importance of directly testing reproducibility across datasets collected at different

15 sites.

16   Naturalistic viewing fMRI, or video fMRI (here; vfMRI; Hasson et al., 2004),

17 has emerged in recent years as an attractive alternative to conventional task- and

18 connectivity-based paradigms. Videos are arguably more ecologically valid, and

19 permit greater compliance in the scanner (Eickhoff, Milham, & Vanderwal, 2020;

20 Vanderwal, Eilbott, & Castellanos, 2019) making them an ideal candidate to use for

21 developmental and clinical samples (e.g., Richardson et al., 2019). While vfMRI data

22 can be analyzed using conventional task- and connectivity- based approaches, a

23 distinct analysis approach that is based on measuring similarity or synchrony

1   among participant brain responses has gained prominence (Saarimäki, 2021). This

2   inter-subject correlation-based approach (ISC; Hasson et al., 2004) presents its own

3   distinct analytic requirements due to the dependencies inherent in similarity

4   measurements (Chen et al., 2017; Nastase et al., 2019). Within the same dataset

5   from the same scanner, vfMRI paradigms can evoke markedly similar responses

6   across subjects in many parts of the brain (e.g., Hasson et al., 2004, 2009, 2010;

7   Byrge, Dubois, et al., 2015; Richardson et al., 2018; Nastase et al., 2019). Video-

8   evoked brain responses have also been shown to be reliable within individual

9   subjects after repeated stimulus presentations, in some regions (for review, see

10   Hasson et al., 2010). Reliable responses are observed most consistently throughout

11   posterior swaths of cortex including visual and auditory primary sensory and

12   association areas and, for some video stimuli, can also extend to include parts of

13   default network and lateral prefrontal cortex (Hasson et al., 2009, 2010; Byrge,

14   Dubois, et al., 2015; Burunat et al., 2016). However, the extent to which brain

15   responses during vfMRI are reproducible across different datasets collected at

16   different sites has not yet been examined.

17   This issue of examining reproducibility of vfMRI across different sites takes

18   on increased importance given recent momentum toward using vfMRI for clinical

19   studies (autism: Hasson et al., 2009, Salmi et al., 2013, Byrge, Dubois, et al., 2015;

20   schizophrenia: Yang et al., 2020; depression: Guo et al., 2015, Gruskin et al., 2020).

21   The idea is to first use vfMRI to establish "normative" or "benchmark" patterns of

22   brain responses to a video stimulus with clinically relevant features. This then

23   makes it possible to quantify the extent to which an individual's brain responses

1   deviate from this reference pattern, in some particular brain area(s) or at some

2   particular moment(s) of the video (Hasson et al., 2010; Eickhoff et al., 2020). The

3   hope is that the combination of rich, dynamic stimuli that engage multiple brain

4   networks simultaneously, the relative ease of standardizing stimuli and protocols

5   across different data sites, and the increased data quantity and quality permitted by

6   greater scan compliance might yield insights into the neural basis for the given

7   condition, facilitate discovery of novel biomarkers (Sonkusare et al., 2019; Eickhoff

8   et al., 2020), and eventually inform diagnosis as well as measure efficacy of

9   interventions (Hasson et al., 2010).

10   Many clinical neuroscience studies are moving to multi-site consortiums

11   (e.g., Di Martino et al., 2017, Loth et al., 2017), which aggregate data collection

12   across different sites to obtain an appropriate sample. However, the weak point in

13   clinical neuroscience studies can often be generalization of findings across different

14   studies, samples, and sites (e.g., Kliemann et al., 2018, King et al., 2019; He et al.,

15   2020). This presumably occurs due to combinations of factors that can include

16   individual variability, methodological and stimulus variation, and differences

17   between scanner equipment and standardization. Using video stimuli can minimize

18   methodological and stimulus variation, as noted. But there is considerable

19   individual variation in brain organization and function within the healthy "control"

20   population (Zilles & Amunts, 2013; Dubois & Adolphs, 2016; Holmes & Patrick,

21   2018), including trait-linked variation in video-evoked brain response similarity

22   (e.g., Salmi et al., 2013; Finn et al., 2018). It is therefore important to test the extent

23   to which the "normative" pattern of brain responding to a video is itself

1 reproducible across different sites, before using it as a clinical reference or

2 benchmark. Such an investigation may also provide insights for fMRI harmonization

3 efforts more generally. This is because stimulus-driven brain responses permit

4 partitioning of variance between exogenously- and endogenously- driven brain

5 function in a way that is not possible for some other types of widely-used fMRI

6 paradigms like resting state functional connectivity.

7 Thus, here we directly examine cross-site consistency of evoked brain

8 responses during video scans collected at two different data sites, Indiana

9 University (Indiana) and California Institute of Technology (Caltech) in independent

10 samples of healthy adults. The primary datasets for this manuscript are carefully

11 matched datasets that were collected on different physical scanners in different

12 states, but with potential sources of cross-site variability tightly controlled: identical

13 scanner models, identical scan protocols, identical preprocessing pipelines, and

14 identical analysis procedures (Table 1). Characterizing the similarity of brain

15 responses across these closely matched datasets in a sample of typical controls will

16 thus suggest a potential upper bound on the levels of cross-site consistency to be

17 expected when the same video stimuli are used and other details are matched as

18 closely as possible. As a further exploratory step, we also examined cross-site

19 similarity of brain responses between two unmatched datasets: the Caltech dataset

20 and an earlier pilot dataset (Pilot) also collected at Indiana University, but several

21 years earlier and prior to a scanner upgrade. This Pilot dataset uses the same video

22 stimuli, but different scanner models, different scan protocols, and differences

23 across numerous dimensions of preprocessing approaches (Table 1). Although the

1   unmatched acquisitions were not designed to disentangle specific sources of cross-

2   site variability, we include that comparison as a case study that is informative about

3   the ranges of similarity possible when sources of cross-site variation vary somewhat

4   more freely – as is the case in some multi-site studies, particularly those pooled

5   from pre-existing datasets. Thus, here we map out where in the brain to expect

6   more consistent responses across sites, and conversely, where variability across

7   matched datasets and unmatched datasets most strongly manifests in vfMRI

8   paradigms. This establishes a key foundation for the clinical use of vfMRI, because

9   confidently identifying atypical video-evoked responses in particular brain regions

10   is ultimately limited by the reliability of vfMRI in that region (see also Elliott et al.,

11   2021).

12

13   [INSERT TABLE 1 ABOUT HERE]

14

15

16

17   **Materials & Methods**

18   **Participants.**

19   *Matched datasets.* The primary matched datasets were collected at two sites,

20   Indiana University and Caltech, between 2017 and 2020, as part of a larger project

21   including both typically developed adults and adults with autism spectrum disorder

22   (ASD). Only data from typically developed adults are included in the current report

23   (N = 49/25 (Indiana/Caltech) participants (mean (SD) age 24.9 (6.5) / 34.2 (4.8),

1   from an original sample of N = 63 /29, prior to data-quality-related exclusions

2   reported below). The current dataset includes predominantly males (40 Indiana, 19

3   Caltech) because its primary purpose is to serve as a matched control for the

4   (mostly male) ASD participants whose data will be reported elsewhere. All subjects

5   provided written informed consent; all experimental procedures were approved by

6   the Institutional Review Boards of Indiana University (IU IRB) and the California

7   Institute of Technology.

8

9   *Unmatched (pilot) dataset.* The pilot dataset was collected between 2015-

10   2016 at Indiana University, prior to a scanner upgrade from a 3T Siemens TIM Trio

11   to a 3T Siemens Prisma.Fit system, and is described in Byrge & Kennedy (2020).

12   This dataset also included both typically developed adults and adults with autism

13   spectrum disorder, and is accordingly skewed male. Only data from typically

14   developed adults (N=25, 22 male; mean (SD) age 25.11 (4.66) years) is included in

15   this report. All subjects provided written informed consent; all experimental

16   procedures were approved by the IU IRB.

17

18   [INSERT TABLE 2 ABOUT HERE]

19

20   **Design.**

21   *Matched datasets.* Participants underwent two scanning sessions separated

22   by approximately one week. Each session consisted of interleaved rest and video

23   scans in a fixed order. A total of ten functional scans were collected (six video scans;

1 four ~16-min. resting-state scans). Table 2 presents an overview of the video scans

2 included for each dataset. For this report, we focus most analyses on Video1 and

3 Video2, because they were also used in the pilot dataset. For a few additional

4 analyses, we also include the remaining four video scans. Resting-state scans were

5 used as comparison scans for some analyses. These and the remaining functional

6 scans will be reported in further detail elsewhere.

7   Stimulus construction for the primary video scans (Video1 and Video2) is

8 described in Byrge & Kennedy (2020); briefly, both scans consisted of sequences of

9 4-6 movie trailers collected from Vimeo (https://vimeo.com) across different

10 genres (e.g. documentary, drama, adventure). Video3 and Video4 were different

11 episodes of the TV sitcom "The Office (Season 1 Episode 6, "Hot Girl"; see also Byrge,

12 Dubois, et al., 2015, and Pantelis et al., 2015; and Season 1 Episode 5, "Basketball").

13 Video5 was a short animated movie, Pixar's "Partly Cloudy," (Reher & Sohn, 2009;

14 see also Richardson & Saxe, 2019). Video6 was an edited excerpt from the episode

15 "Bang! You're dead" from the television series Alfred Hitchcock Presents (1961; see

16 also Hasson et al., 2004). Sample sizes for each video scan are reported in Table 2.

17   Video was back-projected onto a screen that was visible to subjects via a

18 mirror attached to the head coil, with audio provided using Sensimetrics MR-

19 compatible headphones. No video stimulus was provided during resting state scans

20 (the projector was set to a black screen), and wakefulness was monitored via an

21 MR-compatible remote eye tracker camera (Eyelink 1000+, SR Research Ltd.

22 Ottawa, Canada). Subjects were instructed to move as little as possible and to

23 remain awake with eyes open. Scans where problems occurred during acquisition

1   (technical problems, such as muffled audio or issues with projector screen, or

2   participants falling asleep) were also excluded (see Table 2).

3   Anatomical images were acquired following functional runs, during which

4   participants chose to rest or watch a different video.

5

6

7   *Unmatched dataset.* The experimental design for this dataset is described in

8   detail in Byrge & Kennedy (2020). Briefly, this study was also collected across two

9   scan sessions separated by approximately one week, with interleaved rest and video

10   scans. Only the two video scans that used the same stimuli as the primary datasets

11   (Video1 and Video2) were included in this report. Anatomical images were collected

12   following functional scans. See Table 2 for sample sizes.

13

14

15   **Data acquisition, preprocessing, and quality assessment.**

16   *Matched dataset*. MRI images were acquired using identical Siemens 3T

17   Magnetom Prisma.Fit scanners (Siemens Medical Solutions, Natick, MA) at each site,

18   with 64-channel head receive arrays. Scan protocols were matched across sites.

19   Scanner software versions used were VE11B (IU) and VE11C (Caltech, and last 5

20   scans at IU). During functional scans, T_2*-weighted multiband echo planar imaging

21   (EPI) data were acquired using the following parameters: TR/TE 720/30 ms; flip

22   angle = 50 °; 2.5mm isotropic voxels; 60 slices acquired in interleaved order

23   covering the entire brain; multi-band acceleration factor of 6 (Multiband EPI

1 sequence version R16, CMRR, University of Minnesota). Scan lengths were as

2 follows: video 1, 1130 volumes; video 2, 1080 volumes; rest, 1355 volumes. Prior to

3 the first functional scan, spin-echo EPI images were acquired in opposite phase-

4 encoding directions (3 images each with P-A and A-P phase encoding) with identical

5 geometry to the EPI data (TR/TE = 4390 / 37.2 ms; flip angle = 90°) to be used as a

6 fieldmap to correct EPI distortions. High-resolution images of the whole brain were

7 acquired as anatomical references (multi-echo MPRAGE, 0.9mm isotropic voxel size;

8 TR = 2550.0 ms / TEs = 1.63 ms, 3.45 ms, 5.27 ms, 7.09 ms / TI = 1150 ms).

9   An upgrade to the trigger box occurred in the final months of data collection

10 at the IU site, and this sporadically resulted in an intermittent missed trigger and

11 delayed movie start for 35 scans. These scans were identified empirically and

12 adjusted accordingly (see Supplemental Methods); these realignments did not

13 influence the pattern of results reported here, which were effectively identical when

14 conducted with the original (non-realigned) scans.

15   DICOM images were converted to BIDS format (Gorgolewski et al., 2016)

16 before being run through MRIQC (v0.15.2; Esteban et al., 2017) for initial quality

17 assessment using the functional image quality metrics (IQMs) FWHM avg, SNR,

18 TSNR, DVARS std, and GSR. Outliers on these IQMs (the median for that data site

19 plus or minus 1.5 times the interquartile range (IQR) for that IQM for that data site,

20 as appropriate for the measure in question) were flagged for manual review by two

21 of the authors (LB & DK). Following review, the consensus decision was to exclude

22 all such flagged scans from further analyses (see Table 2).

1   After initial quality assessment, preprocessing was conducted using

2   fMRIPrep (Esteban, Markiewicz, et al., 2018). The boilerplate text generated by

3   fMRIPrep, with complete preprocessing details, is included in Supplemental

4   Methods. Briefly, using components from ANTs (Avants et al., 2008) FSL (v. 5.0.9;

5   FMRIB's Software Library, *www.fmrib.ox.ac.uk/fsl*) and Freesurfer (v.6.0.1, Dale,

6   Fischl, and Sereno 1999), anatomical images were bias-corrected, skull-stripped,

7   segmented, and nonlinearly registered to MNI space. Functional scans underwent

8   rigid-body motion correction, fieldmap-based distortion correction, and

9   coregistration to the anatomical reference scan, and confound regressors (head

10   motion parameters, CSF, WM, and whole-brain global signal) were computed.

11   For summarizing motion across a scan as well as identifying epochs of

12   excessive motion, we computed filtered framewise displacement traces (FD_{filt4}) from

13   the fMRIPrep-computed head motion parameters, as the sums of the backwards

14   difference across 4 TRs of motion parameters that had been filtered to exclude

15   respiratory frequencies, as introduced by Power and colleagues (2019) and used

16   previously for the pilot acquisition (Byrge & Kennedy, 2020). FD_{filt4} separates head

17   motions from respiratory fluctuations in multiband acquisitions more effectively

18   than the conventional framewise displacement computations (Power et al., 2019).

19   We excluded all scans with excessive motion, as identified by mean FD_{filt4} exceeding

20   the median plus 1.5 times the IQR of the mean FD_{filt4} across all scans (including

21   scans from ASD participants not included in the current analyses), computed

22   separately at each site, resulting in the following exclusion thresholds: mean FD_{filt4} >

23   0.4808 for Indiana, mean FD_{filt4} > 0.5625 for Caltech (see Table 2). To ensure highest

1   data quality, we also censored time points surrounding excessive motions: 10

2   frames before and 30 frames after any frame with $FD_{filt4} > 3.75$mm; censored time

3   points were treated as missing data in all analyses, and the inclusion or exclusion of

4   censored data did not influence the overall pattern of results.

5   All reports generated by fMRIPrep were inspected by two independent

6   reviewers (two research assistants, one based at each site, trained to conservatively

7   flag any potential issues with anatomical and functional scans and their alignment).

8   All reports flagged by both research assistants were then independently reviewed

9   by both LB & DK and a consensus decision was reached about whether to include or

10 exclude all such flagged scans from the current dataset (see Table 2).

11   Subsequent preprocessing used *xcepengine* version 1.2.1; detailed in detail

12 by Ciric and colleagues (2018). We used the "fc-24p_gsr" pipeline optimized for

13 functional connectivity processing; this configuration is publicly available at

14 https://github.com/PennBBL/xcpEngine/blob/master/designs/fc-24p_gsr.dsn.

15 Briefly, functional data was demeaned and detrended, aligned to the anatomical

16 reference scan, and bandpass-filtered within the range 0.08-0.001 Hz using a

17 Butterworth filter. Then, 36 confound regressors (6 head motion parameters, CSF,

18 WM, global signal, their backwards differences, and then the squares of those 18

19 traces; all temporally filtered in the same way as the data) were regressed from the

20 data, and then the residuals were spatially smoothed with a 2.54mm filter, and then

21 used as the "cleaned" data.

22   For the primary datasets, we examined average BOLD timeseries across

23 several different atlases. We focused exclusively on cerebral cortex, excluding the

1 cerebellum and all subcortical structures, following the largely cortico-centric focus

2 of the inter-subject synchrony literature. The primary atlases used were different

3 parcellation scales of the Schaefer atlas (Schaefer et al., 2018), which subdivides the

4 intrinsic functional connectivity-based Yeo network parcellation of the cortex (Yeo

5 et al., 2011) into 100, 200, 400, 600, 800, and 1000 cortical regions. We also

6 examined the structural Harvard-Oxford Atlas distributed with FSL, which had been

7 previously used to parcellate the pilot dataset. We restricted our analysis of the

8 Harvard-Oxford parcellation to cortical regions of interest (ROIs; 96) only, for

9 consistency with the Schaefer cortical parcellation. For all atlases, we obtained ROI

10 timeseries for each region as the mean of the "cleaned" BOLD signal across all voxels

11 in the given region, at each time point.

12 As an additional data quality assessment, for all video scans, we examined

13 BOLD time series from primary visual cortex and primary auditory cortex in each

14 hemisphere (using the Harvard-Oxford parcellation), in order to identify and

15 exclude scans where technical problems with the stimulus presentation or visual or

16 auditory aspects of the stimulus occurred but were not noted at the time of scanning

17 (e.g., headphone or projector failure, or misalignment between the start of image

18 acquisition and the video). We approached this cautiously and conservatively,

19 because similarity of BOLD time series among scans is also our measure of interest

20 for this report; but, at the same time, extremely low similarity to other participant

21 timeseries in primary sensory areas during long video scans is an indicator that

22 something has gone wrong in the scan acquisition process. Therefore, separately

23 within each dataset, for each video scan, and for each of the four primary sensory

1 regions of interest, we computed pairwise correlations among all participant time

2 series, and computed the median minus three times the interquartile range of

3 median pairwise correlations for each participant as a threshold to identify extreme

4 outlier values suggestive of equipment issues. We excluded scans for which the

5 median pairwise correlation was below this data-driven threshold in at least one of

6 the sensory regions (see Table 2).

7

8 *Exploratory pilot dataset*. This acquisition and preprocessing pipeline is

9 described more completely in Byrge & Kennedy (2020). Briefly, images were

10 collected using a 3 T Magnetom Tim Trio system (Siemens Medical Solutions, Natick,

11 MA) with 32-channel head receive array, running software version VB17. T_2*-

12 weighted multiband EPI data was acquired using the following parameters: TR/TE =

13 813/28 ms; 1200 volumes; flip angle = 60°; 3.4 mm isotropic voxels; 42 slices

14 acquired with interleaved order covering the whole brain; multi-band acceleration

15 factor of 3. Gradient-echo EPI images (10 images each with P-A and A-P phase

16 encoding; TR/TE = 1175/39.2 ms, flip angle = 60°) were used as fieldmaps for EPI

17 distortion correction. High-resolution T_1-weighted images of the whole brain

18 (MPRAGE, .7 mm isotropic voxel size; TR/TE/TI = 2499/2.3/1000 ms) were

19 acquired as anatomical references.

20 Data were preprocessed using an in-house pipeline using FSL (v. 5.0.8;

21 FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl), ANTs (v2.1.0; Avants et

22 al., 2011), and Matlab_R2014b (www.mathworks.com, Natick, MA, USA).

23 Preprocessing steps included rigid-body motion correction, fieldmap-based

1   geometric distortion correction, non-brain removal, weak highpass temporal

2   filtering (>2000s FWHM) to remove slow drifts. Denoising was preformed using

3   FSL-FIX (Salimi-Khorshidi et al., 2014) followed by mean cortical signal regression

4   in a second step (effectively the same as global signal regression, but using the

5   signal across the cortex rather than whole brain; Burgess et al., 2016), with the

6   residuals analyzed as the "cleaned" data. Volumetric registrations were conducted

7   using FSL and ANTs, using a combined affine and diffeomorphic transformation

8   matrix. Region of interest (ROI) timeseries using the Harvard-Oxford Atlas

9   distributed with FSL were obtained as the weighted mean signal from the "cleaned"

10   BOLD signal across voxels within each of the 110 ROIs.

11   As these data were collected using a different repetition time (TR) than the

12   primary dataset (813 ms vs. 720 ms), the final preprocessing step for this report

13   was to resample these time series to match the faster sampling rate of the primary

14   dataset. There are many possible ways to perform such resampling; here, we used

15   Fourier method resampling as implemented in scipy.signal.resample.

16   Differences between the primary and exploratory pilot unmatched

17   acquisitions are summarized in Table 1.

18

19   [INSERT FIGURE 1 ABOUT HERE]

20

21   **Data Analysis.**

22   Naturalistic fMRI data analysis requires evaluating the similarity of brain

23   response time series. Here, we examined similarity of brain responses across sites at

1   two distinct levels: similarity of group-average time series from each site (Fig. 1A),

2   and similarity of pairs of individual subject time series across sites and within each

3   site (Fig. 1B; pairwise inter-subject correlations; ISC; Hasson et al., 2004). At the

4   group level, within each site, we used median time series across subjects within

5   each brain region of interest to isolate the common brain response pattern while

6   reducing the influence of various forms of noise. We take these measurements of

7   across-site similarity at both levels as our measures of cross-site consistency.

8   Unless noted, analyses are repeated on the two primary video trailer scans

9   (Video 1 and Video 2) that were used as stimuli in each of the datasets. Most

10   analyses are conducted across multiple spatial scales using different granularities of

11   the Schaefer parcellation (Schaefer et al., 2018).

12

13   *Statistical comparisons*.

14   *Consistency of brain responses across sites*.

15   As a first broad characterization of the extent to which there is shared signal

16   across sites, we measured similarity between median time series for each site using

17   Pearson correlations. We considered a brain region (ROI) to be responding more

18   consistently across sites during the same video than expected by chance if the

19   correlation was statistically significant following FDR-correction for the granularity

20   of the parcellation in question. (Note that the patterns of results were effectively the

21   same when using a non-parametric, bootstrapped procedure).

22   For individual-level analyses, for each scan and in each brain region, we

23   computed pairwise inter-subject correlations (ISC) as the Pearson correlation

1   among pairs of individual brain responses for all pairs of subjects. We also

2   computed pairwise similarity across different scans (e.g., between Video1 and Rest)

3   for use in null models as described below. We used non-parametric statistical

4   comparisons for this level of analysis, following the recommendations of Chen and

5   colleagues (2016) for pairwise ISC, and pooled pairwise correlations within and/or

6   across sites without collapsing at the individual level.

7   To evaluate whether a brain region responded more consistently across sites

8   than expected by chance at the individual level, we examined only inter-subject

9   correlations between pairs of participants from different sites, and asked whether

10   the magnitudes of those correlations were greater when both participants were

11   watching the same video than when one participant was watching the video and the

12   other participant was resting. We compared correlation magnitudes (absolute

13   values) to avoid exaggerated influence from very small negative correlations, which

14   were expected between video and rest scans. We held the comparison scans fixed at

15   each site, yielding two observed median correlation differences. Specifically, within

16   each brain region, for video scan V and rest scan R, across all across-site pairs of

17   Indiana participants Ii and Caltech participants Cj, we obtained: $\Delta r_{Indiana}$= *median*(|

18    $r(V_{I1},V_{C1})$ |, ..., | $r(V_{Ii},V_{Cj})$ |) – *median*(| $r(V_{I1}, R_{C1})$ | , ..., | $r(V_{Ii}, R_{Cj})$ |) and $\Delta r_{Caltech}$ =

19   *median*(| $r(V_{I1},V_{C1})$ | , ..., | $r(V_{Ii},V_{Cj})$ |) – *median*(| $r(R_{I1}, V_{C1})$ | , ..., | $r(R_{Ii}, V_{Cj})$ |). We

20   then permuted scan type labels (Video vs Rest) 10000 times and computed this

21   same measure to establish a null distribution of median differences, and obtained a

22   one-sided empirical *p*-value (corrected to avoid bias due to finite sampling, Davison

23   & Hinkley, 1997) of observing $\Delta r_{Indiana}$ and $\Delta r_{Caltech}$ by chance. To address multiple

1 comparisons, we applied FDR correction within each parcellation. We

2 conservatively considered a region as responding more consistently than expected

3 by chance only if it survived FDR correction for both $\Delta r_{Indiana}$ and $\Delta r_{Caltech}$ (note also

4 that results were effectively the same for both cases).

5

6 *Differences between sites.*

7 After examining consistency of brain responses between sites, we examined

8 differences. To do this, and to get estimates of variance, we continued at the

9 individual subject level, because group average time series mitigate or even

10 eliminate noise that might be unique to a given site or scanner. Differences between

11 datasets would manifest as differences in within-site similarity vs. across-site

12 similarity. Therefore, for each video, site, and brain region, we computed the

13 observed difference in within vs. across-site ISC as the difference between the

14 median ISC among all subjects at the site in question and the median ISC among all

15 different-site subject pairs. As before, we computed this measure separately for each

16 site, because levels of within-site similarity could be different. We then computed

17 the empirical p-value of observing this median difference as before using a

18 permuted null distribution constructed by shuffling site labels 10000 times and

19 computing the permuted median difference. For this analysis, we conservatively

20 used $\alpha = 0.05$ with no correction for multiple comparisons, in order to increase our

21 sensitivity to detect potential differences between datasets (at the expense of likely

22 false positives). Finally, to contextualize the magnitudes of differences between

23 datasets, we also compared distributions of within- and across- site ISC values using

1   Mann-Whitney rank sum tests and report the common-language effect sizes (CLES;

2   Vargha & Delaney, 2000) across ROIs. Common-language effect sizes in the *within –*

3   *across* direction reported here reflect the proportion of pairs of observations in

4   which within-site pairwise ISC is higher than across-site pairwise ISC. CLES of 0.5

5   indicates no effect, and CLES of 0.56, 0.64, and 0.71 roughly correspond with

6   Cohen's *d* values of 0.2, 0.5, and 0.8, indicating small, medium, and large effect sizes

7   (Ruscio, 2008). (Note that CLES below 0.5 would indicate higher across-site ISC than

8   within-site ISC with comparable interpretations, e.g. CLES of 0.44 would reflect a

9   small effect.)

10

11   **Results**

12

13

14   **1.   SIMILARITY.**

15   **1.1.   Consistent group-level brain responses are evoked by the same**

16   **videos at different sites.**

17   Average brain response time series across a group of participants should

18   capture common patterns of stimulus-evoked brain function while mitigating the

19   effects of physiological noise, scanner noise, and individual differences in brain

20   functioning. The hope, for generalizability of vfMRI studies, would be that these

21   group-level brain responses would be largely similar across sites (as depicted in Fig.

22   1A), especially when acquisitions and processing are matched as closely as they are

23   in these primary datasets. Figure 2 shows correlations between the median time

1   series across IU participants and the median time series across Caltech participants

2   in each brain region while watching the videos, under different spatial scales of the

3   Schaefer atlas (Schaefer et al., 2018), which subdivides the intrinsic connectivity-

4   based Yeo network parcellation of the cortex into 100, 200, 400, 600, 800, and 1000

5   regions. As is evident, average brain responses across Indiana participants were

6   highly similar to average brain responses across Caltech participants, during both

7   video scans, across all parcellation scales examined. Highly similar brain responses

8   were not limited to the primary sensory areas expected to be driven by the stimulus

9   (c.f. visual and auditory timeseries shown in Fig. 2) but extended throughout the

10 cortex (c.f. association timeseries shown in Fig. 2).

11

12 [INSERT FIGURE 2 ABOUT HERE]

13

14   **1.2.   Consistent group-level brain responses are found through most of**

15   **the cortex.**

16   As is evident in Figure 2, highly similar group-level brain responses across

17 scanners were not limited to the coarser parcellations. As parcellation granularity

18 increases, though, some correlation magnitudes decrease – as expected, as regional

19 timeseries approach voxel-level timeseries with correspondingly reduced spatial

20 smoothing – to the extent that it becomes unclear by eye whether brain response

21 similarity across scanners exceeds chance levels in some brain regions. In all

22 parcellations, the median time series at each site was more correlated than expected

23 by chance in more than 99% of brain regions. However, the size of these effects

1   varied across the brain, as can be seen on the color axis, and in some association

2   areas, significant correlations between group-level brain responses at each site

3   were quite small (around r = 0.1, after FDR correction). One might expect such a

4   weak shared signal to be easily dominated by other factors (e.g. endogenous

5   processing, scanner noise, registration inaccuracies) if not for averaging across

6   multiple scans (i.e., participants).

7   Brain regions that did not respond consistently across sites at the group level

8   were detected only in the finer parcellations and included parts of the temporal pole

9   (an area prone to susceptibility artifacts) as well as part of medial prefrontal cortex

10   during one scan.

11

12   **1.3.   Participant-level brain responses are consistent across sites**

13   **throughout most of the cortex.**

14   Average time series across groups of subjects are effective at isolating

15   common response patterns by dampening down individual variability, but they are

16   not representative of any individual brain's functioning and could minimize

17   potential differences in noise properties across scanners. Thus, next we examined

18   consistency of individual brain responses among pairs of participants as they

19   watched the same videos (as depicted in Fig. 1B). Figure 3 (see also Supplemental

20   Figure 1) shows the median of these pairwise inter-subject correlations across sites

21   in the center column, along with pairwise ISCs within each site (left and right

22   columns) for comparison. As expected, based on increased noise and individual

23   variability in participant time series, the range of correlations is shifted lower than

1   in Figure 2, but the general conclusion remains the same: consistent brain responses

2   across sites are widespread throughout the majority of the cortex. Nearly all ROIs

3   (>90% in all six cortical parcellations; ranging from 98% of ROIs in the 100-ROI

4   parcellation and 91% in the 1000-ROI parcellation) were more similar between

5   cross-site pairs of subjects watching the same video than expected by chance, using

6   a null distribution comprised of pairwise brain response similarity in which one

7   participant watches this same video and the other undergoes a resting state scan.

8

9   [INSERT FIGURE 3 ABOUT HERE]

10

11   It is important to note, though, that above-chance similarity relative to

12   resting state does not imply a large effect size, and the median across-site pairwise

13   ISCs in some ROIs that responded at above-chance levels could be exceedingly small,

14   even below 0.01. In other words, although the common stimulus explained some

15   proportion of variance in these time series (and a vanishingly small proportion for

16   these smallest correlations), most variability is left unexplained – and thus left to be

17   explained in future studies of stimulus-level, contextual, state-level, physiological,

18   and phenotypic factors underlying these individual brain responses (see e.g., Chang

19   et al., 2021). This is the case even for the primary sensory areas most strongly

20   driven by the stimulus, where median across-site pairwise correlations could

21   exceed 0.5, which still leaves around 75% of the variance unexplained by the

22   common stimulus. Inspection of the randomly selected example time series in

23   Figure 3 (bottom) and Supplemental Figure 1 suggests the possibility that some

1   specific moments of the stimulus may drive relatively instantaneous similarity amid

2   otherwise dissimilar brain responses in some of the association areas that have low

3   ISC that nonetheless exceeds chance (c.f. Figure 3, PFCd), but future work will be

4   needed to examine that possibility directly.

5

6   **1.4.   Consistent group- and individual- level brain responses are evoked**

7   **by a variety of video stimuli.**

8   Participants in the primary dataset watched different video stimuli sampling

9   a variety of genres: movie trailers (Videos 1 and 2; the main stimulus throughout

10   this manuscript), complete episodes of TV sitcoms ("The Office"; Videos 3 and 4), an

11   animated short film ("Partly Cloudy"; Video 5), and a black-and-white Alfred

12   Hitchcock film ("Bang! You're dead"; Video 6). As is already apparent in Figure 2,

13   cross-site consistency at the group level was high for both Video1 and Video2

14   despite varying stimulus content – each of those videos consisted of sequences of

15   trailers for entirely different movies. Figure 4 (top) shows the correlations between

16   group-average timeseries at both sites for all six video scans in the primary dataset,

17   using the coarsest and finest parcellation scales. As is evident, high cross-site

18   similarity between average time series is a feature of all the video stimuli used, and

19   not limited to movie trailers alone. Note that while the maps look similar –

20   qualitatively, regions that responded highly consistently across scanners during one

21   video also responded consistently during the other video – the patterns are also not

22   identical across different videos. These differences could reflect differences in

23   stimulus content. For instance, reductions in cross-site similarity can be observed

1   in some temporal and frontal regions during the largely silent animated film

2   (Video5). Notably, episodes of The Office were chosen so as to emphasize social

3   features in the video, and cross-site consistency in medial prefrontal cortex appears

4   elevated in Videos 3 and 4 relative to the other scans, potentially reflecting the

5   increased social processing demands of the stimulus. While these video scans also

6   differed in length, scan length did not appear to be the main driver of these

7   differences (see also Supplemental Figure 2, which presents an alternate version of

8   this figure that was randomly downsampled to address length differences, but

9   shows largely similar patterns).

10

11   [INSERT FIGURE 4 ABOUT HERE]

12

13   Figure 3 (top) and Supplemental Figure 1 showed high levels of cross-site

14   consistency at the level of individual subject pairs for Videos 1 and 2. Figure 4

15   (bottom) shows median across-site (in black) and within-site (in color) pairwise

16   inter-subject correlations for each of the video scans in each ROI. One of the colored

17   lines corresponds to the values that would be projected onto a brain map if the data

18   had been collected at a single site (e.g., the red line for Video1 in Figure 4 (bottom)

19   is the same as the Indiana within-site ISC map in Figure 3, upper left). While the

20   colored lines show small intermittent deviations above and below the black line, the

21   larger take-away is that the lines track one another closely. The pattern of median

22   within-site ISC across ROIs for one site is highly correlated with the pattern of

23   within-site ISC for the other site, and both quantities are highly correlated with

1   median across-site ISC (all r > 0.91 across all videos and all parcellations). In other

2   words, brain responses for pairs of subjects at different sites are about as similar as

3   pairs of subjects at the same site. This close tracking is apparent both for brain

4   regions that are more evoked and less evoked by the stimuli, as well as for different

5   video stimuli that drive higher and lower ISC values in the same brain regions (for

6   instance, c.f. x = 41 (part of left hemisphere temporal lobe) for Video5 vs. the other

7   video scans, for the center panel with the 100-scale parcellation). Cross-site

8   consistency of brain responses for this set of stimuli is neither limited to a few

9   sensory regions that are most strongly driven by the video, nor a subset of video

10   stimuli that drive the brain especially strongly, but is instead apparent throughout

11   the different stimuli used here.

12

13   **2.   DIFFERENCES:**

14   **2.1.   Differences across sites are minimal when acquisitions and**

15   **processing are matched.**

16   After establishing that brain responses across the cortex are indeed

17   consistent across sites, the natural question is to ask about differences. If there were

18   no differences between datasets, an individual scan would be just as similar to other

19   scans at the same acquisition site as it is to other scans from a different acquisition

20   site. Arguably, site-level differences could manifest as either increased or decreased

21   similarity among participants at the same site, depending on noise properties. We

22   thus evaluated potential site differences by testing whether within-site pairwise

23   similarity for either site differed from across-site pairwise similarity in any brain

1   region, by comparing the observed differences of the medians to a permuted null

2   distribution in which site labels were shuffled. Because levels of within-site

3   consistency need not be the same for both sites, and therefore differences between

4   within-site and across-site similarity could differ, we considered a brain region to

5   have a site difference if such a difference was observed for either site, and not

6   necessarily for both sites. For this analysis, to conservatively increase sensitivity for

7   detecting any potential differences, we did not correct for multiple comparisons,

8   and thus some false positives are likely.

9   For all video scans, and for all parcellations, the majority of brain regions had

10   no site differences at this conservative threshold. Proportions of brain regions that

11   did have site differences ranged (across parcellation scales) as follows: Video1, 0.13-

12   0.17; Video2, 0.09-0.15; Video3, 0.06-0.11; Video4, 0.06-0.13; Video5, 0.23-0.32;

13   Video6, 0.35-0.47. As noted above, these proportions are likely to be an

14   overestimate. Differences can be observed in Figure 4 (bottom) as gaps between the

15   black and colored lines. Differences are generally small relative to the level of ISC,

16   and are found in regions including those that are strongly driven by the stimulus

17   (e.g., $x = 53$, part of the right hemisphere peripheral visual network, for the center

18   panel with the 100-scale parcellation). Supplemental Table 2 summarizes

19   differences in the distributions of within-site and across-site ISC values. For all

20   videos and all parcellations, median differences between within- and across- site ISC

21   values across ROIs were small (<0.03), with median common-language effect sizes

22   (CLES) across ROIs corresponding with small effect sizes. Maximum CLES across

23   ROIs reflected small or medium effects depending on the video and parcellation

1   (maximum CLES from 0.57 to 0.67), but never large effects. Such differences could

2   arise due to different levels of individual variability or effects of scanning equipment

3   per se (or both). Regardless of the sources, it is important to note that when

4   datasets are closely matched, as they are in this primary acquisition, most cortical

5   regions did not show site differences even at this sensitive threshold, the effect sizes

6   of site differences were predominantly small, and, as noted earlier, patterns of

7   within-site ISC for each site were highly correlated with one another and with the

8   pattern of across-site ISC.

9

10   **2.2.   When acquisitions are not matched, differences become more**

11   **apparent, despite still-widespread consistency.**

12   In an exploratory comparison we also examined cross-site differences

13   between the primary Caltech dataset and a pilot dataset ("Pilot") also collected at

14   Indiana University prior to a scanner upgrade. These unmatched datasets were

15   collected using different scanner models, protocols with numerous differences, and

16   different preprocessing approaches (Table 1), but using the same scanner

17   manufacturer (Siemens) and field strength (3T). As these acquisitions were not

18   designed to systematically test the effects of varying all these parameters, it will not

19   be possible to disentangle the specific sources of any differences identified.

20   Nonetheless, we include this comparison as somewhat more representative of real-

21   world differences between pre-existing datasets collected as participants watch the

22   same video stimulus.

1   Figure 5 (center) depicts consistency between these two unmatched datasets

2   at the level of median time series at each site, along with consistency between the

3   two primary matched datasets (left), and the difference of these quantities (right),

4   for comparison. Median time series from all three sites are presented as well. The

5   same general pattern of results from Figure 2 is evident even though the Caltech and

6   Pilot acquisitions are unmatched: high similarity across datasets at the group

7   average level while participants watch the same video stimulus. Despite this high

8   similarity, though, it is also visually apparent that similarity between the unmatched

9   acquisitions is reduced, relative to the matched acquisitions.

10

11   [INSERT FIGURE 5 AROUND HERE]

12

13   Because potential differences between datasets are expected to manifest

14   most strongly within individual subject data, we tested for differences at the level of

15   pairwise ISCs in the same way as described previously, by testing whether within-

16   site similarity differed from across-site similarity for either site. Pairwise ISCs

17   within and across each of these unmatched datasets are mapped in Figure 6 (top,

18   and Supplemental Figure 3), and also presented as line plots (bottom) to facilitate

19   comparison. Differences in ISC levels are visible, with within-site similarity for the

20   Pilot dataset appearing elevated. To test for differences, as before, to be

21   conservative, we did not correct for multiple corrections, and considered a brain

22   region as having a site difference if differences were observed for either site (and

23   not necessarily both). In contrast to the previous results for the matched datasets

1   (IU vs. Caltech, see 2.1), here, when datasets were unmatched in numerous ways, we

2   observed site differences in most brain regions: 89.6% of regions for Video1, and

3   97.9% for Video 2.

4

5   [INSERT FIGURE 6 AROUND HERE]

6

7   As can be seen in Figure 6, differences between median within-site and

8   median across-site ISC varied across ROIs and varied by site. They appear relatively

9   minimal for the Caltech dataset but more noticeable for the Pilot dataset, and they

10   are not homogenous across the brain. For instance, elevated within-site similarity in

11   the Pilot dataset was found throughout the superior temporal lobe extending into

12   the temporoparietal junction (c.f. Figure 6, x = 9 and 57, left and right posterior

13   superior temporal gyrus), but much less so for many visual areas (c.f. x = 39 and 87,

14   left and right occipital fusiform). For the within-Pilot vs. across-site ISC comparison

15   (pink vs. black lines), these differences ranged across ROIs from 0.008-0.2 (median

16   0.06; IQR 0.06) for Video 1 and from 0.006-0.24 (median 0.08; IQR 0.5) for Video2.

17   CLES for these differences ranged from 0.52-0.92 across ROIS (median 0.66; IQR

18   0.11) for Video1 and from 0.52-0.95 (median 0.71; IQR 0.11) for Video2. For the

19   within-Caltech vs. across-site ISC comparison (orange vs. black lines), the median

20   difference across ROIS ranged from -0.05-0.12 (median 0.02; IQR 0.035) for Video1

21   and -0.06-0.1 (median -0.006; IQR 0.03) for Video2. CLES ranged from 0.32-0.76

22   (median 0.49; IQR 0.1) for Video1 and from 0.32-0.81 (median 0.47; IQR 0.1) for

23   Video2. In contrast to the matched datasets, then, quantitative comparisons of

1   pairwise ISC levels within- and across- sites can reveal differences with medium-to-

2   large effect sizes spanning ROIs.

3   Due to the numerous factors that vary between these unmatched datasets

4   (Table 1), it is not possible to pinpoint the exact cause(s) of the elevated within-site

5   similarity in the Pilot dataset; disentangling these factors is beyond the scope of the

6   current project and a question for future targeted new acquisitions. Nonetheless we

7   present these comparisons as a case study showing how similarity across and

8   within sites can vary when datasets using the same video stimuli are unmatched.

9   And while quantitative differences in ISC levels were prevalent in comparing these

10   unmatched datasets, it is important to observe that qualitatively, the pattern of ISC

11   remained similar across sites and within each site. Figure 6 shows that all three

12   lines increase and decrease in tandem, and indeed they are all highly correlated for

13   both video scans (all $r > 0.92$, $p < 0.0001$). So, while levels of ISC can differ

14   considerably when datasets are unmatched, ROIs with higher ISC at one site also

15   have higher ISC at the other site and across-sites, and vice versa.

16   Altogether, these results indicate that differences in brain responses across

17   sites are more readily apparent when datasets are unmatched, and can be

18   considerable and non-homogeneous across the cortex – but, despite these

19   quantitative differences, video stimuli drive qualitatively consistent patterns of

20   brain responding across sites even when numerous acquisition, processing,

21   hardware, and participant details vary freely.

22

23   **Discussion**

1

2   We find that video fMRI paradigms evoke robustly similar brain responses

3 across different sites and samples of subjects, with consistent brain responses found

4 through most of the cortex. When datasets are matched closely, such that scanner

5 manufacturer, model, imaging protocols, and preprocessing details are the same at

6 each site, differences in brain responses between datasets are minimal. When

7 datasets are unmatched, such that scanner model and acquisition and processing

8 details vary more freely, differences are more prevalent, especially in pairwise

9 comparisons of individual data. Nonetheless, consistency of brain responses across

10 unmatched datasets remains high, although attenuated relative to matched datasets.

11   In the matched datasets, at the level of group-average time series, we find

12 that most regions of the brain (>99%) respond similarly across sites, and this nearly

13 cortex-wide similarity is observed across parcellation granularities (from 100 to

14 1000 ROIs) – it is not an artifact of using a coarse parcellation and therefore

15 spatially smoothing across large swaths of cortex. We find comparable results at the

16 level of individual time series similarity, albeit with the reduced correlation

17 magnitudes expected from pairwise correlations. Procedures adjusting for

18 individual differences in functional specialization and hemodynamic responses

19 (Haxby et al., 2020, Dubois & Adolphs, 2016) could be employed in the future to

20 potentially reveal even higher similarity across sites.

21   Across parcellations, regions with consistent group-level brain responses

22 include some frontal and ventral regions that are not typically observed on

23 individual-level ISC maps. On one hand, this is reminiscent of findings in task-based

1   fMRI that averaging across larger numbers of timeseries "unmasks" the involvement

2   of common task-locked signal in previously unappreciated regions (Gonzalez-

3   Castillo et al., 2012). On the other hand, cross-site similarity between group-level

4   timeseries in some of these regions is, while statistically significant, quite weak, and

5   similar correlations have been interpreted by other groups as showing little

6   evidence of synchronized brain responses (Chang et al., 2021). We see this as a

7   scenario akin to asking "is the glass half empty or half full?". Weak correlations

8   between time series undoubtedly indicate that the signal is predominantly

9   explained by other sources including endogenous processing, intrinsic brain

10   dynamics, and various sources of scanner and physiological noise. Alternative

11   methods for correcting for multiple comparisons that capture underlying data

12   dimensionality and potential dependencies between timeseries could also shift the

13   statistical threshold delineating which ROIs can be considered weakly correlated

14   above chance. Nonetheless, identifying shared signal – albeit weakly shared –

15   between two datasets is stronger evidence than can be provided by one dataset

16   alone that there is something about these video stimuli that can evoke common

17   brain function in such areas, potentially indirectly and potentially only momentarily.

18   Better understanding the aspects of the video stimulus that drive such weakly

19   evoked responses in brain areas more commonly associated with endogenous brain

20   function is an important topic of future study (see also Chang et al., 2021; Yeshurun

21   et al., 2021).

22   The specific moments of the video and specific features of the stimulus that

23   drive the most and least consistent brain responses across sites is also a question

1   for further study. A visual comparison across the different video scans presented in

2   Figure 4 shows clear similarities in the patterns of group-level consistency (top) and

3   pairwise across-site ISC (bottom) evoked by all the different video stimuli

4   employed. In other words, brain regions that respond very consistently across sites

5   during one video tend to also respond very consistently in a different video, and vice

6   versa. This surely reflects fundamental aspects of neural architecture for dynamic

7   audiovisual stimulation, as the most consistent brain regions were the primary

8   sensory areas expected to be most directly driven by the stimulus. Some differences

9   in ISC levels across each full-length scan could arise due to differences in video

10   lengths, which varied considerably. But even after equating for video lengths,

11   differences in group-level brain response consistency across different videos could

12   be observed (Supplemental Figure 2). Presumably, these differences are elicited by

13   specific video stimulus features and idiosyncratic responses to those features, as

14   well as the processing demands they impose on the brain (see also Hasson et al.,

15   2010, for discussion of stimulus-specificity of within-site reliability). For instance,

16   cross-site consistency in medial prefrontal cortex (mPFC) for both episodes of *The*

17   *Office* (Video3 and Video4) appears elevated relative to the other videos. This is

18   noteworthy because *The Office* is a TV show that is characterized by many socially

19   awkward moments and was specifically selected for its increased demands on the

20   social brain (including mPFC; Kennedy & Adolphs, 2012). Further work

21   comprehensively decomposing these videos from low-level stimulus features to

22   high-level semantic properties will be needed to verify this observation and more

1   generally understand how different video stimulus properties influence patterns of

2   consistency across sites.

3   As noted, the comparison between the unmatched datasets was presented as

4   a case study and as an example with which to contrast the high levels of cross-site

5   similarity in the matched datasets. Particularly with increasing data sharing efforts

6   in recent years, this comparison has more real-world relevance for the pooling of

7   some pre-existing vfMRI datasets, which are unlikely to have been as carefully

8   matched as the primary samples in this study. For the unmatched datasets in the

9   current study, we observed quantitative differences in group-level consistency and

10 pairwise ISC, but qualitatively, the patterns of pairwise ISC remained highly similar

11 across and within each site. For these unmatched datasets, differences in the

12 acquisition and processing varied considerably (Table 1), including participants,

13 scanner model, acquisition parameters including voxel size, sampling rate,

14 multiband parameters, and sequences used for anatomical scans and fieldmaps, and

15 preprocessing choices including denoising methodology, filtering, and smoothing.

16 Many if not all of these factors could influence cross-site consistency of brain

17 responses (e.g., He et al., 2020; Friedman et al, 2006; Yu et al., 2018). It is also

18 important to note that the levels of consistency observed in the unmatched datasets

19 are not intended to suggest a lower bound. All datasets in this study used the same

20 scanner manufacturer (Siemens) and field strength (3T), and it is reasonable to

21 expect that cross-manufacturer or cross-magnet comparisons could potentially

22 further affect consistency. A full disentangling of the specific combinations of factors

23 that gave rise to the more prevalent differences observed in the unmatched datasets

1 is beyond the scope of the current project, which was not designed to test these

2 factors systematically. An important question for future study would be to unpack

3 these factors by parametrically varying the differences between these datasets, and

4 to include comparisons across different scanner manufacturers and different field

5 strengths. This would also guide the development of statistical harmonization

6 methods for pooling existing video fMRI data (as in Yu et al., 2018; Yamashita et al.,

7 2019, for resting state data), which could span a variety of manufacturers and even

8 field strengths.

9   Even for the matched datasets, our existing data does not allow us to

10 conclusively separate effects caused by different scanners from other factors that

11 covaried between the matched datasets. Those factors were intentionally

12 minimized, but do include both different physical scanners and different individual

13 subjects. Some aspects of the differences that were observed between these

14 matched datasets could thus have been driven by participant variability rather than

15 scanner differences. To fully decouple individual variability from scanner variability,

16 a new data acquisition with traveling subjects that are repeatedly scanned at

17 different locations (as has been done for resting state designs; Noble et al., 2017)

18 would be required. This would be an important direction for future work.

19

20   **Conclusion**

21   In sum, we find similar group-level brain responses spanning the cortex

22 when participants at different sites watch the same video stimulus, and these highly

23 similar average time series occur with both matched and unmatched datasets. When

1   datasets are carefully matched such that the acquisition and processing is effectively

2   identical, differences between datasets at the level of pairwise similarity of

3   individual brain responses are minimal, and some such differences could reflect

4   individual variability rather than scanner-specific effects. When dataset parameters

5   vary more freely, differences between sites are more prevalent, which points to the

6   importance of both careful control for such differences in analyses and of the

7   development of harmonization protocols specific to ISC analyses of video fMRI data

8   for at least some purposes. Nonetheless, the overarching conclusion indicates high

9   levels of consistency in video-evoked fMRI data across these different sites, across

10   matched and unmatched datasets alike. The ability to quantify this consistency

11   highlights one of the unique features of video fMRI and holds promise for further

12   development of this approach to studies of individual differences in healthy and

13   clinical populations alike.

14

15

16

17   **Author contributions**

18

19   LB, RA, and DPK conceptualized the project. HC & MT developed MRI protocols,

20   coordinated them across sites, and continuously conducted scanner quality

21   assurance. LB, DK, and Indiana University and Caltech personnel collected data. LB,

22   DK, and YH preprocessed data and ensured data quality. LB & DPK developed the

1 analysis approach and LB analyzed the data. LB drafted the manuscript with input

2 from DPK and all co-authors provided feedback and approved the final version.

3

4

5

6 **References**

7

8 Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee. 2008. "Symmetric

9 Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated

10 Labeling of Elderly and Neurodegenerative Brain." Medical Image Analysis 12 (1):

11 26–41. https://doi.org/10.1016/j.media.2007.06.004.

12

13 Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson,

14 M., … & Rieck, J. R. (2020). Variability in the analysis of a single neuroimaging

15 dataset by many teams. *Nature*, *582*(7810), 84-88.

16

17 Burgess, G. C., Kandala, S., Nolan, D., Laumann, T. O., Power, J. D., Adeyemo, B., … &

18 Barch, D. M. (2016). Evaluation of denoising strategies to address motion-correlated

19 artifacts in resting-state functional magnetic resonance imaging data from the

20 Human Connectome Project. Brain Connectivity, 6(9), 669-680.

21

1   Burunat, I., Toiviainen, P., Alluri, V., Bogert, B., Ristaniemi, T., Sams, M., & Brattico, E.

2   (2016). The reliability of continuous brain responses during naturalistic listening to

3   music. *NeuroImage*, *124*, 224-231.

4

5   Byrge, L.*, Dubois, J.*, Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015).

6   Idiosyncratic brain activation patterns are associated with poor social

7   comprehension in autism. Journal of Neuroscience, 35(14), 5837-5850. (* = equal

8   contribution).

9

10   Byrge, L., & Kennedy, D. P. (2020). Accurate prediction of individual subject identity

11   and task, but not autism diagnosis, from functional connectomes. *Human brain*

12   *mapping*, *41*(9), 2249-2262.

13

14   Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K. M., Greenstein, N., Chen, P. H. A., &

15   Manning, J. R. (2021). Endogenous variation in ventromedial prefrontal cortex state

16   dynamics during naturalistic viewing reflects affective experience. *Science Advances*,

17   *7*(17), eabf7129.

18

19   Chen, G., Shin, Y. W., Taylor, P. A., Glen, D. R., Reynolds, R. C., Israel, R. B., & Cox, R. W.

20   (2016). Untangling the relatedness among correlations, part I: nonparametric

21   approaches to inter-subject correlation analysis at the group level. *NeuroImage*, *142*,

22   248-259.

23

1 Ciric, R., Rosen, A. F., Erus, G., Cieslak, M., Adebimpe, A., Cook, P. A., ... &

2 Satterthwaite, T. D. (2018). Mitigating head motion artifact in functional

3 connectivity MRI. *Nature Protocols*, *13*(12), 2801-2826.

4

5 Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (No.

6 1). Cambridge university press.

7

8 Di Martino, A., O'connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., ... &

9 Milham, M. P. (2017). Enhancing studies of the connectome in autism using the

10 autism brain imaging data exchange II. *Scientific data*, *4*(1), 1-15.

11

12 Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from

13 fMRI. *Trends in cognitive sciences*, *20*(6), 425-443.

14

15 Eickhoff, S. B., Milham, M., & Vanderwal, T. (2020). Towards clinical applications of

16 movie fMRI. *NeuroImage*, *217*, 116860.

17

18 Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences

19 for spatial extent have inflated false-positive rates. *Proceedings of the national*

20 *academy of sciences*, *113*(28), 7900-7905.

21

22 Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., ... &

23 Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI

1   measures? New empirical evidence and a meta-analysis. *Psychological Science*,

2   *31*(7), 792-806.

3

4   Elliott, M. L., Knodt, A. R., & Hariri, A. R. (2021). Striving toward translation:

5   strategies for reliable fMRI measurement. *Trends in cognitive sciences*.

6

7   Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J.

8   (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from

9   unseen sites. *PloS one*, *12*(9), e0184661.

10

11   Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... &

12   Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional

13   MRI. Nature methods, 16(1), 111-116.

14

15   Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., & Constable, R. T. (2018). Trait

16   paranoia shapes inter-subject synchrony in brain activity during an ambiguous

17   social narrative. *Nature Communications*, *9*(1), 1-13.

18

19   Friedman, L., Glover, G. H., & FBIRN Consortium. (2006). Reducing interscanner

20   variability of activation in a multicenter fMRI study: controlling for signal-to-

21   fluctuation-noise-ratio (SFNR) differences. *Neuroimage*, *33*(2), 471-481.

22

1   Gonzalez-Castillo, J., Saad, Z. S., Handwerker, D. A., Inati, S. J., Brenowitz, N., &

2   Bandettini, P. A. (2012). Whole-brain, time-locked activation with simple tasks

3   revealed using massive averaging and model-free analysis. *Proceedings of the*

4   *National Academy of Sciences*, *109*(14), 5487-5492.

5

6   Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... &

7   Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and

8   describing outputs of neuroimaging experiments. *Scientific data*, *3*(1), 1-9.

9

10 Gruskin, D. C., Rosenberg, M. D., & Holmes, A. J. (2020). Relationships between

11 depressive symptoms and brain responses during emotional movie viewing emerge

12 in adolescence. *NeuroImage*, *216*, 116217.

13

14 Guo, C. C., Nguyen, V. T., Hyett, M. P., Parker, G. B., & Breakspear, M. J. (2015). Out-of-

15 sync: disrupted neural activity in emotional circuitry during film viewing in

16 melancholic depression. *Scientific reports*, *5*(1), 1-12.

17

18 Hall, D., Huerta, M. F., McAuliffe, M. J., & Farber, G. K. (2012). Sharing heterogeneous

19 data: the national database for autism research. *Neuroinformatics*, *10*(4), 331-339.

20

21 Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject

22 synchronization of cortical activity during natural vision. *science*, *303*(5664), 1634-

23 1640.

1

2   Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., & Behrmann,

3   M. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed

4   under continuous real-life viewing conditions. Autism Research, 2(4), 220-231.

5

6   Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during

7   natural stimulation. *Trends in cognitive sciences*, *14*(1), 40-48.

8

9   Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment:

10   Modeling shared information encoded in idiosyncratic cortical topographies. *Elife*, *9*,

11   e56601.

12

13   He, Y., Byrge, L., & Kennedy, D. P. (2020). Nonreplication of functional connectivity

14   differences in autism spectrum disorder across multiple sites and denoising

15   strategies. *Human brain mapping*, *41*(5), 1334-1350.

16

17   Holmes, A. J., & Patrick, L. M. (2018). The myth of optimality in clinical neuroscience.

18   *Trends in cognitive sciences*, *22*(3), 241-257.

19

20   Ioannidis, J. P. (2005). Why most published research findings are false. *PloS*

21   *medicine*, *2*(8), e124.

22

1 King, J. B., Prigge, M. B., King, C. K., Morgan, J., Weathersby, F., Fox, J. C., ... &

2 Anderson, J. S. (2019). Generalizability and reproducibility of functional connectivity

3 in autism. *Molecular Autism*, *10*(1), 1-23.

4

5 Loth, E., Charman, T., Mason, L., Tillmann, J., Jones, E. J., Wooldridge, C., ... & Buitelaar,

6 J. K. (2017). The EU-AIMS Longitudinal European Autism Project (LEAP): design and

7 methodologies to identify and validate stratification biomarkers for autism

8 spectrum disorders. *Molecular autism*, *8*(1), 1-19.

9

10 Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared

11 responses across subjects using intersubject correlation. *Social Cognitive and*

12 *Affective Neuroscience*, 14 (6), 667–685.

13

14 Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., ... & Yeo, B. T.

15 (2017). Best practices in data analysis and sharing in neuroimaging using MRI.

16 *Nature neuroscience*, *20*(3), 299-303.

17

18 Nickerson, L. D. (2018). Replication of resting state-task network correspondence

19 and novel findings on brain network activation during task fmri in the human

20 connectome project study. *Scientific reports*, *8*(1), 1-12.

21

1   Noble, S., Scheinost, D., Finn, E. S., Shen, X., Papademetris, X., McEwen, S. C., ... &

2   Constable, R. T. (2017). Multisite reliability of MR-based functional connectivity.

3   Neuroimage, 146, 959-970.

4

5   Pantelis, P. C., Byrge, L., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). A specific

6   hypoactivation of right temporo-parietal junction/posterior superior temporal

7   sulcus in response to socially awkward situations in autism. *Social cognitive and*

8   *affective neuroscience*, *10*(10), 1348-1356.

9

10 Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M.

11 R., ... & Yarkoni, T. (2017). Scanning the horizon: towards transparent and

12 reproducible neuroimaging research. *Nature reviews neuroscience*, *18*(2), 115.

13

14 Power, J. D., Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., & Jones, R. M. (2019).

15 Distinctions among real and apparent respiratory motions in human fMRI data.

16 *NeuroImage*, *201*, 116041.

17

18 Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development

19 of the social brain from age three to twelve years. *Nature communications*, *9*(1), 1-

20 12.

21

1   Richardson, H. (2019). Development of brain networks for social functions:

2   Confirmatory analyses in a large open source dataset. *Developmental cognitive*

3   *neuroscience*, *37*, 100598.

4

5   Ruscio, J. (2008). A probability-based measure of effect size: robustness to base

6   rates and other factors. *Psychological methods*, *13*(1), 19.

7

8   Saarimäki, H. (2021). Naturalistic stimuli in affective neuroimaging: a review.

9   *Frontiers in Human Neurosciences.*

10

11   Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith,

12   S. M. (2014). Automatic denoising of functional MRI data: combining independent

13   component analysis and hierarchical fusion of classifiers. Neuroimage, 90, 449-468.

14

15   Salmi, J., Roine, U., Glerean, E., Lahnakoski, J., Nieminen-von Wendt, T., Tani, P., ... &

16   Sams, M. (2013). The brains of high functioning autistic individuals do not

17   synchronize with those of others. *NeuroImage: Clinical*, *3*, 489-497.

18

19   Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo

20   BTT. Local-Global parcellation of the human cerebral cortex from intrinsic

21   functional connectivity MRI. Cerebral Cortex, 29:3095-3114, 2018

22

1   Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience:

2   critically acclaimed. *Trends in cognitive sciences*, *23*(8), 699-714.

3

4   Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet:

5   Naturalistic paradigms in developmental functional neuroimaging. *Developmental*

6   *cognitive neuroscience*, *36*, 100600.

7

8   Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common

9   language effect size statistics of McGraw and Wong. *Journal of Educational and*

10   *Behavioral Statistics*, *25*(2), 101-132.

11

12   Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., ... & Imamizu,

13   H. (2019). Harmonization of resting-state functional MRI data across multiple

14   imaging sites via the separation of site differences into sampling bias and

15   measurement bias. *PloS biology*, *17*(4), e3000042.

16

17   Yang, Z., Wu, J., Xu, L., Deng, Z., Tang, Y., Gao, J., ... & Wang, J. (2020). Individualized

18   psychiatric imaging based on inter-subject neural synchronization in movie

19   watching. *NeuroImage*, *216*, 116227.

20

21   Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: where the

22   idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3),

23   181-192.

1

2 Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., ... & Sheline, Y. I.

3 (2018). Statistical harmonization corrects site effects in functional connectivity

4 measurements from multi-site fMRI data. *Human brain mapping*, *39*(11), 4213-

5 4227.

6

7 Zilles, K., & Amunts, K. (2013). Individual variability is not noise. *Trends in cognitive*

8 *sciences*, *17*(4), 153-155.

9

10 Zuo, X. N., Biswal, B. B., & Poldrack, R. A. (2019). Reliability and reproducibility in

11 functional connectomics. *Frontiers in neuroscience*, *13*, 117.

12

13

14

15 **Tables**

16 Table 1.

17 Similarities and differences between the matched and unmatched datasets.

	Matched datasets		
	IU	**Caltech**	**Pilot**
		Unmatched datasets	
Participants			
population	HC	HC	HC
sample	different	different	different
MRI acquisition			
scanner manufacturer	Siemens	Siemens	Siemens
field strength	3T	3T	3T

scanner model	Prisma.Fit	Prisma.Fit	TIM Trio
scanner location	Bloomington, IN	Pasadena, CA	Bloomington, IN
MRI protocols	matched	matched	unmatched
EPI resolution (spatial)	2.5mm iso	2.5mm iso	3.4mm iso
EPI resolution (temporal)	0.72s TR	0.72s TR	0.813s TR
multiband acceleration factor	6	6	3
Experiment			
video stimulus	same (V1-6)	same (V1-6)	same (V1-2)
stimulus presentation code	same	same	same
Data preprocessing & analysis			
preprocessing pipeline	same	same	different
denoising approach	GLM with GSR	GLM with GSR	GLM+ICA-FIX, then GSR
temporal filtering	bandpass	bandpass	detrending
spatial smoothing	2.54mm	2.54mm	none
analysis code	same	same	same
Personnel			
experimenter	different	different	different
data analyst	same	same	same

1

2   *Table presents the main similarities and differences between the matched (IU &*

3   *Caltech; dark gray) and unmatched (Caltech & Pilot; light gray) datasets. Table*

4   *organization corresponds roughly to the taxonomy of reproducibility in neuroimaging*

5   *from Nichols and colleagues (2017). The primary comparison between matched*

6   *datasets is situated between "Near replicability" and "Intermediate replicability" of*

7   *generalization over materials and methods in that taxonomy. The exploratory*

8   *comparison between unmatched datasets is situated between "Intermediate*

9   *replicability" and "Far replicability"; for that comparison, the Pilot acquisition was*

10 *resampled temporally to match the sampling rate of the primary matched datasets.*

1   *HC, healthy control adults. No participants overlapped between datasets. Entries listed*

2   *as "same" and "different" for brevity are further detailed in Methods.*

3

4   Table 2.

5   Video scans and sample sizes for matched and unmatched datasets.

6

	Video1 movie trailers ~13.5 min.			Video2 movie trailers ~13 min.			Video3 The Office ~22 min.		Video4 The Office ~22 min.		Video5 Partly cloudy ~5.6 min		Video6 Bang ~8 min.	
	IU	Cal	Pilot	IU	Cal	Pilot	IU	Cal	IU	Cal	IU	Cal	IU	Cal
Initial sample	61	29	29	56	28	29	61	28	56	28	56	28	54	28
Scan issues	1	1	0	2	2	0	1	2	1	4	0	1	0	2
MRIQC outlier	5	1	n/a	1	1	n/a	4	1	1	0	4	0	1	1
Motion	3	1	4	3	0	4	2	2	2	1	2	1	1	3
Registration	1	0	0	4	1	0	2	0	3	0	2	1	3	0
ISC outlier	3	1	n/a	1	1	n/a	3	1	0	0	2	0	2	0
Final sample	48	25	25	45	23	25	49	22	49	23	46	25	47	22

7

8   *Table presents video scans, initial sample sizes, and exclusions for matched datasets*

9   *(IU, Cal) and unmatched datasets (Cal, Pilot). IU, Indiana. Cal, Caltech. Video1 and*

10   *Video2 are the primary scans analyzed here because those video stimuli were used in*

11   *all three datasets. Scan issues include technical problems (muffled sound, projector*

12   *issues, missing image data) and participant sleep. Quality assurance workflow differed*

13   *for the matched and unmatched datasets and MRIQC and ISC outlier exclusions were*

14   *not applicable to the pilot dataset. Columns with a white background denote scans*

15   *collected during the first session; columns with a gray background denote scans*

1   *collected during a second session approximately one week after the first. Video scans*

2   *1-4 were all preceded by rest scans.*

3

4

1 **Figure Legends**

2
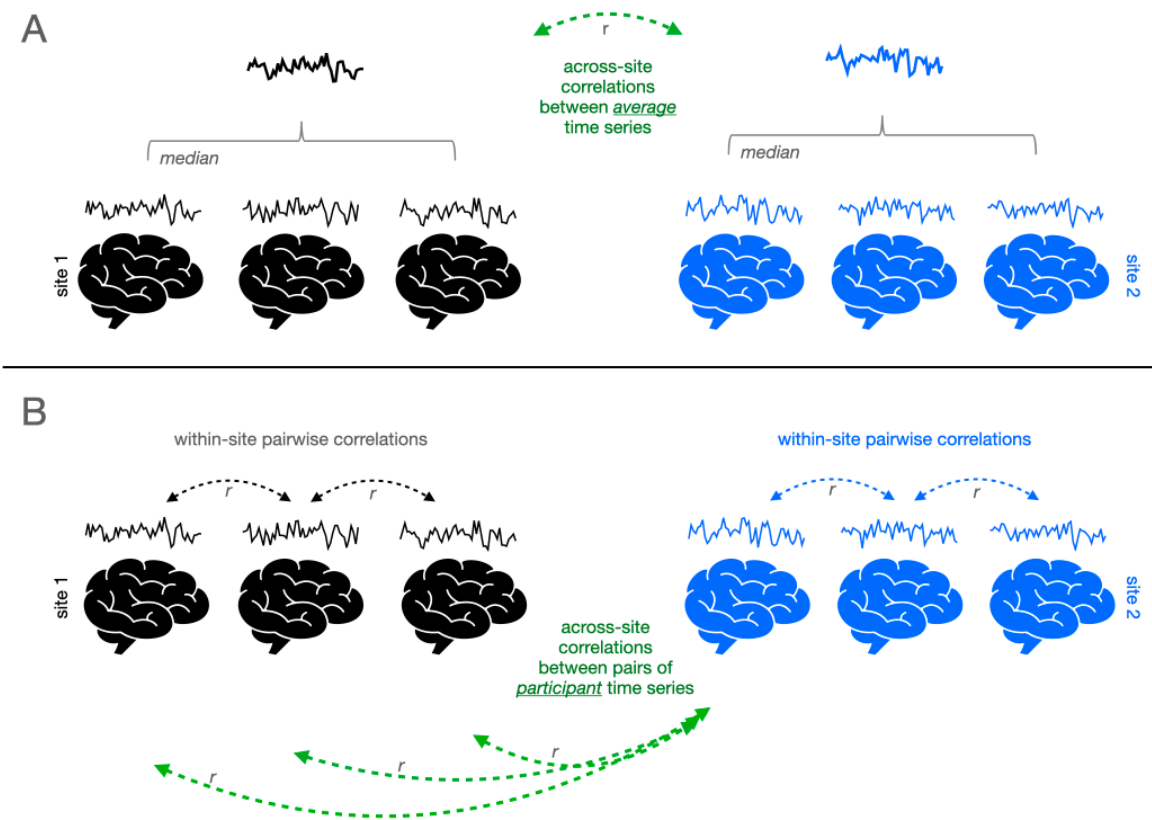
3

4 *Figure 1. Schematic of approach for examining consistency of video-evoked brain responses across sites*

5 *(black and blue) at the level of the group (A) and individuals (B). Example individual time series depict*

6 *the fMRI BOLD signal averaged across a given region of interest, across the duration of the video. To*

7 *examine consistency across sites at the group level (A), the average of all these individual time series is*

8 *computed for each site (bolded timeseries), and then the correlation between those average site-level*

9 *time series is computed (green arrow). To examine consistency across sites at the individual level (B),*

10 *correlations between pairs of time series from individual participants at different sites are computed*

11 *(green arrows), and for some analyses compared to correlations between pairs of participant time*

12 *series from the same site (black arrows, or blue arrows).*
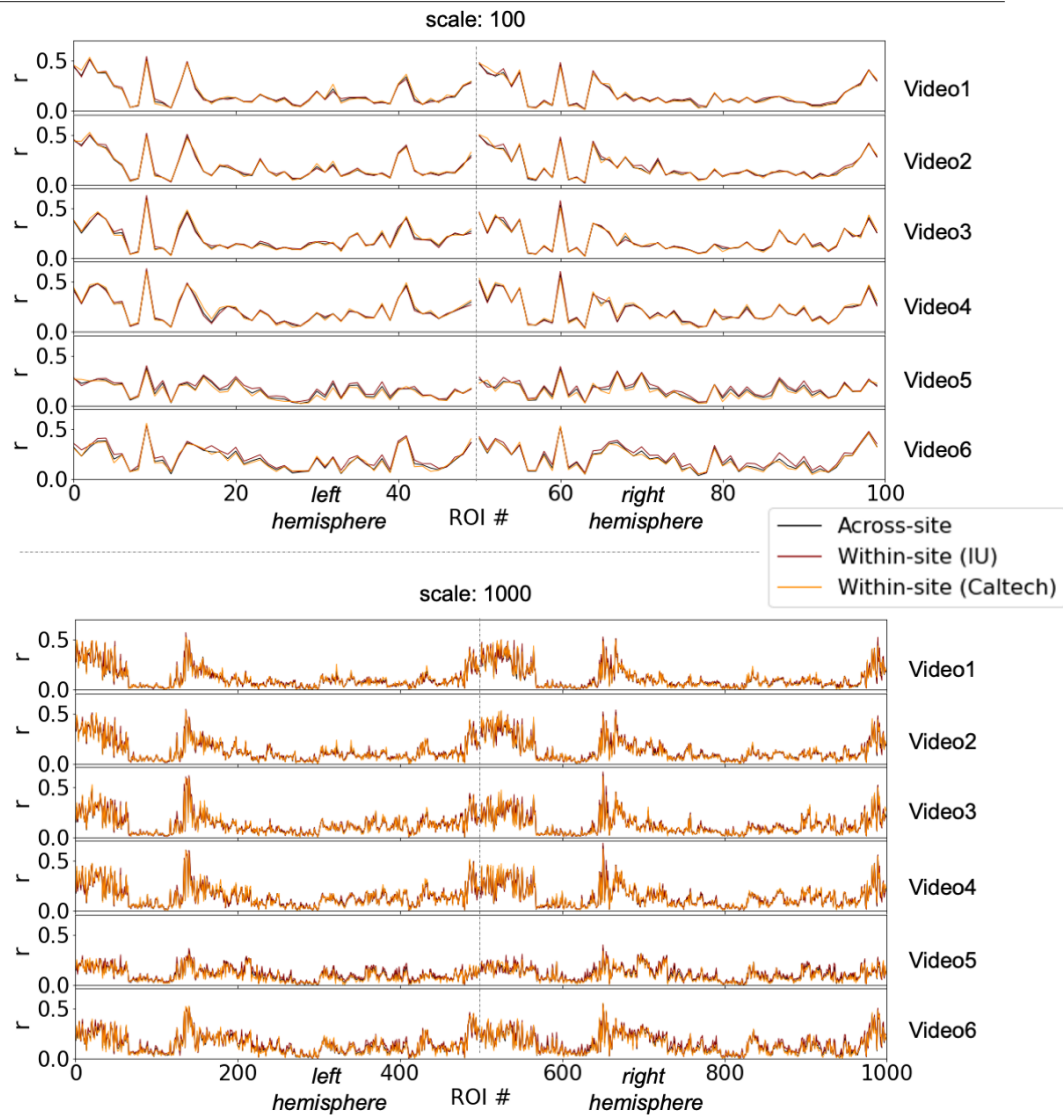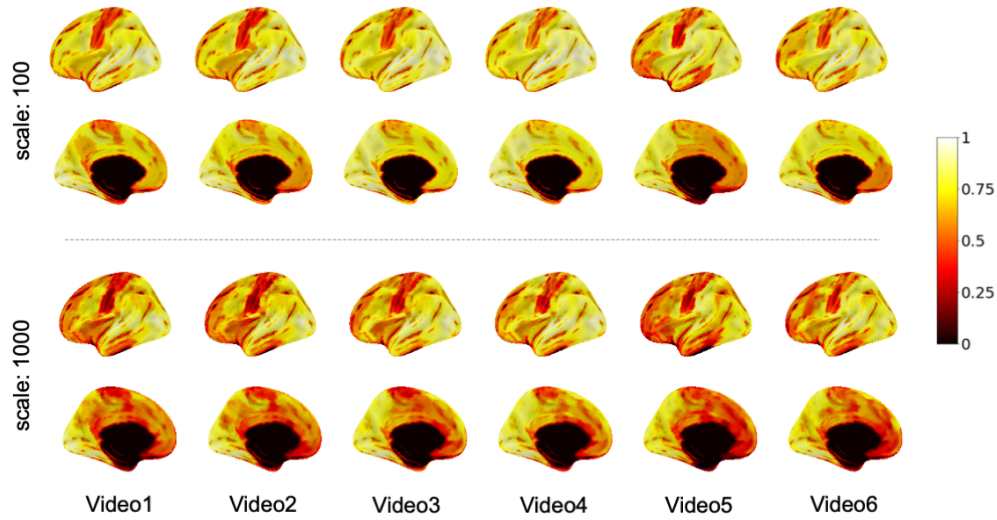
Video 1
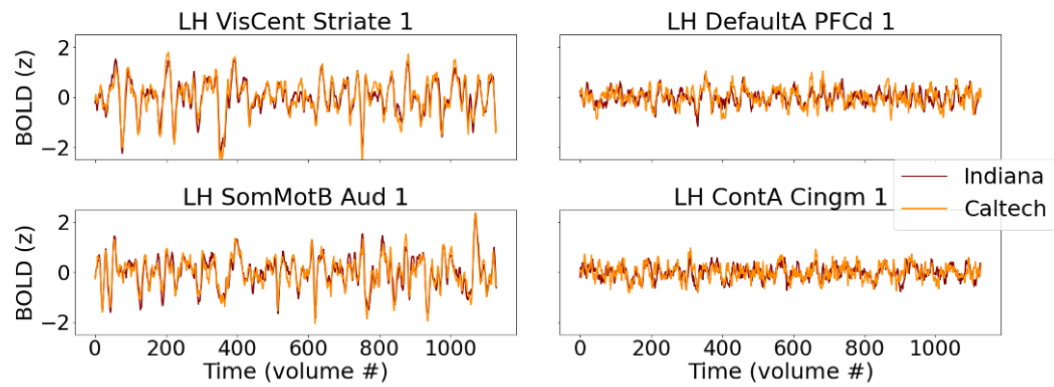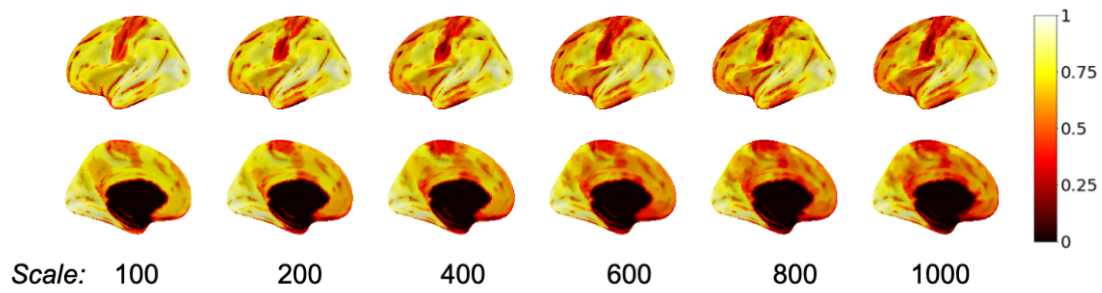
Video 2
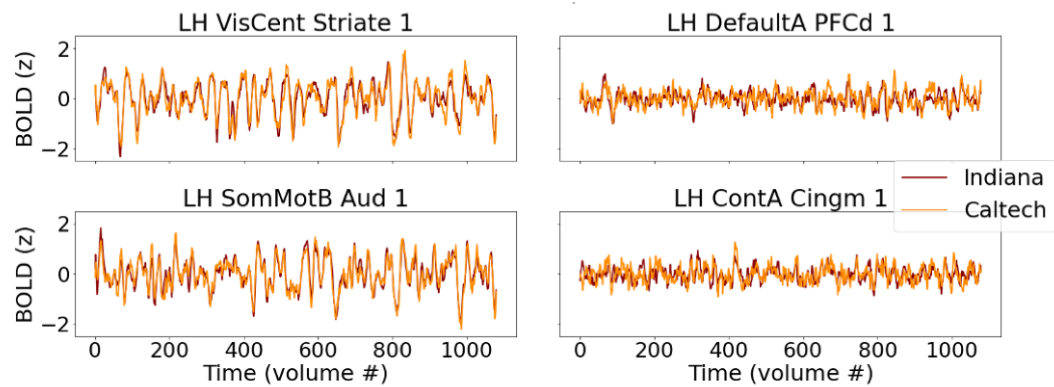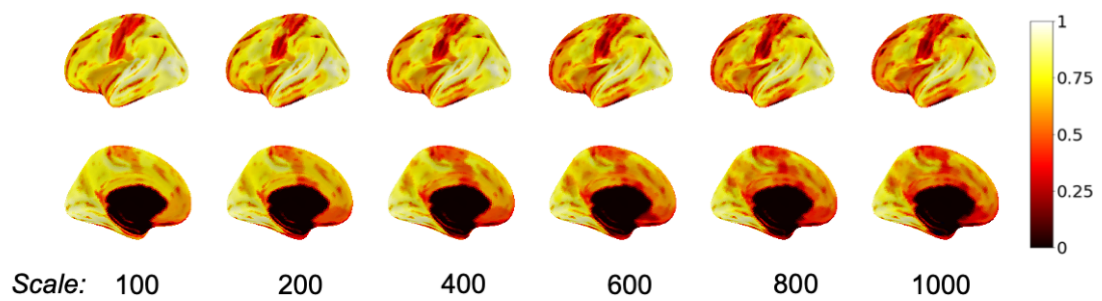
1   *Figure 2. Consistency of average group-level brain responses across sites while participants watched the*

2   *same videos (see Fig. 1A), for matched datasets. Brain visualizations depict correlations between*

3   *median time series across all participants at each site, in each brain region, under different scales of the*

4   *Schaefer parcellation (100-1000 ROIs). The Schaefer parcellation is a cortical parcellation; black along*

5   *the midline in medial views here and elsewhere indicate missing data, not low correlations. Line plots*

6   *depict median timeseries across participants at each site in primary sensory areas (left) and association*

7   *areas (right) using the 400-region Schaefer parcellation during Video 1 (top) and Video 2 (bottom). All*

8   *figures depict the left hemisphere; the pattern of results for the right hemisphere is effectively the same.*
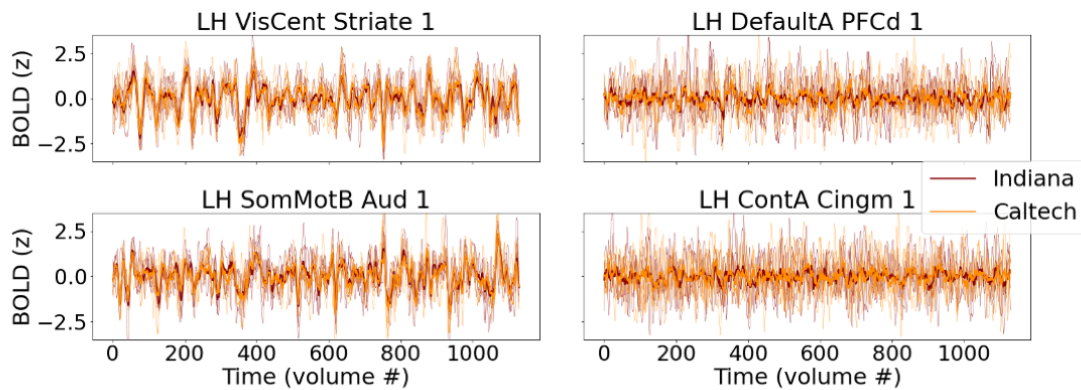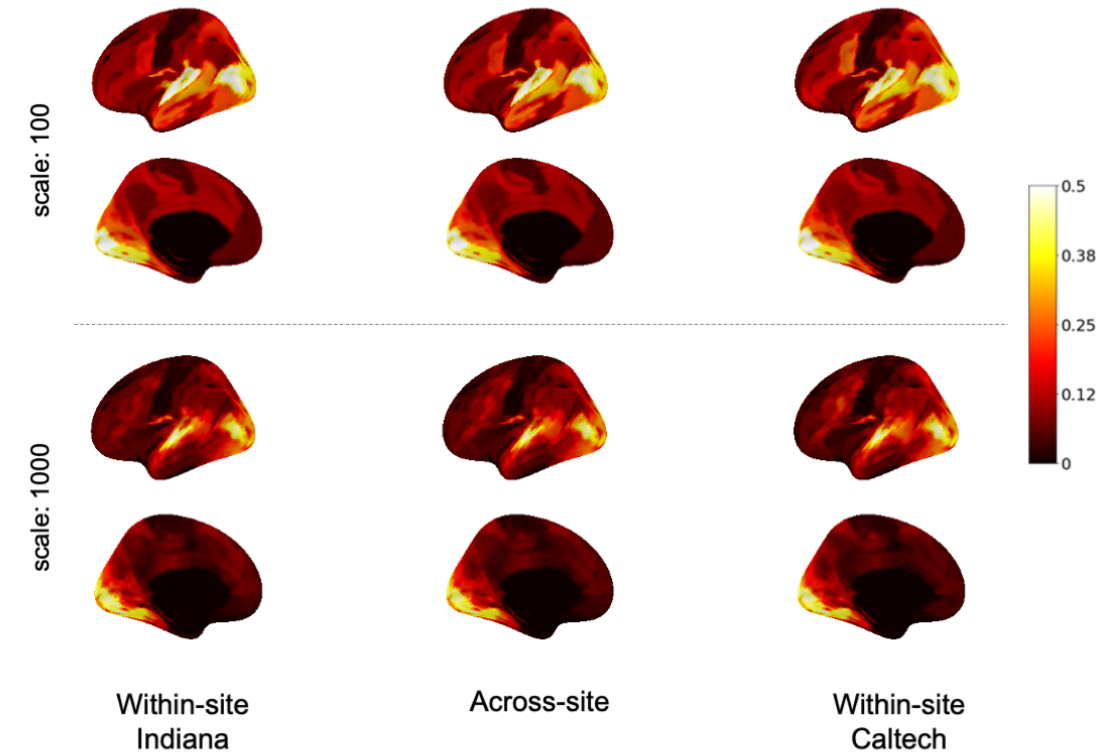
Figure 3. Similarity of individual participant brain responses within and across sites during Video1 (see Fig. 1B) for matched datasets. Top: brain maps depict magnitudes of medians of pairwise correlations between participant brain response time series in each region, for the most coarse (top) and the most fine (bottom) scales of the Schaefer cortical parcellation. The left and right columns show correlations among pairs of participants at the same site (left: Indiana; right: Caltech). The center column shows correlations among pairs of participants spanning different sites. While absolute values are depicted

1 *here for readability, nearly all median correlations were positive, except one temporal pole ROI with a*

2 *near- zero median correlation of -0.0036. As in Figure 2, black along the midline in medial views*

3 *indicates missing data, not low correlations. See also Figure 4 (bottom panel, top plot) for a line version*

4 *of this same data, and Supplemental Table 2 for characterization of differences and effect sizes. Bottom:*

5 *line plots depict timeseries from 5 randomly-selected participants at each site in primary sensory areas*

6 *(left) and association areas (right), along with the site-level median time series shown in Figure 2. These*

7 *line plots use the same mid-scale parcellation as Figure 2 (Schaefer 400x17). See also Supplemental*

8 *Figure 1 for the equivalent figure for Video2, which is similar but supplemental for space purposes.*

1

1   *Figure 4. Consistency in brain responses across sites while participants watched the same video, across a*

2   *variety of different videos, in the matched datasets. Top: brain maps show cross-site consistency at the*

3   *group level (as in Figure 2, see also Fig. 1A). Only parcellation scales of 100 and 1000 ROIs are shown for*

4   *space considerations. As in Figure 2, black along the midline in medial views indicates missing data, not*

5   *low correlations. Scan lengths vary across different videos. See also Supplemental Figure 2 for a version*

6   *of this figure that randomly downsamples to equate for scan length. Bottom: line plots show median*

7   *pairwise ISC among pairs of participants within each sites and across sites (see Fig. 1B), for the most*

8   *coarse and most fine parcellation scales. Please see Supplemental Table 1 for ROI labels, which are*

9   *omitted for readability. The line plots for Video 1 and 2 are the same values plotted on brains in Figure 3*

10   *(top) and Supplemental Figure 1. Note that individual data points are connected with a line to facilitate*

11   *comparing overall patterns, but these plots are not time series. Rather, each data point reflects median*

12   *similarity across pairs of timeseries. When values for a given ROI are similar across different scans (e.g.*

13   *x = 60 for top two line plots), that reflects comparable levels of similarity across entirely different brain*

14   *response time series for different videos.*

Video 1

Video 2

1
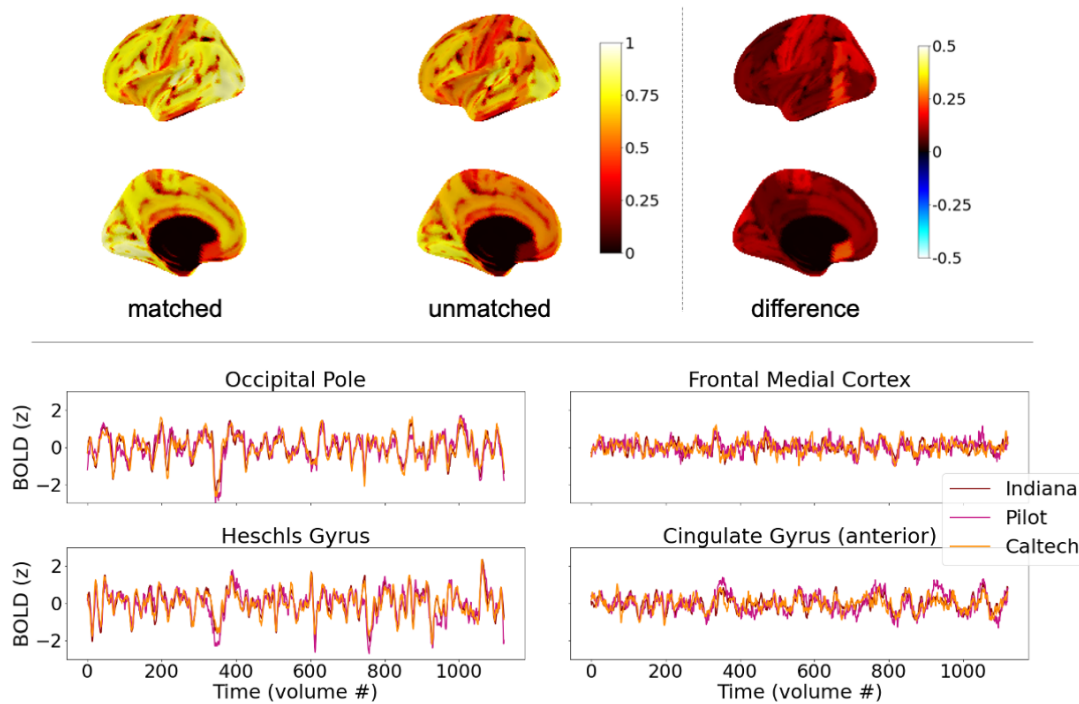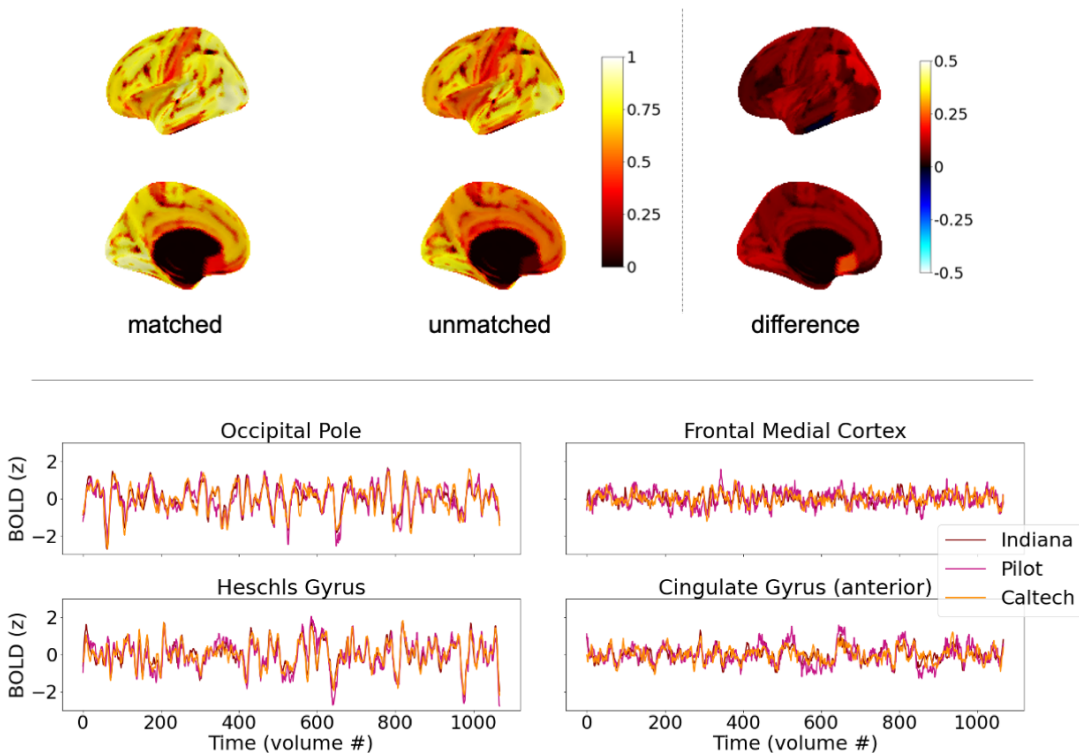
1   *Figure 5. Exploratory comparison of brain response consistency for matched datasets and unmatched*

2   *datasets, at the level of average time series (see also Fig. 1A), for Videos 1 and 2. Left and center brain*

3   *maps show correlations between median time series for all participants at each site (as in Fig. 2, top);*

4   *right shows the difference of the two maps. Matched datasets are Indiana and Caltech (as in Fig. 2);*

5   *unmatched datasets are Pilot and Caltech. Black along the midline in medial views indicates missing*

6   *data, not low correlations. Similarity of average time series across datasets is high in general, but*

7   *highest when acquisitions are matched. Time series figures show median time series for each site in*

8   *sensory areas (left) and association areas (right). Alll panels use the Harvard-Oxford 96-ROI cortical*
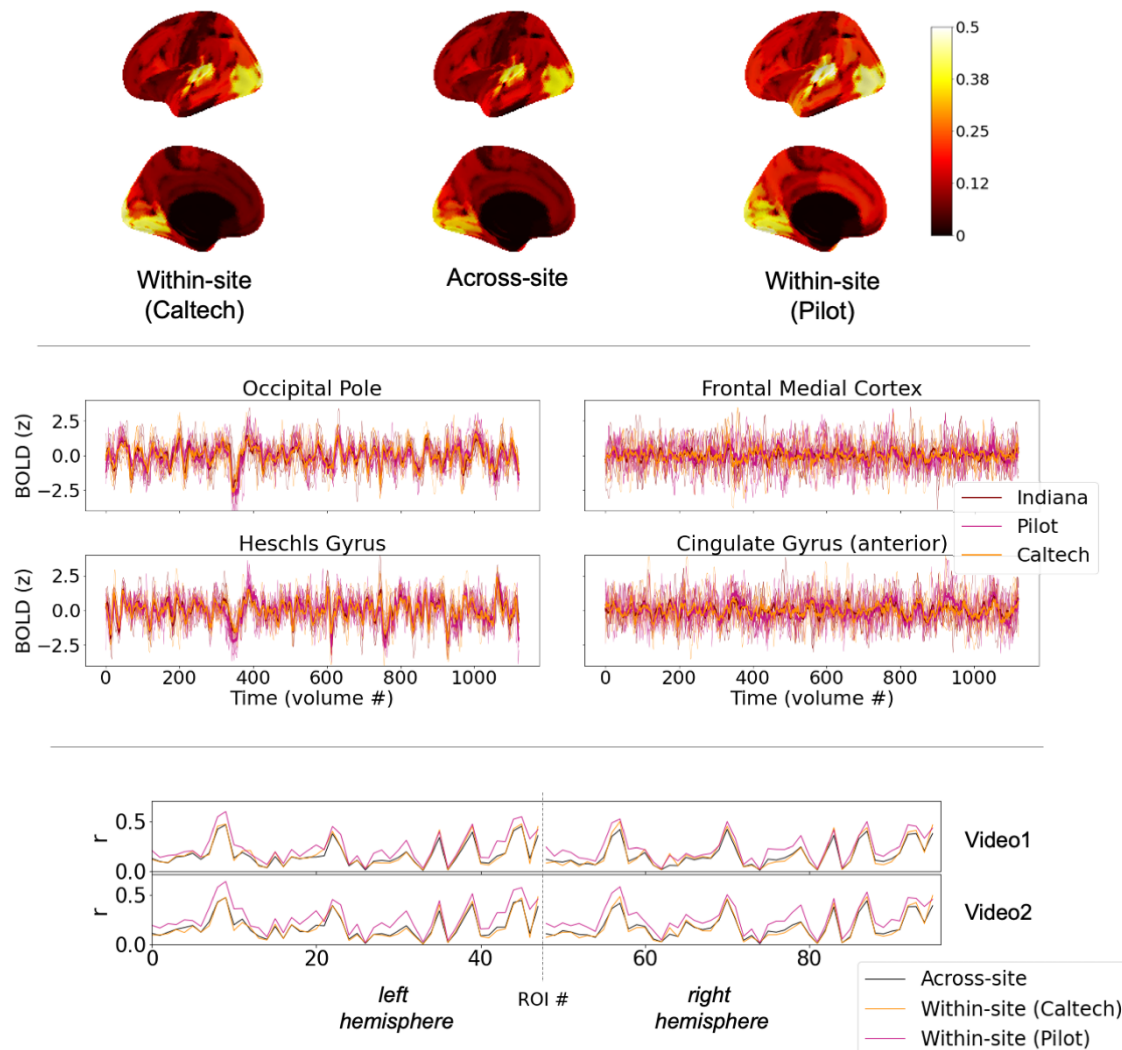
9   *parcellation.*

1

Figure 6. Exploratory comparison of brain response consistency across unmatched datasets, at the level of individual time series (Fig. 1B), for Video1. Top: brain maps depict magnitudes of medians of pairwise correlations between participant brain response time series in each region. The left and right columns show correlations among pairs of participants at the same site (left: Caltech; right: Pilot). The center column shows correlations among pairs of participants spanning different sites (as in Fig. 1B, green). Black along the midline in medial views indicates missing data, not low correlations. Center: Time series plots show five randomly-selected individual time series for each site in sensory areas (left) and association areas (right), along with the median time series across all participants at that site

1   *superimposed in bold. Bottom: Within- and across-site ISC values from top panel (Video1) and*

2   *Supplemental Figure 3 (Video2) presented as a line plot, to facilitate comparison. See text for summary*

3   *of differences and effect sizes, and see Supplemental Table 1 for ROI labels, which are omitted for*

4   *readability. As in Figure 4 (bottom), individual data points are connected with a line, but these plots are*

5   *not time series. Rather, each data point reflects median similarity across pairs of timeseries. When*

6   *values for a given ROI are similar across the two different scans, that reflects comparable levels of*

7   *similarity across entirely different brain response time series evoked by different videos. All panels use*

8   *the Harvard-Oxford 96-ROI cortical parcellation.*

9
10